ACADEMY OF SCIENCES OF THE ESTONIAN S.S.R.
INSTITUTE OF LANGUAGE AND LITERATURE

CONTENTS

INDEX OF AUTHORS

ADAPTIVE VARIABILITY AND ABSOLUTE CONSTANCY IN SPEECH SIGNALS: TWO THEMES IN THE QUEST FOR PHONETIC INVARIANCE

Björn Lindblom

Department of Linguistics, University of Stockholm, Sweden

ABSTRACT

Our topic is the classical problem of reconciling the physical and linguistic descriptions of speech: the invariance issue. Evidence is first presented indicating the possibility of defining phonetic invariance at the articulatory, acoustic or auditory levels of the speech signal. However, as we broaden the scope of our review, we find that attempts to define phonetic invariance in terms of absolute physical constancies tend to lose ground to theories that recognize signal variability as an essentially systematic and adaptive consequence of the informational mutuality of natural speaker-listener interactions. We reach this conclusion not only by examining experimental data on on-line speech processes but also by analyzing typological evidence on how the phonetic structure of consonant systems varies with inventory size in a lawful manner.

INTRODUCTION

Traditionally the problem of invariance in phonetics can be said to be that of proposing physical descriptions of linguistic entities that have the characteristic of remaining invariant across the large range of contexts that the communicatively successful real-life speech acts present to us.

Many of us share the conviction that taking steps towards the solution of this problem will be crucial if we are to acquire a deeper theoretical understanding of the behavior of speakers and listeners as well as develop more advanced systems for speech-based man-machine communication (Perkell&Klatt 1986).

The present paper will attempt to address some of the questions that we typically encounter in the search for invariance. We shall do so by summarizing research undertaken mostly in our own laboratory in Stockholm. Although thus deliberately limiting the scope of our review we hope that the issues raised will nevertheless be of sufficient interest to stimulate general discussion.

IS PHONETIC INVARIANCE ARTICULATORY?

A few decades ago phoneticians began to interpret phonetic events by comparing articulators to highly damped oscillatory systems. More recently, such models have acquired an important role within the framework of action theory (Kelso, Saltzman and Tuller 1986). In the sixties it was hoped that a lot of the variability that speech signals typically exhibit - e g reductions and vowel-consonant coarticulation (Öhman 1967) - could be explained in terms of the spatial and temporal overlap of adjacent "motor commands" (MacNeilage 1970). Articulatory movements were seen as sluggish responses to an underlying forcing function which was assumed to change, usually in a step-wise fashion, at the initiation of every new phoneme (Henke 1966). Owing to variations in say stress or speaking tempo, different contexts would give rise to differences in timing for a given sequence of phoneme commands. Articulatory and acoustic goals would not always be reached, the so-called 'undershoot' phenomenon (Stevens and House 1963). But since such undershoot appeared to be lawfully related to the duration and context of the gestures (Lindblom 1963), the underlying articulatory "targets" of any given phoneme - 'die Lautabsicht' - would nevertheless, it was maintained, remain invariant. Accordingly, at that time it seemed possible to argue that phonetic invariance might be articulatory.

Duration-dependent undershoot still seems to to be a phonetically valid notion for biomechanical reasons. But it is clearly not as inevitable a phenomenon as was first thought. Current experimental information indicates that in fast speech articulatory and acoustic goals can be attained despite short segment durations (cf Engstrand 1987, Gay 1978, Kuehn and Moll 1976). Furthermore undershoot has been observed in unstressed Swedish vowels that exhibit long durations owing to 'final lengthening' (Nord 1986). Such deviations from simple duration-dependence appear to highlight the reorganizational abilities of the speech production system. One way of resolving the problem posed by these somewhat contradictory results might be obtained if it were shown that when instructed to speak fast, subjects have a tendency to "overarticulate", thus avoiding undershoot to some extent, whereas when destressing they are more prone to "underarticulate" (cf discussion below of hypo- and hyper-speech). The demonstration of language-specific patterns of vowel reduction

(cf Delattre's 1969 discussion of English, French, German and Spanish) becomes particularly relevant in the context of addressing such questions.

In summary, the original observations of 'undershoot' carried the implication that the invariant correlates of linguistic units were to be found, not in the speech wave nor at an auditory level, but upstream from the level of articulatory movement. Phonetic invariance was accordingly associated with the constancy of underlying "spatial articulatory targets" (for reviews of the target concept see e g MacNeilage 1970, 1980). However, subsequent experimentation - some of which we already hinted at above - has revealed that the notion of segmental target must be given a much more complex interpretation.

This conclusion is reinforced particularly strongly by studies of **compensatory articulation**. Let us summarize some results from an experiment using the so-called "bite-block" paradigm (Lindblom, Lubker, Lyberg, Branderud, Holmgren in press). Native Swedish speakers were asked to pronounce monosyllables and bi- and trisyllabic words under two conditions: normally and with a large bite-block between their teeth. They were instructed to try to produce the bite-block utterances with the same rhythm and stress pattern as the corresponding normal items. Real Swedish words as well as "reiterant" nonsense forms were used: To exemplify, one of the metric patterns was: - '- - . This pattern would occur in the lists as "begabba" and /ba'bab:ab/. Measurements were made of the duration of the consonant and vowel segments of the normal and the bite-block versions of the reiterant speech samples. The question was thus whether subjects would be able achieve the bilabial closure for the /b/ segments in spite of the abnormally low and fixed jaw position and whether they would be able to do so reproducing the normal durational patterns.

We found that the timing in the bite-block words deviated systematically but very little from the normal patterns and concluded that our subjects were indeed capable of compensating. To explain the results we suggested that a representation of the "desired end-product" - the metric pattern of the word - must be available in some form to the the subjects' speech motor systems and that the successful compensations implied a reorganization of articulatory gestures that must have been controlled by such an output-oriented target representation. These results are in agreement with those reported earlier by Netsell, Kent and Abbs (1978). Moreover, they are completely analogous to the previous demonstrations that naive speakers are capable of producing isolated vowels whose formant patterns are normal at the first glottal pulse in spite of an unnatural jaw opening imposed by the use of a "bite-block" (Lindblom, Lubker and Gay 1979, Gay, Lindblom and Lubker 1981).

These results bear on the recent discussion of speech timing as "intrinsically" or "extrinsically" controlled. Proponents of action theory (Fowler, Rubin, Remez and Turvey 1980) approach the physics of the speech motor system from a dynamical perspective with a view to reanalyzing many of the traditional notions that now require explicit representation in extant speech production models such as 'feedback loop', 'target' etc. Their writings convey the expectation that many aspects of the traditional "translation models" will simply fall out as consequences of the dynamic properties intrinsic to the speech motor system. In the terminology of Kelso, Saltzman and Tuller (1986, 55) "...., both **time** and **timing** are deemed to be intrinsic consequences of the system's dynamical organization." Methodologically, action theory is commendable since, being committed to interpreting phonetic phenomena as fortuitous (intrinsic) consequences rather than as controlled (extrinsic) aspects of a speaker's articulatory behavior, it guarantees a maximally thorough examination of speech production processes. However, it is difficult to see how, applying the action theoretic framework to the data on compensatory timing just reviewed, we could possibly avoid postulating some sort of "temporal target" representation which is (i) extrinsic to the particular structures executing the gestures and which is (ii) responsible for extrinsically tuning their dynamics. Speech production is a highly versatile process and sometimes appears strongly listener-oriented.

The plasticity of the speech motor system is further illustrated by an experiment recently done by Schulman (forthcoming) invoking a "natural bite-block" situation. This condition is provided by loud **speech** in which a more open mandible tends to be used than in normally spoken syllables.

Whether rounded or not the vowels of loud test words produced by Schulman's talkers were found to exhibit almost three times as large jaw openings as the corresponding segments in the normal words. In the context of compensatory articulation two observations call for special comments. Why do not speakers compensate for the greater jaw opening in the loud vowels the way they do in the bite-block experiments? Schulman shows that they do not since the fundamental frequency and (as predicted by articulatory-acoustic nomograms) the first formant of the loud vowels are shifted upwards by about one Bark whereas the other formants do not undergo comparable modification. (Below we shall relate the F1 and F0 shift to the results of a perceptual experiment).

The other finding of interest is the fact that loud vowel durations increase whereas loud consonant durations tend to decrease (cf Fonágy and Fonágy 1966). What does that result mean? The normal-loud vowel duration differences look suspiciously similar to the durational differences between normal open and close vowels which have been observed for many languages (Lehiste 1970). Finding that the duration of the EMG recorded from the anterior belly of the digastric correlated with both mandibular displacement and vowel duration Westbury and Keating (1980) suggest that this temporal variation among vowels, although non-distinctive, must be seen as present in the neuromuscular signals controlling their articulation. An alternative interpretation would be to regard the differences as automatic consequences of an interaction between an invariant underlying "vowel duration command" and articulatory inertia (cf Keating 1985 for further discussion). In (Lindblom 1967) we reported some evidence in favor of the latter interpretation, the "extent of movement hypothesis" (Fischer-Jørgensen 1964). We also found that the durational consequences of more extensive articulatory gestures were sometimes actively counteracted.

The question whether the open-close vowel duration difference is an intrinsic or extrinsic phonetic phenomenon is accordingly somewhat controversial. Schulman's findings bear on the problem. He constructed a model of loud speech based on the observation that loud movements appear to be "exaggerated" versions of the normal movements. Assuming that the lips and the jaw are linear mechanical systems and that loud differs from normal speech solely in terms of the amplitudes of the underlying excitation forces he performed a linear scaling of all articulatory parameters recorded for normal syllables (vertical displacements of upper and lower lips and jaw) and combined the scaled curves so as to derive the vertical separation of the lips - the parameter that determines the open-closed state of the mouth opening. By using the value of this parameter at opening and closing in the normal syllables as his criterion he was then able to predict the durations of vowel and consonant segments for loud speech. He found that linear scaling eliminated stop closures entirely or produced much too long vowels.

The implication of this result is that it clearly attributes the durational differences to a superposition effect, that is the interaction arising from the superposition of the lip and the jaw movements. Schulman concludes that, unless the effect of opening and closing of the jaw had been actively counteracted, loud and normal vowel durations would have differed even more than they actually did.

Let us remark in the present context that, while it appears reasonable to suggest, as do Westbury and Keating, that the acoustic vowel duration differences are probably reflected at a level of neuromuscular control, there is also evidence indicating that the function of neural control signals may be a compensatory rather than a positive one, that is a function opposite to that suggested by Westbury and Keating.

The preliminary implication of all work touching the theme of compensatory articulation appears to be that - whether we use "target" with reference to segmental attributes, segment durations or patterns of speech rhythm - the term is better defined, not in terms of any simple articulatory invariants, but with respect to the acoustic output that the talker wants to achieve. If phonetic invariance is not articulatory could it be acoustic then?

IS PHONETIC INVARIANCE ACOUSTIC?

The suggestion that the speech signal contains absolute physical invariants corresponding to phonetic segments and features has received a lot of attention thanks to the work by Stevens and Blumstein (Stevens and Blumstein 1978, 1981; Blumstein and Stevens 1979, 1981). The idea has been favorably received by many, for instance Fowler in her attempts to apply the perspective of **direct perception** to speech (Fowler 1986).

Others have been provoked to emphasize the inadequacy of the non-dynamic nature of the Stevens template notion (Kewley-Port 1983) and the substantial context-dependence that the stop consonants of various languages typically display even in samples of carefully enunciated speech (Öhman 1966).

Recent work by Krull and Lacerda in our Stockholm laboratory uses the method of quantifying the extent of consonant-vowel coarticulation in the form of linear "locus equations". These relationships are obtained by plotting formant frequencies at CV_2- and V_1C-boundaries as a function of the formants for V_2 and V_1 respectively. Acoustic theory indicates that for the consonant-vowel combinations in question near-linear relationships should be expected. Such diagrams show clearly that, although a "locus" pattern can exhibit considerable variation, it is predictable from information on stop consonant identity and adjacent vowel context. Here coarticulation stands out as the salient fact and the lack rather than the presence of absolute acoustic invariance tends to be reinforced.

Incidentally, let us note that, if it exists, acoustic invariance is a strange notion since talkers can only monitor it through their senses and listeners can only access it through their hearing system. Why should sensory and auditory transduction be assumed to have a transfer function of one imposing no transformation? Is it the case that what people really mean when they talk about acoustic invariance is in fact "auditory" invariance? Let us look at some psycho-acoustic results.

IS PHONETIC INVARIANCE AUDITORY?

We mentioned earlier a perceptual result that offers a rather curious parallel to Schulman's findings. It is the "Traunmäller effect" which is a demonstration of the transforms required to preserve the perceptual constancy of vowel quality under changes in (i) vocal effort and (ii) vocal tract size. It is also somewhat reminscent of the findings on F0-Fi interrelationships in soprano vowels (Sundberg 1975).

Effort and vocal tract variations can be dramatically illustrated by synthetically modifying a naturally spoken /i/. When all formants and F0 are shifted equally along a Bark scale an /i/-like vowel is perceived but the voice changes from an adult's to a child's. When both Fi and F0 are varied in such a way that Fi-F0 is kept constant on a Bark scale - and the upper formant complex is left unchanged - an /i/-like vowel is perceived. This is remarkable in view of the fact that Fi reaches a value more typical of a low-pitched /ɑ/. One's impression is that the speaker remains the same but that she "shouts".

Note the parallel between Schulman's and Traunmäller's results. Are the findings causally related? Do we explain the lack of formant compensation in loud speech in terms of the Traunmäller effect? Or do we account for the vowel quality results in terms of the "Schulman" effect?

Of importance for the present discussion is the fact that behavioral constancies have been demonstrated and that they imply that at least in this case phonetic invariance must be defined at a level of auditory representation.

Let us return for a moment to the alleged invariance of the release spectra of stop consonants. Diana Krull collected perceptual responses from Swedish listeners to burst fragments obtained from $V_1C:V_2$ words (Krull 1987). One hundred test words were generated by constructing all possible combinations of V_1 or V_2 = short /i e a o u/ with C: = /b: d: rd: g:/. Confusion matrices for the burst stimuli demonstrate the drastic coarticulation effects. By and large, listener responses can be accounted for in terms of the acoustic properties of the stimuli. This is shown in her attempts to predict the confusions from auditorily based "perceptual distance" computations.

A related study has been carried out by Lacerda (1986). We can characterize one part of his research as variations on the theme struck by Flanagan in his early "difference limen" experiments on vowel formant frequencies (Flanagan 1955). Lacerda's question was: How well can listeners discriminate four-formant stimuli that differ solely in terms of the frequency of F2. His work permits us to compare a psycho-acoustic task: the discrimination of F2 in brief tone bursts with formant patterns static - with a "speech task": the discrimination of the onset of F2-transitions in /da/-stimuli.

The results indicate that the subjects' ability to discriminate on the psycho-acoustic task is in close agreement with Flanagan's findings whereas their performance on the /da/-stimuli is drastically impaired. One interpretation is that the discrimination change is related to the fact that intra-category discrimination is considerably worse than inter-category discrimination (Liberman, Harris, Hoffman and Griffith 1957).

With reference to the invariance issue it is important to note the following. Krull's results on stop perception indicate that the coarticulatory spectral variability of the stop releases is rather accurately reflected in the confusions that her listeners made of such brief sounds. This is fully compatible with Lacerda's results on tone bursts. Note that in Lacerda's speech-task test however, the variability does not seem to be as faithfully mirrored in the listeners' percepts for apparently they treat stimuli easily discriminable in psycho-acoustic tests as "the same". Whether it is the listener invoking the "speech mode" or it is the interaction of the dynamic stimulus properties and speech-independent auditory processing is an issue still worth addressing. However, our main point is this: The invariance that we discern in these findings is not acoustic. It clearly presupposes auditory processing.

IMPLICATIONS OF SPEAKING STYLE: THE HYPER-HYPO DIMENSION

Everyday experience indicates that speaking is a highly flexible process. We are capable of varying our style of speech from fast to slow, soft to loud, casual to clear, intimate to public. We speak in different ways when talking to foreigners, babies, computers and hard of hearing persons. And we change our pronunciation as a function of the social rules that govern speaker-listener interactions (Labov 1972).

Above we considered principally three types of phonetic invariance: articulatory, acoustic and auditory invariance. What are the implications of variations in speaking style for the invariance issue? For the purpose of our discussion let us give phonetic invariance a strong literal interpretation which is rather extreme but nevertheless not too far from working hypotheses explored previously by various investigators: "All the information is in the signal, particularly in its dynamics". For such a view of invariance to be correct - let us call it the strong version of absolute physical invariance - the following must be true: Talkers vary their speaking style and thereby contribute to increasing the variability of the speech wave but in utterances that are intelligible linguistic units will always exhibit a core of invariant physical information that will remain undestroyed so as to be successfully used by a listener.

We recently undertook a literature survey+ in order to systematize the types of speech materials that have been used in acoustic phonetic studies published during the past ten years in J Acoust Soc Am, J of Phonetics, Language and Speech, and Phonetica. A total of over 700 articles were selected as preliminarily relevant. We ended up choosing 216 as meeting our criterion of "descriptive study of speech based on quantitative acoustic phonetic measurements".

Of special interest to us was to ascertain the relative proportions of studies investigating "self-generated" speech (including e g spontaneous conversation) on the one hand and speech samples chosen by the experimenter (e g list readings, nonsense words etc) on the other. Not surprisingly, we found that the majority of studies, over 90%, use experimenter-controlled speech samples. The reason is clear. A satisfactory experimental design presupposes good control of the variables involved. This is less of a problem if the experimenter determines the test items but for "real speech" with its immense number of variables there is no established methodology that will guarantee such control. So rather than drown in an ocean of "unknown factors" our strategy tends naturally to become one of resorting to "given" test materials and read speaking modes.

One way of justifying this widely used procedure is to argue that first we will solve the problem of phonetic invariance in "lab speech". Then we will get to work on "natural speech". Another outlook might be to suggest that, although we lack the supplementary methodology required by "ecological" speech, the excessive use of "lab speech" introduces an undesirable bias in our data bases as well as in our theoretical intuitions about invariance and other key issues - a bias that might make us underestimate the problem of speech variability in spite of the fact that it is readily acknowledged by all workers in the field and has already, it would appear, been rather massively documented. Consequently the situation ought to be balanced.

We have recently been persuaded by the latter point of view and are currently recording (1) "self-generated speech" produced under natural conditions and (2) parallel "citation form" speech based on the syllables, words and phrases that occur in the spontaneous materials. Data are currently being collected by Rolf Lindgren, Diana Krull and myself using this two-pronged approach involving comparisons of reference pronunciations ("citation form" speech) with samples of "self-generated speech". A few preliminary observations can be made that bear on the present discussion (cf also Lindblom and Lindgren 1985).

The reductions that we have found in spontaneous speech - and often escape the trained phonetic ear even after spectrographic evidence has been examined - are sometimes drastic. Speaking style has marked effects on the acoustic patterns of words. The vowel space shrinks in casual style and is expanded in "hyperspeech" modes. The diphthongization of tense Swedish vowels is enhanced and is particularly apparent in clear speech. Contrast in VOT for voiced and voiceless stops increases and decreases as we compare hyper- and hypo-forms respectively. Locus equations show a smaller slope (=less vowel-dependence) for citation form pronunciations than for spontaneous speech which we interpret to indicate that vowel-consonant coarticulation is counteracted in hyperspeech (more invariance) but tolerated in hypospeech (less invariance). Although preliminary the observations made so far suggest that the prospects for any strong version of absolute physical invariance to be substantiated seem most unfavorable.

SPEECH UNDERSTANDING: (IN)DEPENDENCE OF SIGNAL INFORMATION

At the Department of Romance Languages at Stockholm University a test is used to measure how proficient native Swedish students are in understanding spoken French in which the task of the students is to listen to triads of stimuli consisting of two identical sentences and one minimally different and to indicate the odd case.

Montre leur ce chapeau s'il te plaît
Montre leur ce chapeau s'il te plaît
Montre leur ces chapeaux s'il te plaît

Native speakers of French have no problems of course with such sentences whereas Swedish listeners knowing no French have a lot of trouble. However, when the key information - e g the ce/ce/ces triad - is presented as fragments gated from the original sentences the performance of the Swedish subjects improves radically (Dufberg and Stöök forthcoming).

This test can serve to remind us that perception is a product of two things: signal-dependent and signal-independent information. While I am perfectly capable of discriminating the French minimal contrasts as auditory patterns I would quickly lose those patterns in a sentence context unless I have a sufficiently good command of French - that is access to signal-independent 'knowledge' whose interaction with the signal is a part of forming of the final percept.

The speech literature is full of experimental data indicating that processes not primarily driven by the signal play an

---

+I am indebted to Diana Krull for doing the preliminary selections and to Natasha Beery of the Phonology Laboratory, University of California, Berkeley for the statistical analyses.

important role in the perception of speech. There will not be time to do justice to all the research bearing on this issue. Let me just recall some well-known paradigms: Perception of speech in the presence of various disturbances (noise and distortion). The improvement of identification as the signal gets linguistically richer (Miller, Heise and Lichten, Pollack and Pickett 1964 and by Miller&Isard). Detection of deliberate mispronunciations (Cole 1973). Word frequency effects (Howes, Savin). Restoration (Warren 1970, Ohala and Feder 1986). Phoneme monitoring (Foss&Blank). Word recognition from word fragments (Grosjean 1980, Nooteboom 1981). Fluent restorations in shadowing mispronunciations (Marslen-Wilson and Welsh 1978). Verbal transformations (Warren). Intelligibility of lip-reading from video-recordings supplemented by "hummed speech" - an audio signal processed to contain primarily rhythm and intonation cues (Risberg 1979). Inferences from historical sound changes (Ohala 1981).

CONCEPTUALIZING SPEAKER-LISTENER INTERACTIONS

Our review of experimental evidence bearing on the invariance issue has been selective but should nevertheless provide a rough indication of a panoply of alternative positions and their respective pro's and con's. We have considered the suggestion that the invariance of phonetic segments be defined: (i) at an articulatory level (e g the "spatial target" hypothesis); (ii) at an acoustic level (e g spectral properties of stops); (iii) at an auditory level (e g perceptual constancy of vowel quality). Which of these alternatives should we put our money on?

When pursued experimentally articulatory, acoustic or auditory definitions of invariance have the methodological virtue of encouraging a maximally thorough search at these particular levels. But in seeking a broader theoretical understanding of speech communication we would stand little to gain from spending effort on choosing between levels. Such an appoach misreads the evidence which, when viewed in a broader perspective, strongly points to the conclusion that: The invariance problem is not a phonetic issue at all for ultimately invariance can be defined only at the level of listener comprehension.

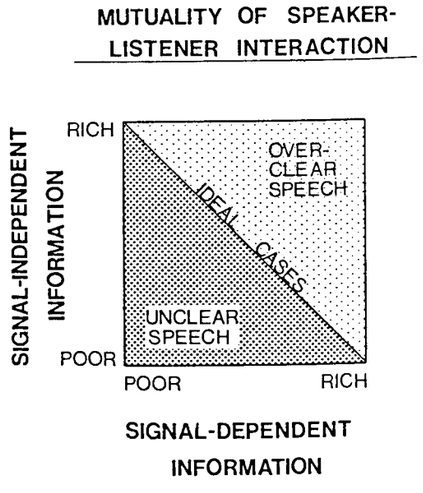
We can convince ourselves of the correctness of that point by considering the following phrase in English: /lesnsevn/. We can hear this utterance either as LESS THAN SEVEN, or as LESSON SEVEN. In the appropriate contexts (say "How many are coming", and "What is our topic to-day?") the listener will not be aware of any ambiguity. At which phonetic level do we find the physical correlates of the initial segments of the word "than"? Needless to say there ARE no such correlates in this particular case. The conclusion seems inescapable: We should not put our money on any of the above alternatives. We must seek a more general theory.

The experimental data on production indicates that the behavior of the speech motor system is shaped primarily by two forces - plasticity (listener-oriented reorganization) and economy (talker-oriented simplification) - which interact on a short-term basis so as to generate signals that may be "rich or poor" in explicit physical information.

The evidence on perception has identified two major sources of information: signal-dependent and signal-independent processes and suggests that on a short-term basis percepts arise from the latter (i e "context") modulating the former in an analogously "rich or poor" manner.

One possible way of schematizing the logical possibilitites of these conceptual simplifications is shown in the diagram of the enclosed figure. This is not a very rigorous scheme but seems useful, at least pedagogically, in contrasting some of the ideas currently entertained in phonetics (cf J of Phonetics, January issue 1986).

This graph states that for speech to be intelligible the sum of explicit physical information and signal-independent information must be above a threshold, that is the 135 degree line. In the ideal case this sum equals a constant the x- and y-values of specific speech samples falling right on that line. Points above the line are associated with what might be termed "over-clear" speech, points below it with "unintelligible" speech.



It appears reasonable to assume that in the real-life situations utterances can vary tremendously with respect to how socially and communicatively successful they prove to be. For our present purposes let us focus on speech samples from hypothetically successful real-life speaker-listener interactions and assume that they produce data points clustering near and above the slant line. What would such a result imply? It would mean that there is a complementary relation between the amounts of information contributed by signal attributes on the one hand and "context" on the other. When speakers come close to the slant line it would indicate first of all that they are capable of varying their speech output in a plastic way (cf evidence on hypo-hyper-speech modes and other instances of reorganization of speech motor control) and secondly that, while perhaps not being perfect 'mind-readers', they are at least capable of adapting their speech on-line to the short-term fluctuations in the listener's access to "context" or signal-independent information (cf experimental documentation of numerous cases showing that listeners are in fact capable of successfully coping with highly context-dependent reduced and coarticulated speech stimuli). The possibility of such complementarity in real speech emerges also from some recent measurements reported by Hunnicutt (1985) as well as from Lieberman's 1963 study.

If we hypothesize that this strategy - let us call it the STRATEGY OF ADAPTIVE VARIABILITY - comes near the way real speakers actually behave when they are communicatively successful, we obtain a natural way of resolving some of the paradoxes that surround the invariance issue. For it follows that intra-speaker phonetic variation - along a hyper-hypo-continuum as well as along other dimensions - is the characteristic that we should expect the units of ecological speech to exhibit - not absolute physical invariance.

The proposed way of thinking about the issue does not, of course, rule out finding physical speech sound invariance in restricted domains of observation but it does explain why our quest for a general concept of phonetic invariance has been largely unsuccessful. And, in a pessimistic vein, it predicts in fact that it will continue to be so.

Our reasoning leads us back to a conclusion already drawn by MacNeilage in his 1970 review of the invariance issue:

> "...the essence of the speech production process is not an inefficient response to invariant central signals, but an elegantly controlled variability of response to the demand for a relatively constant end (p 184)".

If, as suggested here, we take the "relatively constant end" to be defined neither articulatorily, acoustically nor auditorily but specified only with reference to "the level of listener comprehension" MacNeilage's formulation still captures the "essence of the speech production process" satisfactorily.

Let us pause to reflect on some of the implications of the two theories contrasted in our discussion: Absolute Physical Invariance versus Adaptive Variability. The former, if proved correct, would transform what currently looks like instances of massive variability into artefacts. For this theory says in fact that there simply IS NO variability of linguistic units; There seems to be but that is merely a result of our presently inadequate conceptual and experimental tools. Further note that if we push the notion of absolute constancy to its extreme another implication can be noted, namely that the transmission of information by speech - an undeniably biological process - is basically non-adaptive.

The Theory of Adaptive Variability, on the other hand, says exactly the opposite. This is a theory for which it is easier to find support within the general study of the biology of motor control and perception. It is precisely by emphasizing the adaptive nature of speech processes that we obtain a principled way of investigating phonetic variation and its origin.

ON-LINE PROCESSES IN THE LIGHT OF TYPOLOGICAL EVIDENCE ON CONSONANT SYSTEMS

Some time ago Nooteboom did an experiment on word retrieval and was able to show that listeners perform better if presented with the first halves of words than on the corresponding second-half fragments (Nooteboom 1981). For an explanation he suggested that, since word recognition is a real-time left-to-right process, word beginnings are less predictable than word endings. Consequently left-to-right context can be much more easily used than right-to-left context.

He concluded his paper by raising the question whether this asymmetry - that he takes to be a universal feature of the perceptual processing of any language - might have left its imprint on how lexical information is organized in the languages of the world. He predicted (p 422) that: "(1) in the initial position there will be a greater variety of different phonemes and phoneme combinations than in word final position, and (2) word initial phonemes will suffer less than word final phonemes from assimilation and coarticulation rules."

One basic assumption is that variations in perceptual predictability correlate with signal "distinctiveness". Hence "the greater variety of different phonemes and phoneme combinations" in the initial as compared with the final position of words. Restating the

idea we can say that a larger paradigm goes with a RICHER signal inventory. The other side of the coin is of course that a smaller paradigm - such as that attributed to word endings - goes with a POORER signal inventory. In suggesting that the presence of assimilation and coarticulation should vary inversely with the need for keeping items distinct Nooteboom tacitly formulates a hypothesis that comes close to the theory of Adaptive Variability described here. Note that the theory Absolute Physical Invariance does not offer us any basis at all for making predictions about a possible interplay between language structure and on-line processing. Why? As stated earlier according to that theory there IS no phonetic variation, there only seems to be. The idea of language structure adapting to the on-line constraints of speaking and listening only becomes a possibility once we recognize the existence and systematic nature of phonetic variation. Only from that point of departure will we be able to address the question of what feeds the processes of phonological innovation.

We shall not be in a position to present the typological data needed to test Nooteboom's hypothesis. However, we shall conclude our paper by presenting some other data that do bear on it and strongly encourage further examination of the underlying ideas.

In collaboration with Ian Maddieson we recently undertook an analysis of the consonant inventories of 317 languages, carefully selected so as to constitute a reasonable sample of the "languages of the world". Our corpus was that of UPSID, the UCLA Phonetic Segment Inventory Database (Maddieson 1984). The data consists of lists of systems whose elements (allophones of major phonemes) are specified in phonetic transcription.

Inventory sizes range from 6 to 95 consonants per system. The materials lend themselves to testing a paraphrase of Nooteboom's hypothesis: Is the phonetic structure of consonant systems independent of their size? Or is it systematically related to that dimension? If there is a systematic size-dependence what is it?

There is neither time nor space to give the details of the analysis. They will be published elsewhere (Lindblom, MacNeilage and Studdert-Kennedy; Lindblom and Maddieson forthcoming). Fortunately, Nooteboom's perspective provides us with a way of summarizing the main findings.

It turns out that small paradigms statistically favor segments with both phonatory and articulatory properties that can be classified as basic or elementary. Medium-sized paradigms tend to include consonants invoking more elaborated gestures in addition to a core of basic elements. The largest systems use both these types but also combinations of elaborated gestures that we label complex articulations. To exemplify, plain /p t k/ are classfied as "basic" articulations whereas ejective /pˀ tˀ kˀ/ or aspirated /pʰ tʰ kʰ/ invoke "elaborated" mechanisms. A segment such as /t͡sʰ/ is "complex" since it shows more than one elaboration: both of place (retroflexion) and source features (aspiration). Logically a six-consonant system could use the ejective set for its stop series. Small systems never do in our material whereas medium-sized and large systems do. Moreover, the "complex", multiply elaborated segments are most frequent in the large inventories. The basic rule is that a less simple consonant tends not to be recruited without the presence of parallel more simple ("basic" or "elaborated") series (cf the notion of 'implicational hierarchy' of traditional terminology). The claim we make is accordingly that we see a positive correlation between paradigm size and the number of elements that a sound pattern selects from a dimension of "articulatory complexity".

The validity of our analysis naturally hinges on the success with which we can give non-circular, independently motivated definitions of "articulatory complexity". When it comes to the details of the analysis that problem is a topic for future quantitative phonetic theory. For the moment we believe that the major trends are rather gross effects that can be convincingly demonstrated by the force of the examples. They permit us to make the following generalization: Small consonant paradigms invoke 'unmarked' phonetics, large paradigms 'marked' phonetics. That is of course exactly what Nooteboom's hypothesis predicts and it takes a few steps towards an explanation for why seven-consonant systems do not show inventories like the following (Ohala 1980):

[ɗ kʼ tsʲ ɬʲ m r ɬ]

We take the present typological data on consonant systems as providing strong evidence in favor of (a) language structure evolving as an adaptation to the constraints of the on-line processes of speaker-listener interaction; and for (b) the correctness of a theory of Adaptive Variability as an account of those processes.

REFERENCES

Blumstein S and Stevens K N (1979): "Acoustic Invariance in Speech Production: Evidence from Measurement of the Spectral Characteristics of Stop Consonants", J Acoust Soc Am 72, 43-50.

Blumstein S and Stevens K N (1981): "Phonetic Features and Acoustic Invariance in Speech", Cognition 10, 25-32.

Cole R A (1973): "Listening for Mispronunciations: A Measure of What We Hear during Speech", Perception and Psychophysics 13, 153-156.

Delattre, P (1969): "The General Phonetic Characteristics of Languages: An Acoustic and Articulatory Study of Vowel Reduction in Four Languages", Mimeographed Report, University of California, Santa Barbara.

Engstrand, O (1987): "Articulatory Correlates of Stress and Speaking Rate", accepted for publication in J Acoust Soc Am.

Flanagan, J (1955): "A Difference Limen for Vowel Formant Frequency", J Acoust Soc Am 27:613-614.

Fischer-Jørgensen E (1964): "Sound Duration and Place of Articulartion", Zeitschrift für Sprachwissenschaft und Kommunikationsforschung 17:175-207.

Fonágy I and Fonágy J (1966): "Sound Pressure Level and Duration", Phonetica 15:14-21.

Fowler C A, Rubin P, Remez R E and Turvey M T (1980): "Implications for Speech Production of a General Theory of Action", 373-420 in Butterworth, B (ed): Language Production, vol I, London:Academic Press.

Gay, T (1978): "Effect of Speaking Rate on Vowel Formant Movements", J Acoust Soc Am 63(1):223-230.

Gay T, Lindblom B and Lubker J (1981): "Production of Bite-Block Vowels: Acoustic Equivalence by Selective Compensation", J Acoust Soc Am 69(3), 802-810.

Grosjean, F (1980): "Spoken Word Recognition and the Gating Paradigm", Perception and Psychophysics 28, 267-283.

Henke, W J (1966): Dynamic Articulatory Model of Speech Production Using Computer Simulation, Doctoral dissertation, M.I.T.

Hunnicutt, S (1985): "Intelligibility versus Redundancy - Conditions of Dependency", Language and Speech 28(1):47-56.

Keating, P (1985): "Universal Phonetics and the Organization of Grammars", 115-132 in Fromkin, V A (ed): Phonetic Linguistics, Orlando, FL:Academic Press.

Kelso J A S, Saltzman, E L and Tuller, B (1986): "The Dynamical Perspective on Speech Production: Data and Theory", J of Phon 14:1, 29-59.

Kewley-Port, D (1983): "Time-varying Features as Correlates of Place of Articulation in Stop Consonants", J Acoust Soc Am 73:322-355.

Krull, D (1987): "Evaluation of Distance Metrics Using Swedish Stop Consonants", paper submitted to the XIth ICPhS, Tallinn, Estonia.

Kuehn, D P and Moll, K L (1976): "A Cineradiographic Study of VC and CV Articulatory Velocities", J of Phon 4:303-320.

Labov, W (1972): Sociolinguistic Patterns, Philadelphia:University of Pennsylvania.

Lehiste, I (1970): Suprasegmentals, Cambridge, MA:MIT Press.

Lieberman, P (1963): "Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech", Language and Speech 6:172-187.

Liberman A M, Harris K S, Hoffman H S and Griffith B C (1957): "The Discrimination of Speech Sounds within and across Phoneme Boundaries", J of Experimental Psychology 54:358-368.

Lindblom, B (1963): "Spectrographic Study of Vowel Reduction", J Acoust Soc Am 35:1773-1781.

Lindblom, B (1967): "Vowel Duration and a Model of Lip Mandible Coordination", STL-QPSR 4/1967, 1-29 (Dept of Speech Communication, RIT, Stockholm).

Lindblom B, Lubker J and Gay T (1979): "Formant Frequencies of Some Fixed-Mandible Vowels and a Model of Speech Motor Programming by Predictive Simulation", J of Phonetics 7, 147-161.

Lindblom B, Lubker J, Lyberg B, Branderud P and Holmgren K (in press): "The Concept of Target and Speech Timing", to appear in Festschrift for Ilse Lehiste.

Lindblom, B and Lindgren R (1985): "Speaker-Listener Interaction and Phonetic

Variation", Perilus IV, Dept of Linguistics, University of Stockholm.

Lindblom B, MacNeilage P and Studdert-Kennedy M (forthcoming): Evolution of Spoken Language, Orlando, FL:Academic Press.

Lindblom, B and Maddieson, I (1988): "Phonetic Universals in Consonant Systems", to appear in Hyman, L M and Li, C N (eds): Language, Speech and Mind, Croom Helm.

MacNeilage, P (1970): "Motor Control of Serial Ordering of Speech", Psychological Review 77:182-196.

MacNeilage, P (1980): "Speech Production", Language and Speech 23(1), 3-24.

Maddieson, I (1984): Patterns of Sound, Cambridge:Cambridge University Press.

Marslen-Wilson, W D and Welsh, A (1978): "Processing Interactions and Lexical Access during Word Recognition in Continuous Speech", Cognitive Psychology 10, 29-63.

Netsell R, Kent, R and Abbs J (1978): "Adjustments of the Tongue and Lip to Fixed Jaw Positions during Speech: A Preliminary Report", Conference on Speech Motor Control, Madison, Wisconsin.

Nooteboom, S G (1981): "Lexical Retrieval from Fragments of Spoken Words: Beginnings vs Endings", J of Phonetics 9, 407-424.

Nord, L (1986): "Acoustic Studies of Vowel Reduction in Swedish", STL-QPSR 4/1986, 19-36 (Dept of Speech Communication, RIT, Stockholm).

Ohala, J J (1980): "Chairman's Introduction to Symposium on Phonetic Universals in Phonological Systems and their Explanation", 184-185 in Proceedings of the IXth International Congress of Phonetic Sciences 1979, Institute of Phonetics, University of Copenhagen.

Ohala, J J (1981): "The Listener as a Source of Sound Change", 178-203 in Masek, C S, Hendrick, R A and Miller, M F (eds): Papers from the Parasession on Language and Behavior, Chicago:Chicago Linguistic Society.

Ohala, J J and Feder, D (1986): "Speech Sound Identification Influenced by Adjacent "Restored" Phonemes", J Acoust Soc Am 80.S110.

Öhman, S (1966): "Coarticulation in VCV Utterances: Spectrographic Measurements", J Acoust Soc Am 39:151-168.

Öhman, S (1967): "Numerical Model of Coarticulation", J Acoust Soc Am 41:310-320.

Perkell, J and Klatt, D (1986): Invariance and Variability in Speech Processes, Hillsdale, N J:LEA.

Pollack, I and Pickett, J M (1964): "Intelligibility of Excerpts from Fluent Speech: Auditory vs Structural Context", J Verb Learn and Vert Beh 3:79-84.

Risberg, A (1979): Doctoral dissertation, RIT, Stockholm.

Schulman, R (forthcoming): "Articulatory Dynamics of Loud and Normal Speech", submitted to J Acoust Soc Am.

Stevens, K N and House A S (1963): "Perturbation of Vowel Articulations by Consonantal Context: An Acoustical Study", J Speech & Hearing Res 6:111-128.

Stevens K N and Blumstein S (1978): "Invariant Cues for Place of Articulation in Stop Consonants", J Acoust Soc Am 64, 1358-1368.

Stevens K N and Blumstein S (1981): "The Search for Invariant Correlates Phonetic Features", in Eimas, P and Miller J (eds): Perspectives on the Study of Speech, Hillsdale, N J:LEA.

Sundberg, J (1975): "Formant Technique in a Professional Singer", Acustica 32(2), 89-96.

Traunmüller, H (1981): "Perceptual Dimension of Openness in Vowels", J Acoust Soc Am 69, 1465-1475.

Warren, R (1970): "Perceptual Restoration of Missing Speech Sounds", Science 167, 392-393.

Westbury, J and Keating P (1980): "Central Representation of Vowel Duration", J Acoust Soc Am 67, Suppl 1, S37 (A).

STRESS - PRESSURE CHANGES OR LARYNGEAL ACTIVITY?

Ilkka Raimo
Dept. of Phonetics
University of Turku
SF-20500 Turku
FINLAND

Olli Aaltonen
Dept. of Phonetics
University of Turku

Erkki Vilkman
Phoniatric Dept.
Tampere University
Central Hospital

ABSTRACT

We have investigated the physiological background of sentence stress production in normal and whispered Finnish, and in simulated sentences produced with excised human larynges. In normal phonation the EMG activity of the cricothyroid muscle and subglottal pressure (esophageal pressure) registered from two subjects showed a clear association with stress. In whisper the contribution of the cricothyroid muscle was negligible. The same sentence used with the living subjects was simulated by means of excised larynges. The contours could be obtained either by changing subglottal pressure only or by laryngeal adjustments only. All in all, it seems that sentence stress is not produced by any single factor but by complex interactions of physiological subsystems.

INTRODUCTION

The change in fundamental frequency is generally assumed to signal to a listener the word which has been emphasized by the speaker [1]. However, there are different views about the physiological mechanisms underlying these fundamental frequency variations. Müller [2] observed that a change in fundamental frequency can be brought about by a change in subglottal air pressure or by a change in the tension of the vocal folds. This basic distinction has been widely studied with a variety of experiments, and evidence supporting one or the other has been presented.

The essential role of the cricothyroid muscle and rotation in cricothyroid articulation in fundamental frequency (F0) regulation is well accepted (see [3] for a review). EMG studies have shown that the activity of the cricothyroid muscle correlates with fundamental frequency peaks also in sentences (e.g. [4,5,6]). Ladefoged [7] has stated that an increase in the flow of air out the lungs results an increase in fundamental frequency signalling stress. However, the subglottal pressure has been rather ineffective in fundamental frequency regulation. Such ratios as 2-5 Hz/cmH_2O has been reported in chest register [8,9]. Monsen et al. [10] have concluded that both mechanisms are involved in the production of stress, and it may be that different languages make use of these two controlling mechanisms in a different manner.

The purpose of the present study was to study the effects of the laryngeal and subglottal mechanisms on the fundamental frequency variations connected with sentence stress in Finnish. In Experiment I we studied the EMG activity of two intrinsic laryngeal muscles and esophageal pressure. The measurements were made in two different conditions: in normal phonation and in whisper. Vocal fold vibration is avoided in whisper, and the periodic voice replaced by aperiodic noise. Even with no fundamental frequency it is still possible to identify a stressed word. In Experiment II we used excised human larynges to study independently the effects of laryngeal adjustments and subglottal pressure changes on fundamental frequency signalling stress in sentence. The use of excised larynges in voice physiological studies has a long traditions. Within certain limits excised human larynges have been noticed to produce vibrations comparable to those of living subjects [2,11]. The main limitation is that the action of the thyroarytenoid muscle cannot be simulated.

MATERIAL AND METHODS

Experiment I

Two healthy, male native speakers of Finnish participated voluntarily in the investigation.

For the acoustical analysis the speech samples were recorded using a high-quality tape recorder (Teac) and an acoustic microphone in normal room acoustics. Esophageal pressure records were obtained from an air-filled (2 ml) balloon sealed to a catheter. The balloon was passed through the nose. The catheter was connected to a pressure meter (Frökjaer-Jensen Manophone). The signal from the meter was tape recorded (Racal).

EMG activities were recorded from the cricothyroid (CT) and the thyroarytenoid (VOC) muscles for normal phonation and whisper using bipolar hooked-wire electrodes. The percutaneous and submucous method was applied.

Fundamental frequency (F0) (Frökjaer-Jensen Frequency Meter) and absolute values of intensity (Frökjaer-Jensen Intensity Meter) as well as the esophageal pressure were recorded on paper. The peak value of the first, stressed syllable for each word of the three-word test utterance was measured manually.

The EMG signals were full-wave rectified and averaged (n=5) using a computer-based (Hewlett Packard) EMG data processing program developed at the Department of Biology of Physical Activity at the University of Jyväskylä [12]. The line-up point for averaging was the onset of phonation of each utterance. The peak values of EMG activity were measured from the averaged data for each word of the test utterance.

The spoken material consisted of five versions of the sentence /nalle meni ma:lle/ ("Teddy went to the country") in which the stress was varied in the following way: 1) no extra stress, 2) the first word stressed,...5) all the words stressed. The subjects repeated each version five times both for normal speech and whisper.

See [13] for further details.

Experiment II

Two normal fresh larynges taken from autopsies of males were used.

In the dissection the extralaryngeal structures were removed. Also the epiglottis and ventricular folds were dissected.

The acoustical samples were tape recorded. Electroglottographic signals were recorded using a electroglottograph (Frökjaer-Jensen EG 830) and coin-shaped electrodes attached symmetrically to the thyroid cartilage with a screw (c.f. [14]). The subglottal pressure signals were obtained using a pressure meter (Frökjaer-Jensen Manophone) connected to an outlet in the subglottal space. The pressure signal was tape-recorded (Racal). The air flow was monitored visually by means of a lead pearl flow meter attached to the air intake.

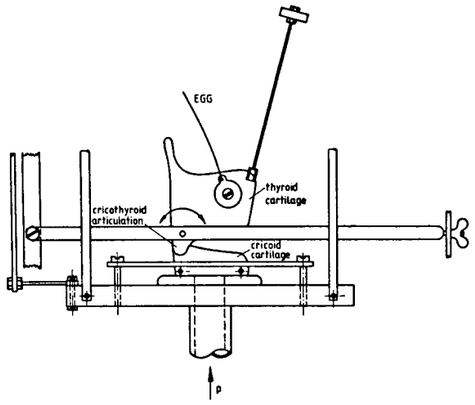


Figure 1. The apparatus used in Experiment II. EGG=electroglottograph, p=pressure. See [15] for further technical details.

The same sentences as in Experiment I were used as models and the F0 contours and the timing of this sentence were simulated to produce stress patterns 1-4.

The F0 contours were produced as follows: (1) by changing manually the laryngeal adjustments only. The flow was constant (approximately 300 ml/s). In this case the glottis was closed and opened manually and the intended stress was produced by rotating the thyroid cartilage; (2) by changing the subglottal pressure by regulating the air flow to the subglottal space. The thyroid cartilage was fixed. The glottis was closed using a forceps.

Perceptually the most natural (n=21) F0 contours were chosen by the authors and recorded on separate tape. The recordings were played to a panel of naive listeners. (n=20) whose task was to determine the position of the stress in each sentence. On the basis of this listening test the best complete sets (four sentence types) of both laryngeally and pressure induced stress patterns was formed for further analysis. The F0 (F-J Frequency Meter), intensity (F-J Intensity Meter) and subglottal pressure curves were recorded on plotting paper (Mingograph). The peak values of each variable were measured from the curves. The measuring points were the approximated site of the first stressed syllable of each word of the simulated three word sentence.

RESULTS

Experiment I

As expected, the stressed words differed acoustically from the non-stressed words in normal speech by prominent peaks in the F0 and intensity curves. In whisper stress was signalled by intensity. Each word purposely stressed by the speakers, was accompanied by a peak in the intensity curve. However, intensity did not show the same downward slope in whisper as in normal phonation. For instance, the last word in an utterance without any extra stress could exhibit the highest intensity peak.

The peak values of the esophageal pressure (p_{oes}) were nearly equal for both normal phonation and whisper. The most striking difference between them was that the differences between the peak values of the p_{oes} for the stressed and the non-stressed words were greater in whisper than in normal phonation. It is also worth noting that the pressure peaks increased towards the end of the utterances in whisper when all the words were stressed. The correlations of the p_{oes} with the acoustic variables of stress proved to be significant for both normal speech and whisper.

The cricothyroid muscle (CT) showed a peak of activation immediately preceding the peaks in normal speech. The correlations of the CT activity with F0, intensity and p_{oes} were also significant. In whisper the correlation of CT and intensity was non-significant. There were recordings in which the peaks in the CT muscle activity did not exceed the level of rest discharge for stressed words, even though they occasionally showed a phasal relationship with stress. In general, the average activation pattern of the cricothyroid muscle was lower and less variable in whisper than in normal speech.

Contrary to CT, the thyroarytenoid muscle (VOC) exhibited more similar stress patterns of the EMG activity for normal phonation and whisper. However, the two conditions differed from each other as to the correlations of the VOC-EMG with p_{oes} and intensity. The correlations between the VOC-EMG and intensity as well as between the VOC-EMG and p_{oes} were lower in whisper than in normal phonation.

The activity of the thyroarytenoid muscle also showed interindividual differences. There was a significant correlation between F0 and VOC-EMG only for one speaker. For both subjects the VOC-EMG correlated significantly with intensity. See [13] for further details.

Experiment II

The basic flow level in the laryngeally induced stress was lower than for the flow-induced stress. Consequently, intensity and subglottal pressure are higher in the latter case. Higher flow rate (approximately 500 ml/s) was needed for instance for getting an abrupt enough attack. In both cases, however, the phonation type represented chest register phonation and also the higher pressure values were within physiological limits (Fig. 2).

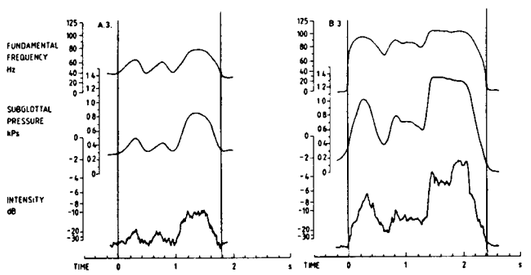


Figure 2. An example of the F0 contours produced by an excised larynx. The last word of the simulated sentence /nalle meni ma:lle/ has been emphasized. The left hand panel shows curves obtained by changing the laryngeal adjustments. The right hand panel illustrates results of changing the flow.

The differences between the peak values of the variables within each sentence for both experimental conditions were calculated. The changes in the fundamental frequency are greater in laryngeal stress than in pressure induced stress even if the subglottal pressure changes in the latter case were somewhat larger. The ratio of fundamental frequency and subglottal pressure change was 63.1 Hz/kPs for laryngeally induced stress and 26.6 Hz/kPs in the stress induced by air-flow changes.

The listeners perceived stress quite reliably irrespective of whether it was produced by changing the laryngeal adjustments only or by changing pressure only. The correlations between the peak values of F0, intensity and subglottal pressure were significant in both conditions. The intercorrelations were lower for laryngeally induced stress.

DISCUSSION

The peak values of the acoustical variables, and the cricothyroid muscle EMG activity exhibited a significant correlation. The thyroarytenoid muscle showed significant correlation with F0 only for one subject. This may be partly due to individual differences in strategies of producing stress and to technical difficulties in obtaining EMG data free of movement artifacts. The contribution of the thyroarytenoid muscle in stress production in normal speech has qualitatively been reported before (e.g. [4]). However, based on correlation analysis of the EMG data of one subject, Atkinson [6] stated that the thyroarytenoid muscle plays a minor role in controlling fundamental frequency in speech.

The minor role of the cricothyroid muscle in whisper is supported by an X-ray study of the larynx according to which the distance between the anterior tips of the cricoid and the thyroid cartilage (an estimate of the vocal fold length) was the same in whisper as in respiration [16]. The low level of cricothyroid activity in whisper when there is a need for higher effort level for stress production is comprehensible when the function of this muscle is considered: the vocal folds are abducted by cricothyroid muscle twitch [17].

Warren [18] observed higher intraoral pressures in whisper than in normal phonation. If the esophageal pressure used in the present study is a gross estimate of tracheal pressure, then our esophageal pressure peak values were only occasionally higher in whisper than in normal speech. Still, the peak values of the subglottal pressure in whispered stress production are high enough to elicit vibrations of the vocal folds in normal phonation. Biomechanically the increased thyroarytenoid muscle activity in whisper may cause the necessary extra medial compression or internal stiffening needed to hinder the vocal folds from vibrating during whispered stress production.

The correlation of the intensity of whisper and the EMG activity of the thyroarytenoid muscle was positive but nonsignificant. VOC EMG maxima, however, often appeared in the vicinity of the whispered stressed word. The correlation of the peak values of the esophageal pressure with the intensity of whisper was significant. It can be assumed that the thyroarytenoid muscle activity together with other laryngeal muscles (e.g. the lateral cricoarytenoid) and the actively increased subglottal pressure reflect the extra effort needed to cause perceivable stress in whisper. Correspondingly the low subglottal pressure values for unstressed words, as compared to normal phonation, may reflect low glottal resistance and respiratory activity.

The results of the experiments with the excised larynges are well in line with earlier findings. The increase in glottal resistance due to cricothyroid muscle twitch has been reported before in living subjects [19] and with excised larynges (e.g. [15]). The ratio of fundamental frequency and subglottal pressure change was also in the limits of earlier findings concerning chest voice phonation (e.g. [9]). However, the mechanism through which the subglottal pressure affects the fundamental frequency is not clear. It has been set forth that even in this case the reason would be laryngeal [20].

From the phonetic point of view the most important result is that perceivable sentence stress can be produced with excised larynges both by means of laryngeal and pressure adjustments. It has been suggested that the results of studies on voice production should be interpreted in terms of systems physiology because there are considerable interactions between the contributing subcomponents (e.g. [21]). All in all it seems that the dichotomic question: stress - vocal fold tension or subglottal pressure? is not justified on the basis of the present study.

REFERENCES

[1] Lehiste I: Suprasegmentals. MIT Press, Cambridge 1970

[2] Müller J: Handbuch der Physiologie des Menschen.Teil I. Hölscher, Coblenz 1837

[3] Hollien H, Hicks J W: Mechanisms for the control of vocal frequency. In: Speech Communication Papers Presented at the 97th meeting of Acoust Soc Am 12-16 June 1979. pp. 97-100. Eds. J J Wolf, D H Klatt. MIT Press, Cambridge 1979

[4] Hirano M, Ohala J, Vennard W: The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation. J Speech Hear Res 12: 616-628, 1969

[5] Collier R: Physiological correlates of intonation patterns. J Acoust Soc Am 58: 249-255, 1975
[6] Atkinson J: Correlation analysis of the physiological factors controlling fundamental voicefrequency. J Acoust Soc Am 63:211-222, 1978
[7] Ladefoged P: A course in phonetics. Harcourt, Brace, Jovanovitch, New York 1975
[8] Hixon T J, Klatt D H, Mead J: Influence of forced transglottal pressure on fundamental frequency. J Acoust Soc Am 49: 105, 1971
[9] Baer T: Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes. J Acoust Soc Am 65: 1271-1275, 1979
[10] Monsen R B, Engebretson A M, Vemula N R: Indirect assessment of the contribution of subglottal pressure and vocal-fold tension to changes of fundamental frequency in English. J Acoust Soc Am 64: 65-80, 1978
[11] van den Berg J W, Tan T S: Results of experiments with human larynges. Pract Oto-rhino-laryng 21: 425-450, 1959
[12] Viitasalo J. Komi PV: Signal characteristics of EMG with special reference to reproducibility of measurements. Acta Physiol Scand 93: 531-539, 1975
[13] Vilkman E, Aaltonen O, Raimo I, Ignatius J, Komi PV: On stress production in whispered Finnish. J Phonetics. In press.
[14] Lecluse FLE: Elektroglottografie. Dissertation. Erasmus University, Rotterdam 1977
[15] Vilkman E: An apparatus for studying the role of the cricothyroid articulation in voice production of excised human larynges. Folia Phoniat. In press.
[16] Fink B, Demarest R: Laryngeal biomechanics. Harvard University Press, Cambridge 1978
[17] Arnold GE: Physiology and pathology of the cricothyroid muscle. Laryngoscope 71: 687-753, 1961
[18] Warren D W: Aerodynamics of speech production. In: Contemporary issues in experimental phonetics. pp. 105-137. Ed. N J Lass. Academic Press, New York 1976
[19] Yanagihara N, von Leden H: The cricothyroid muscle during phonation. Ann Otol Rhinol Laryngol 75: 987-1007, 1968
[20] Titze I, Talkin D: A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. J Acoust Soc Am 66: 60-74, 1979
[21] Müller E M, Abbs J H, Kennedy J G: Some systems physiology considerations for vocal control. In: Vocal fold physiology. pp. 209-227. Ed. K N Stevens, M Hirano. University of Tokyo Press, Tokyo 1981

ACKNOWLEDGEMENTS

The study was financially supported by a grant from the Kordelin Foundation. The authors wish to express their gratitude to Professor Jaakko Lehtonen, Professor Paavo V. Komi and MD Jaakko Ignatius for technical and practical support.

STUDIES OF GLOTTAL EXCITATION USING INVERSE FILTERING AND AN ELECTROGLOTTOGRAPH

Melvyn J. Hunt

National Research Council of Canada
National Aeronautical Establishment
Building U61, Montreal Road
Ottawa, K1A OR6, Canada

Abstract

Glottal excitation has been studied in steady vowels produced by three subjects, two male and one female. Electroglottograph waveforms are shown together with simultaneous glottal airflow waveforms and waveforms for individual formants, both derived by interactive inverse filtering. Modal voice shows formant excitation concentrated on the instant of closure. Falsetto voice shows a triangular or sinusoidal airflow waveform. Breathy voice shows appreciable formant excitation both on closure and at the centre of the open phase. Creaky voice shows appreciable excitation at the start of the open phase as well as its end, and there is often an alternation in spectral content of the excitation from cycle to cycle causing the relative intensities of formants to vary. In examples of extreme creak the airflow waveforms are complex and difficult to interpret, but they are similar to the electroglottograph waveforms.

1. Introduction

The purpose of this paper is to present the results of a study of several modes of phonation. Glottal airflow waveforms derived by inverse filtering are compared with a waveform indicating the area of contact of the vocal folds. By leaving one formant uncanceled in the inverse filtering, the excitation of individual formants is also shown.

2. Method

Subjects spoke in an anechoic chamber, and two-channel recordings were made on a Revox A77 tape recorder of the output from a B&K condenser microphone and a laryngograph (electroglottograph) [1]. The laryngograph measures r.f. impedance across the larynx and hence the area of contact of the vocal folds [2]. The two signals were simultaneously digitized at 20kHz, sharply low-pass filtered at 5kHz without phase distortion and downsampled to 10kHz. Low-frequency phase distortions introduced by the recording process were removed from both signals by an automatic method [3].

The effect on the speech signal of radiation from the lips, which corresponds approximately to differentiation, was countered by integrating the signal. The filtering effect of the vocal tract on the glottal airflow was removed by interactive inverse filtering [4]. In this process, the formants in the speech signal are canceled by an equal number of antiresonances (typically five), whose frequencies and bandwidths are intended to match those of the corresponding formant exactly. The interactive system allows the parameters of one antiresonance to be adjusted at a time. The frequency and bandwidth of this antiresonance are determined by an a/d converter that frequently samples the values of two adjustable potentiometers. The speech signal filtered by the fixed and varying antiresonances is displayed on a graphics screen (DEC VT11). When fewer than a thousand samples are being plotted, the filter parameter update, the filtering, and the replotting can all be carried out within 50ms, giving the user the impression of a continuously changing display as he turns the knobs connected to the potentiometers. By adjusting the antiresonance corresponding to each formant in turn, the effect of the vocal tract can be tuned out in about one minute.

In modal voice at least, the spectrum of the glottal airflow waveform falls off at roughly 12dB per octave. Consequently, if the airflow is differentiated twice, the spectrum is flattened and the main feature in the waveform is an impulse at the instant of glottal closure. It is predominantly this impulse that excites the formants, which in this representation have roughly equal amplitude. When individual formants are shown in this paper, they are shown in a signal corresponding to the doubly differentiated airflow (or, equivalently, the singly differentiated speech signal).

3. Speech Material

Three speakers, two male (AF and MH) and one female (EA), each produced examples of steady unnasalized vowels with modal, falsetto, breathy, and creaky voice quality. Examples of several vowels were provided, but analysis has been concentrated largely on the schwa, since this vowel has formants that are well separated from each other and the first formant is fairly high in frequency and thus well separated from the fundamental frequency (F_0).

4. Discussion of the Validity of the Method

It is sometimes alleged that inverse filtering is a subjective process in which the user obtains the airflow waveform he expects. There may be some truth to the allegation as far as fine detail in the flow is concerned, but we believe that the gross shape of the flow waveform corresponds to the actual flow. The criterion normally said to be used to determine the antiresonance parameters is the flatness of the closed-glottis phase, since there is by definition no airflow during this period. This criterion is not useful for phonation modes in which the glottal closure is very brief. With the interactive method used here, however, it is still possible to estimate the parameters because a continuous range of parameters can be surveyed quickly and it becomes evident that only one set of values results in a maximally simple airflow. If the parameters of any one antiresonance are disturbed, the corresponding formant will appear in the waveform.

Further support for the validity of the method comes from the following considerations. Pairs of individuals working independently choose substantially the same sets of parameters and thus obtain the same airflow patterns. Modal-voice airflow waveforms from two recording sessions with the same speaker look similar. When there is an identifiable closed-glottis phase of suitable length, covariance-method linear predictive coding (LPC) [5] derives a very similar set of parameters and hence the same airflow waveform as an operator using the interactive method. In tests with synthetic speech, it is possible to recover the known excitation waveform exactly. The surprisingly complex airflow waveforms in the extreme creak illustrated in the next section show a strong similarity to their corresponding laryngograph waveforms. Finally, distinguishing features of the different phonation types are consistent across speakers.

5. Results

In the figures the laryngograph waveform is labeled as L_X, the glottal airflow as F_G, the doubly differentiated flow as F_G'', and the n'th formant as F_n. Unless otherwise indicated in the figure cap-

tions, the vowel is a schwa and the duration of the waveforms is 20ms. The waveforms derived from the speech signal are delayed relative to the laryngograph signal by about 0.5ms because of the sound propagation time to the microphone.

i) Modal Voice

The waveforms for the two male speakers (Figs 1,2) showed a prolonged closed phase, there was a clear impulse at closure in F_G'', and formant excitation was concentrated at this point. The female speaker (Fig 3) showed a much shorter closed phase, F_G'' was more complex, but formant excitation still appeared to be concentrated on the instant of closure.

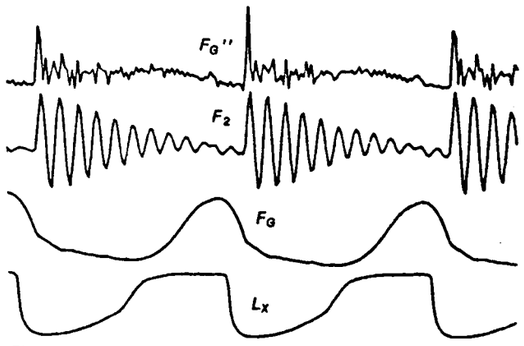


Fig 1. MH modal voice. Excitation at instant of closure, appearing as an impulse in F_G''.

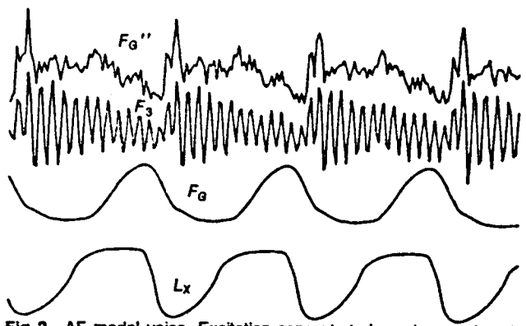


Fig 2. AF modal voice. Excitation concentrated on closure, though less clearly than for MH.

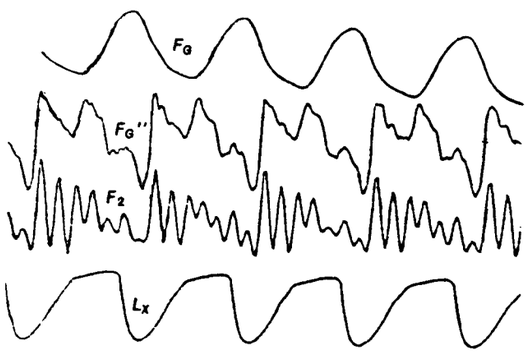


Fig 3. EA modal voice. F_G'' not impulse-like.

ii) Falsetto

For all three speakers (Figs 4,5,6) F_G has a triangular or sinusoidal waveform consistent with an excitation spectrum falling off with frequency more quickly than in modal voice. Any closed phase is very short. F_G'' is reminiscent of the modal voice example from the female speaker, EA, suggesting that its form may be due to a high F_0 rather than to a property of falsetto voice.

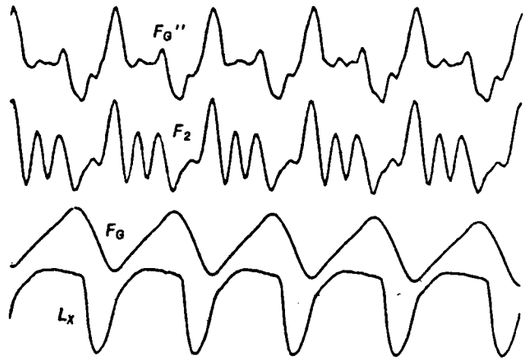


Fig 4. MH falsetto. F_G is triangular. F_G'' is not impulse-like.

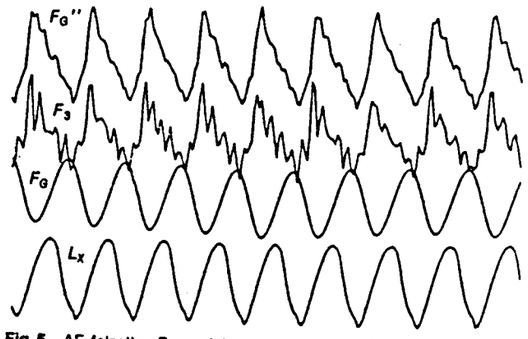


Fig 5. AF falsetto. F_G and L_X are sinusoidal. F_G'' is not impulse-like.

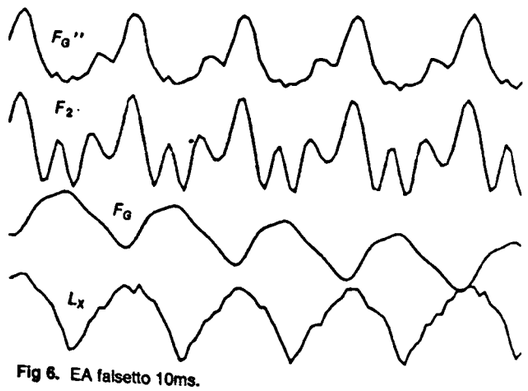


Fig 6. EA falsetto 10ms.

iii) Breathy Voice

F_G and L_X show brief closure or partial closure. In F_G'' the instant of closure is less impulse-like than in modal voice. Single-formant plots for the two male speakers (Figs 7,8) indicate that excitation is strongest at two instants: at closure and at the point of maximum airflow in the centre of the open phase. Excitation for the female speaker looks more noise-like (Fig 9), though it may peak on closure and at maximum flow.

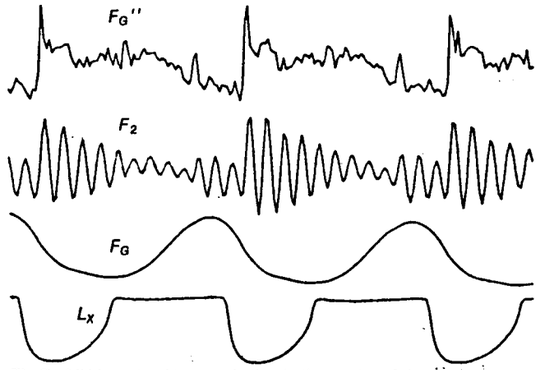


Fig 7. MH breathy. Note excitation in the centre of the open phase.

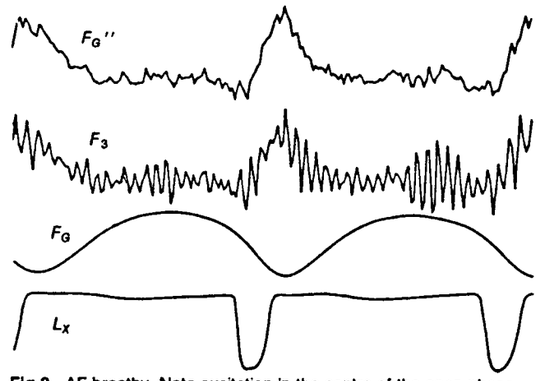


Fig 8. AF breathy. Note excitation in the centre of the open phase.

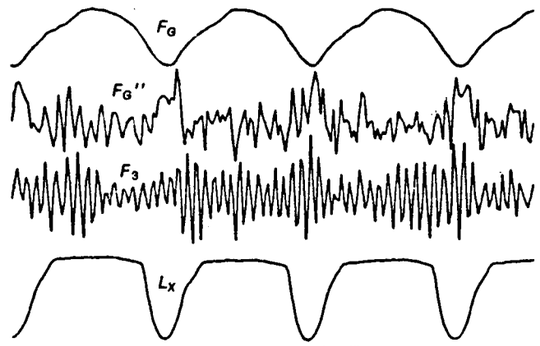


Fig 9. EA breathy. Excitation seems noise-like.

iv) Creaky Voice and Creak

Subjects produced a range of phonation types ranging from a creaky voice whose F_G was superficially similar in form to that for modal voice to a full creak with a glottal cycle so long that individual pulses are perceived rather than F_0 and with F_G and L_X waveforms that are complex and difficult to interpret. MH produced exclusively creaky voice, EA exclusively creak, and AF examples of both extremes as well as intermediate examples. In all cases, the waveforms varied more from cycle to cycle than in other phonation modes.

Creaky voice often showed alternation between two glottal cycles, the cycles differing in their airflow waveform as well as in their duration. In one of the two cycles there was often strong excitation at the point of glottal *opening* (Fig 10). The details of closure generally differed in the two cycles, leading to differences in the excitation spectrum and hence to the relative intensities of formants. In examples of creaky [u] from MH and AF (Figs 11,12) there were series of cycles in which a strong F_1 and a weak F_2 alternated with a weak F_1 and a strong F_2. Since the formant waveform from a previous excitation has not decayed to a negligible value when the next excitation comes along, there will be an interaction between the two waveforms that depends on their relative phases and hence on the formant frequency. This interaction could conceivably account for the observed weak/strong alternation. However, when a single excised piece of flow waveform containing just one excitation event is used to excite a resonance corresponding to F_1 or F_2 the effect is still seen. It seems, then, that the alternation in formant intensities really is due to an alternation in the spectral content of the glottal airflow.

The schwa examples from AF and EA (Figs 13,14) are complex, extreme creak. As noted in Section 4, the L_X and F_G waveforms are similar. The F_G'' waveform shows periods in which an impulse is followed by a long phase of little activity. Sets of formants excited in this period show prolonged steady exponential decays (Fig 15), which offer an opportunity of accurate formant bandwidth determination. They seem to indicate bandwidths much narrower than values normally assumed.

Examples of [i] and [a] vowels from AF (Figs 16,17) are of complexity intermediate between the creaky [u] and the extreme creak in the schwa. They perhaps offer a means of bridging the gap between the two extremes and thus interpreting the extreme creak waveforms.

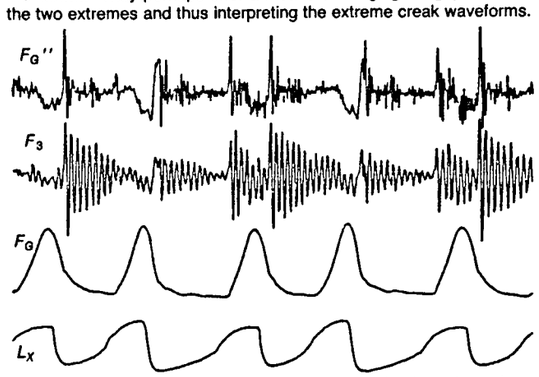


Fig 10. MH creaky 40ms. Excitation on opening in alternate cycles.

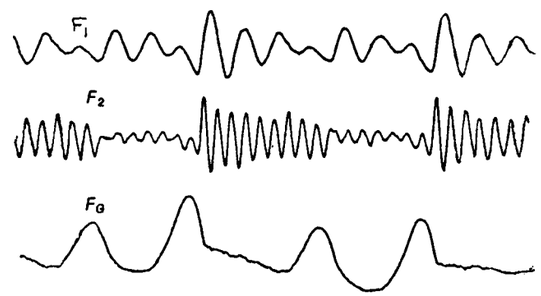


Fig 11. MH creaky [u] 40ms. F_1 and F_2 alternate in intensity.

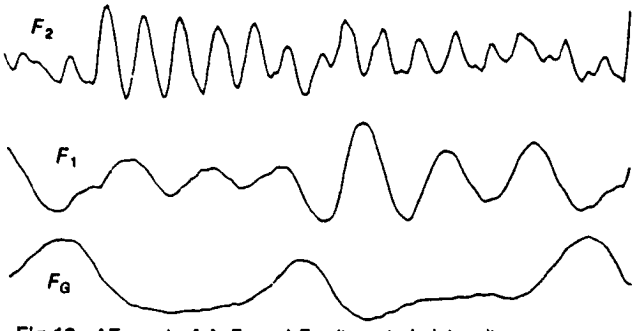


Fig 12. AF creaky [u]. F_1 and F_2 alternate in intensity.

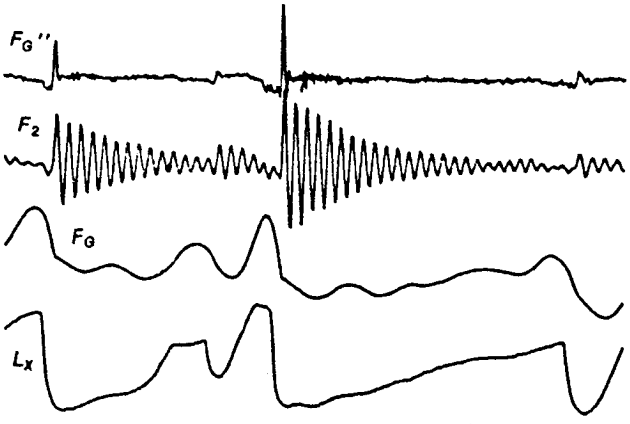


Fig 13. AF creak 40ms. Complex airflow parallels L_X.

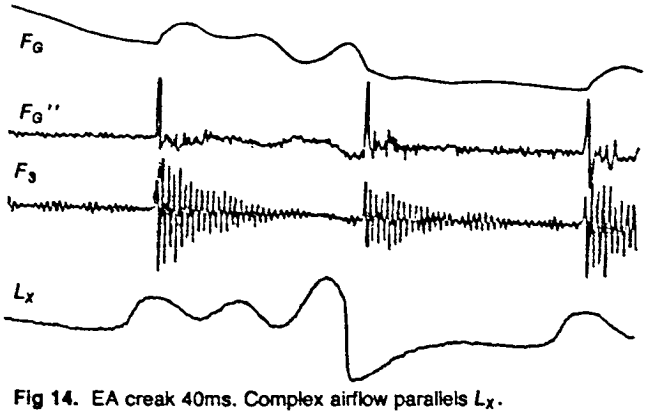


Fig 14. EA creak 40ms. Complex airflow parallels L_X.

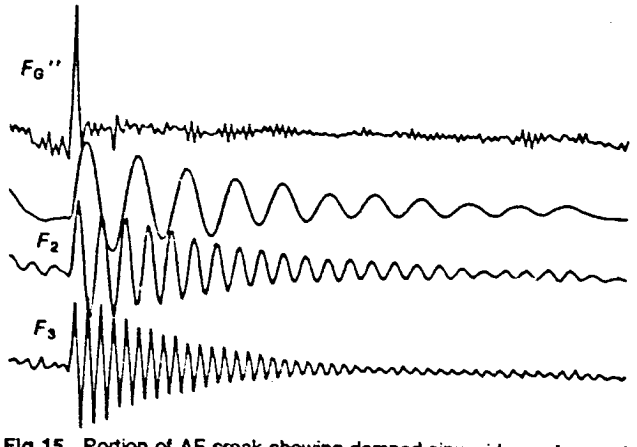


Fig 15. Portion of AF creak showing damped sinusoid waveforms of first three formants.

Fig 16. AF creaky [i] 40ms. F_G only.

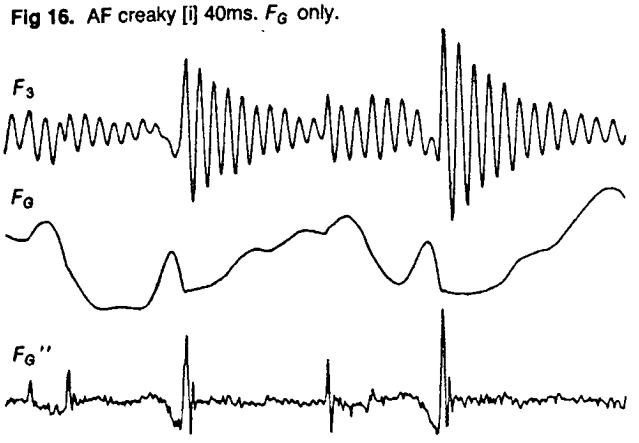


Fig 17. AF [a] creak.

6. Discussion of Results

Justifications of LPC usually assume that excitation of voiced speech is concentrated on the instant of glottal closure and that the excitation waveform approximates an impulse for speech preemphasized at 6dB/octave. The results presented here suggest that these assumptions may be valid only for modal voice, perhaps only for *male* modal voice. Pitch-synchronous LPC [6,7], in which a negative-going impulse in the L_X waveform is used to detect closure and hence excitation, would be likely to encounter difficulties with creaky or breathy phonations having excitation on opening or in the middle of the open phase. Tests of pitch-synchronous LPC with the speech discussed here support this expectation.

Some algorithms for determination of F_0 detect excitation points, while others detect repetitions of patterns in the waveform or, equivalently, harmonic structure in the spectrum. These two approaches would give different results from each other with breathy or creaky phonations. Moreover, neither approach would consistently indicate the frequency of opening and closing of the glottis.

Acknowledgments

The work reported here was carried out mainly while I was employed at the U.K. Joint Speech Research Unit. I am grateful to John Bridle and to Prof. Adrian Fourcin for their help.

References

1. A.J. Fourcin & E. Abberton, "First applications of a new laryngograph," *Medical and Biological Illustration*, vol. 21, pp. 172-182.
2. D.G. Childers, D.M. Hicks, G.P. Moore & Y.A. Alsaka, " A model for vocal fold vibratory motion, contact area, and the electroglottogram," *J. Acoust. Acoust. Soc. Am.*, pp. 1309-1319, Vol 80 No. 5, 1986.
3. M.J. Hunt, "Automatic Correction of Low-frequency Phase Distortion in Analogue Magnetic Recordings," *Acoustics Letters*, Vol.2, pp.6-10, 1978.
4. M.J. Hunt, J.S. Bridle & J.N. Holmes, "Interactive Digital Inverse Filtering and its Relation to Linear Prediction Methods," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tulsa OK, May 1978, pp.15-18.
5. J.D Markel & A.H. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976.
6. A.K. Krishnamurthy & D.G. Childers, "Two-channel Speech Analysis," *IEEE Trans. Acoust, Speech Signal Processing* Vol 34, pp.730-743, 1986.
7. M.J. Hunt & C.E. Harvenberg, "Generation of Controlled Speech Stimuli by Pitch-Synchronous LPC Analysis of Natural Utterances," paper A4-2, Vol 1, *Proc. Int. Congress on Acoustics*, Toronto, Canada, July 1986.

AN ELECTROMYOGRAPHIC INVESTIGATION OF LARYNGEAL MUSCLE ACTIVITY IN MODERN STANDARD CHINESE TONES

Laurent SAGART — Centre de Rech. Ling. sur l'Asie Orientale, CNRS, EHESS., Paris, FRANCE

Pierre HALLE — Labo. de Psychologie expérimentale, CNRS, EHESS., Paris, FRANCE

Bénédicte de BOYSSON-BARDIES — Labo. de Psychologie expérimentale, CNRS, EHESS., Paris, FRANCE

ABSTRACT

The activity patterns of the cricothyroid and sternohyoid muscles in Modern Standard Chinese tones were electromyographically investigated in two subjects. Average activity profiles by tone and by segmental syllable indicate cricothyroid activity is well correlated to Fo with a latency time of 80-100 ms. The sternohyoid participates both in Fo lowering and segmental articulation, with strong activity peaks preceding consonant release when the following vowel is back and/or low.

INTRODUCTION

The present study is part of an ongoing electromyographic (EMG) investigation <14> of modern standard Chinese (MSC) tones. In terms of the pitch contours which characterize them, the four lexical tones of MSC may be roughly described as: tone 1 (T1): high level, tone 2 (T2): mid to high rising, tone 3 (T3): mid-low to low falling, tone 4 (T4): high to low falling. for detailed acoustic descriptions see <7,10,11>. The activity patterns of two laryngeal muscles: the Cricothyroid (CT), the main regulator of vocal fold tension, and the Sternohyoid (SH), an extrinsic laryngeal muscle shown by various EMG studies <13,5> to be active in Fo falls and low Fo, but also believed to be involved in segmental articulation, were investigated. The subjects were two female students in linguistics in their late twenties, both native speakers of MSC. Subject CYC came from Taipeh and subject FJQ from Beijing. The corpuses consisted of MSC syllables in all tones placed in a carrier sentence ([wo3 niæn4 X ts↓4] "I read the character X") to avoid contamination by non-speech muscular activity. The target syllables were meaningful MSC words belonging to minimal series having the same segmentals, with lexical items in the four tones. The same minimal series were used with both subjects, and additional material was also used for subject FJQ.

METHOD

Two thin hooked platinum wire electrodes were percutaneously inserted into both subjects' CT and SH muscles according to a technique described in <3>. The subjects were then made to perform various manoeuvers (opening jaw against opposing force, swallowing, holding breath) to check on the insertion of the electrodes in the desired muscles. As the subjects read the corpuses aloud in a sound-treated room, the EMG signal from the electrodes and the acoustic signal were simultaneously recorded by means of a 7-track AMPEX recorder. Due to displacement of some electrodes after the checking manoeuvers, the signal from subject CYC's SH muscle proved unpossible to interpret. As a result, only signals from subject FJQ's CT and SH and subject CYC's CT could be analyzed.

The audio and EMG signals were digitized at 8 KHz (after low-pass filtering at 3.5 KHz and 6 dB/ octave analog preemphasis for the audio signal) and stored on disk in a Solar 16-40 computer. Fo was extracted by means of a cepstral method with framelength set to 312 ms. and frame period set to 10 ms. The EMG signal was undersampled to 1 KHz, and the absolute values were then integrated over a 75 ms. Hamming window sliding by 4 ms. steps. Programs were designed for the purpose of displaying the audio and EMG signals together with the Fo and integrated EMG curves. The tone-carrying part, consisting of the main vowel and any segment following it in the same syllable <8> was visually identified on the audio tracing. Based on a technique created by Kratochvil <9> for obtaining average Fo and Ao profiles for tones, Fo and integrated EMG were measured by hand in twelve regularly spaced points of time (numbered -2 to 9) for each target-syllable, points 1 and 6 corresponding to the onset and endpoint respectively of the tone-carrying part as determined on the audio tracing, with intervals of 20% of the duration of the tone-carrying part between any two adjacent points. These 24 measurements summarized the evolution of the EMG and Fo curves for each tone-carrying part and the margins on both sides of it. Mean values and standard deviations were calculated for each of these 24 points to obtain average Fo and EMG profiles by tone and by segmental syllable.

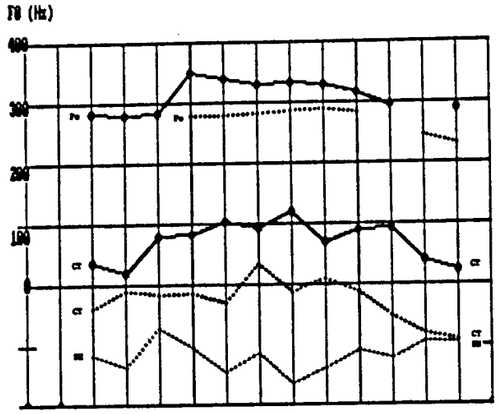


Fig. 1.a: tone 1.

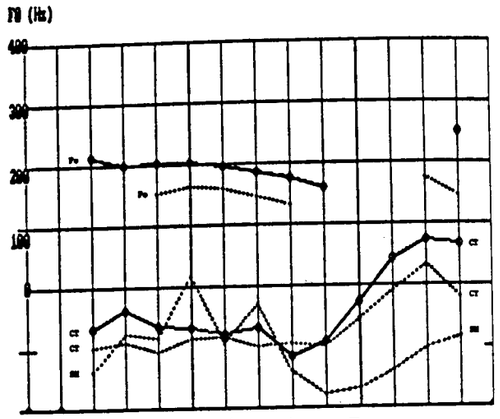


Fig. 1.c: tone 3.

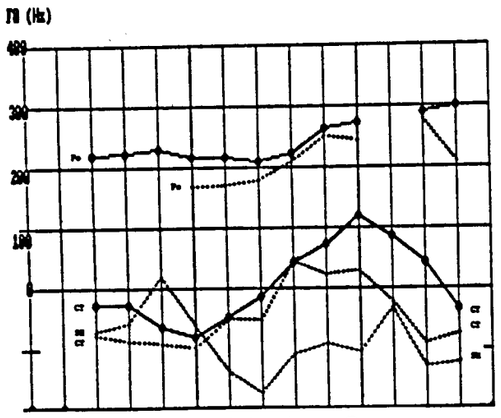


Fig. 1.b: tone 2.

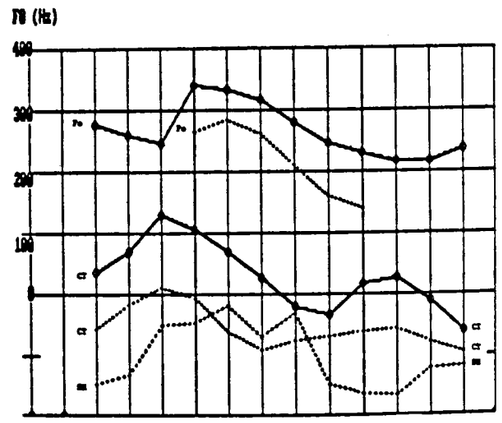


Fig. 1.d: tone 4.

Figs. 1.a-d: average Fo, cricothyroid and sternohyoid activity profiles by tone. Each curve averages Fo or muscle activity in five test syllables with segmentals [ma, ?ai, u, tu, thu]. Solid line: subject CYC. Dotted line: subject FJQ. Thick vertical lines indicate onset (left) and endpoint (right) of tone-carrying part.

RESULTS AND DISCUSSION.
Figs. 1.a-d show Fo and muscle activity profiles for each tone. With speaker FJQ, the CT and SH profiles were obtained from the same tokens.

The cricothyroid: similar evolutions of the CT are observed with both subjects: Fo rises (second part of T2, and any rises preceding the onsets of T1 and T4) are preceded by increases in CT activity. High CT activity also accompanies T1, a tone with high Fo throughout. In contrast, portions of tones characterized by Fo falls (T3, T4) are preceded by decreases in CT activity.

Latency time, the interval between muscle activity and Fo response, is best estimated by cross-correlation methods but a rough estimate can be arrived at by measuring the interval between remarkable points on the integrated EMG curve and the corresponding points on the Fo curve. For both subjects, latency times thus measured range between 50 and 160 ms, with values most frequently situated in the 80-100 ms range. This is in agreement with the finding in <4> of a mean latency time of 94 ms. for this muscle. In Figs.1.a-d, these values correspond to between one and two times the interval between two adjacent points. Accordingly, patterns of CT activity occurring on points -1 to 1 are relevant to the control of Fo at the beginning of the tone-carrying part of the test syllable, but patterns occurring after point 4 cannot relate to the test syllable. Note the increase in CT activity in the vicinity of point 6 in figs 1. a-d. This, we believe, relates to the production of T4 in the following syllable, "ts˧4".
Average profiles by segmental syllable indicate no clear effect of segmentals on the level of CT activity (although syllable structure may affect the location of CT peaks in relation to segmental events). In particular we do not observe the correlation between vowel timbre and peak level of CT which Auteserre & al. <2> suggested might account for the intrinsic frequency of vowels.
This overall pattern is stable across utterances, and also across repetitions of the same utterance. It characterizes not only the test syllables but also the carrier sentences. It corresponds well to the activity pattern of the CT as described for other tonal or non-tonal languages: Swedish <6>, Dutch <4>, Thai <5>, French <2> etc. Typical examples of CT activity in the 4 tones are shown in figs. 2.a-d.

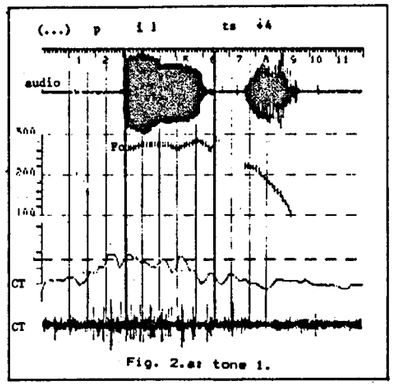


Fig. 2.a: tone 1.

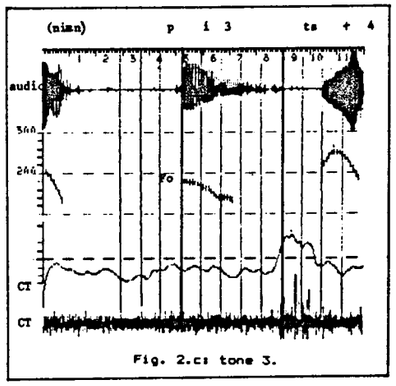


Fig. 2.c: tone 3.

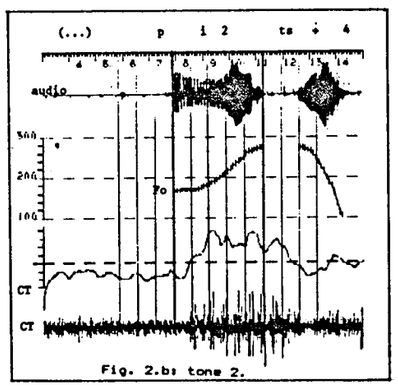


Fig. 2.b: tone 2.

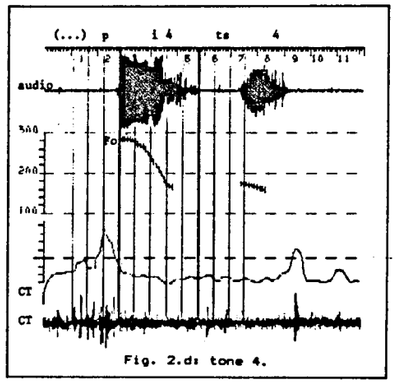


Fig. 2.d: tone 4.

Figs 2. a-d: Fo and cricothyroid activity in syllable [pi] in the 4 tones (subject FJQ). Test syllables are inserted in carrier sentence: [wo3 niæn4 X ts˧4] "I read the character X". From top to bottom: audio signal, Fo, integrated EMG curve and raw EMG signal.

The sternohyoid. In spite of very high standard deviation values, the average SH profiles by tone in figs. 1.a-d show some correlation between low or falling Fo and increased SH activity, in particular in T3 and T4. Conversely, low SH activity accompanies T1, the high level tone, and T2, the rising tone. However, the Fo shoulder at the end of T2 is often preceded by an SH peak.
All profiles regardless of tone also display an increase between points -2 and 0, corresponding to sharp activity peaks

shortly before vowel onset in individual test syllables. This suggests that part of the activity of the SH is unrelated to pitch control. Average profiles by segmental syllable (Fig.3) indicate the level of activity around vowel onset depends on the nature of segmental material, vowel timbre in particular: highest levels occur with [a], lowest with [i], intermediate levels with [u]. These observations of the activity pattern of the SH are consistent with published accounts of its role in speech. Regarding Fo control, Ohala <12> claimed the strap muscles, among which the SH, lower Fo indirectly by lowering the larynx, which in turn reduces the vertical, not the antero-posterior, tension in the vocal folds. Regarding segmental articulation, Ohala and Hirose <13> claimed that the SH also participates in tongue-lowering, tongue-backing and jaw-opening gestures by fixing or lowering the hyoid bone when muscles linking the hyoid bone and structures above it are also contracting (as the anterior belly of the digastric in jaw-opening gestures). Fo lowering will occur only if the hyoid bone and the larynx are free to move downward, that is, if the hyoid bone is not simultaneously pulled upward by muscles above it.

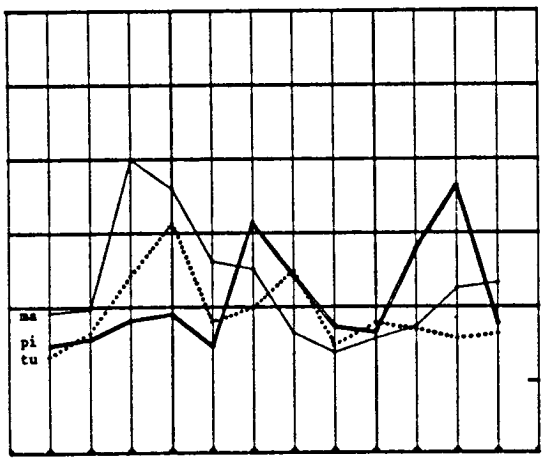


Fig.3: Average sternohyoid activity profiles by segmental syllable (subject FJQ). Each curve averages SH activity in four test syllables with T1, T2, T3 and T4. Simple line: syllable [ma]. Dotted line: syllable [tu]. Double line: syllable [pi].

CONCLUSION

Involvement of the SH in both Fo control and segmental articulation is not the only source of variability in our data: SH activity patterns can differ widely in repetitions of the same utterance. While instability of muscle activity patterns is a normal result in EMG studies, the stability of CT patterns is more remarkable. Two reasons may be invoked to account for it: (a) the CT specializes in Fo control and is not simultaneously involved in other tasks, and (b) although other muscles (among which the SH) play a role in Fo control, none is so efficient in regulating vocal cord tension.

ACKNOWLEDGEMENTS

The EMG recordings were made at INSERM (U3), Hôpital de la Salpétrière, Paris, under the supervision of Dr. Chevrie-Muller. The electrodes were inserted by Dr. Lacau Saint-Guily and the signal-collecting apparatus was installed by Mr. B. Maton. The authors gratefully acknowledge the assistance received.

REFERENCES

<1> Atkinson, J.E. (1973) Physiological factors controlling Fo: results of a correlation analysis. Abstract in JASA 54, 1.319 (A).

<2> Auteserre, D., A. Di Cristo & D. Hirst (1986) Approche physiologique des intonations de base du francais: cricothyroidien et fréquence fondamentale. Proceedings of 15èmes Journées d'étude de la parole (Aix-en-Provence, 1986), pp.37-41.

<3> Basmadjian, J.V. and G.A. Stecko (1962) A new bipolar indwelling electrode for electromyography. Journal of applied physiology, 17, p. 859.

<4> Collier , R. (1975) Physiological correlates of intonation patterns. JASA 58,1:249-255.

<5> Erickson, D. (1976) A physiological analysis of the tones of Thai. U. of Connecticut PhD.

<6> Gårding, E., O. Fujimura & H. Hirose (1970) Laryngeal control of Swedish word tones. Ann. Bull. RILP, 4:45-54.

<7> Howie, J.M. (1976) Acoustical studies of Mandarin vowels and tones. Cambridge: Cambridge University Press.

<8> Howie, J.M. (1974) On the domain of tone in Mandarin. Phonetica 30:129-148.

<9> Kratochvil, P. (1981) Tone and stress discrimination in normal Peking dialect speech. Proceedings of the second conference on Chinese Language Use. Canberra 1981.

<10> Kratochvil, P. (1984) Phonetic tone sandhi in Beijing dialect stage speech. Cahiers de Linguistique Asie Orientale XIII, 2:135-174.

<11> Kratochvil, P. (1985) Variable norms of tones in Beijing prosody. Cahiers de Linguistique Asie Orientale XIV, 2:153-174. Paris.

<12> Ohala, J.(1972) How is pitch lowered? Abstract in JASA 52,1 (part 1).

<13> Ohala, J. & H. Hirose (1970) The function of the sternohyoid muscle in speech. Ann. Bull. RILP 4:41-44.

<14> Sagart, L., P. Hallé, B. de Boysson-Bardies et C. Arabia-Guidet (1986) Tone production in Modern Standard Chinese: an electromyographic investigation. Cahiers de Linguistique Asie Orientale, XV, 2:205-221. Paris.

VELAR AND GLOTTAL ACTIVITY IN A SPEAKER OF ICELANDIC

PHILIP HOOLE

Institut für Phonetik
Ludwig-Maximilians-Universität
Schellingstraße 3
D-8000 Munich 40

ABSTRACT

The realization of Icelandic preaspirated plosives and voiceless nasals is examined, particularly the timing and extent of the devoicing gesture and of velar excursion. The segmental status of preaspiration is discussed and it is suggested that the voiceless nasals can be quite simply regarded as "ordinary" nasals coarticulated with a following preaspirated plosive.

INTRODUCTION

In this paper we examine an aspect of Icelandic that is particularly interesting from the point of view of interarticulatory timing - the realisation of the preaspirated plosives and voiceless nasals (henceforth abbreviated to HC and NH). These sounds have also been of central interest in the phonological analysis of Icelandic.
They are both subject to considerable restrictions on where they can occur in the word: HC occurs only after short vowel medially or finally ("seppi", "löpp") and in combination with other consonants only before /l/ and /n/ ("epli"). NH is subject to dialectal variation but in southern Icelandic occurs principally before /p, t, k/ following short vowel, i.e in the position in which HC also occurs.
The emergence of these two groups of sounds has also been linked diachronically /8/ through a shift in the timing of the devoicing gesture on the plosive. Synchronic analyses have varied, however. Pétursson /7/, in a phonemic description, analyzes preaspiration as the phoneme /h/ based on his phonetic investigations; and based on the existence of minimal pairs such as "dempi" vs. "dembi" ([tempi] vs. [tempi]) assumes the existence of separate voiceless and voiced nasal phonemes. There is thus no apparent relationship between these two groups of sounds. In other analyses /1/ this relationship is preserved, preaspiration being introduced by rule as an auto-segment, and modifying the adjacent nasal if one is present.
The questions just briefly touched on provided the motivation for examining these sounds in as homogeneous an environment as possible. This also gave the opportunity to return to a question raised in our earlier /4/ investigation of German where varying oral articulations in word-pairs such as "fette" vs. "feste" seemed to be combined with a constantly timed glottal gesture when viewed from the onset of the preceding vowel.

SUBJECT, MATERIAL AND METHOD

A male speaker of Southern Icelandic acted as subject. Based on the considerations outlined above the speech material was chosen to permit contrast of preaspirated plosives, unaspirated plosives, voiceless nasal plus plosive and voiced nasal plus plosive. This led to the following list of 10 words, the sounds of interest all being in medial, post-stress position.

"hitti"	preaspirated
"hiti"	unaspirated
"henti"	voiceless nasal
"hendi"	voiced nasal
"seppi"	preaspirated
"sepi"	unaspirated
"dempi"	voiceless nasal
"dembi"	voiced nasal
"hetta"	preaspirated
"sempinn"	voiceless nasal

This list was subjected to six different randomizations, the words being embedded in the sentence frame:
"ég segi ..... þá"
The following procedure was followed for the recordings: Glottal activity was assessed by the transillumination technique and velar activity by Künzel's photoelectric method /5/. For obvious technical reasons velar and glottal activity had to be recorded separately. Precautions were taken to ensure that both recordings were spoken at roughly the same rate. A third recording of oral air-pressure will not be discussed further here; however, it enabled the audio-signal measurements to be based on 18 tokens per word (3 recordings x 6) rather than 6 tokens per word as in the case of the physiological signals.

RESULTS

The results will be presented in three sections: first, the temporal measurements made on the speech signal (summarized in Table 1); and then the results for glottal and velar activity.
Figure 1 shows ensemble averages for glottograph and velograph signals and audio envelope for eight of the ten words. Note that the velar and glottal signals were not, of course, recorded simultaneously. The audio envelope represents an average over both recordings.

Segment Durations

The words will be treated as consisting of a maximum of three basic segments (cf. Table 1):

1. the vowel preceding the consonantal group,
2. the nasal or preaspiration section,
3. the occlusion phase of the plosive.

Vowel length divided the material into two groups, one group consisting only of "sepi" and "hiti", in which the (diphthongized) vowel was 120-140 ms in length (segment 2 being completely lacking here, of course) and a second group containing all other words, in which the length of the vowel was roughly 60-85 ms. As far as segment 2 is concerned, the length of preaspiration was clearly shorter than that of the nasals:
50-70 ms vs. 110-150 ms. Within the nasals the voiceless sounds tended to be longer than the voiced counterparts.
The length of the plosive occlusion tended to counterbalance the length of segment 2 with much shorter plosives following the nasals than in the preaspirated or unaspirated conditions:
ca. 80 ms vs. ca. 120-140 ms.
The relative lengths of preaspiration and the plosive occlusion are comparable to those in /3/ and /10/, while in /7/ the preaspiration was sometimes as long as the following occlusion. Clearly this did not occur here. Nonetheless, there seems to be a feeling in the literature that preaspiration is so long that it cannot be a mirror-image of postaspiration but rather is an independent segment. We believe this argument has been over-valued (but see discussion).
The fact that occlusions for preaspirated and unaspirated plosives are not reliably distinguishable in length also agrees with /3/.

Table 1

Main segment durations averaged over all three recordings (n = 18).

	vowel		nasal or preasp.		occlusion	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
hitti	75	9.0	51	12.5	146	20.7
hiti	142	14.5	---	---	134	16.4
henti	63	10.4	141	13.5	92	11.2
hendi	63	6.0	115	11.7	92	12.7
seppi	59	5.7	76	10.0	123	14.9
sepi	125	14.4	---	---	134	16.1
dempi	68	4.4	143	10.5	87	15.3
dembi	86	5.2	108	12.2	79	20.8
hetta	70	7.2	67	8.4	136	15.9
sempinn	62	4.5	130	10.6	83	15.1

Glottal Activity

Fig. 1 shows that the amplitude of the devoicing gesture for HC and NH was very large, being greater even than for the pre-vocalic fricatives (e.g. /s/ in "seppi"). This agrees with /10/ while Löfqvist /6/ generally noted rather smaller glottal openings for HC. The unaspirated plosives (e.g. "sepi") had a clear devoicing gesture but of restricted amplitude while for the plosives following the voiced nasals no evidence of glottal opening could be seen. Pre-vocalic /h/ (e.g. "hitti") also had a devoicing gesture, of course, but also of restricted amplitude so that, in view of the unconstricted vocal tract in these sounds, they were often not completely voiceless.
The above remarks refer to aspects of the glottal gesture that are so prominent as to be unaffected by non-linearities in the glottographic signal. Any differences in glottal activity between HC and NH are clearly more subtle, however. In the nasals the glottal gesture, particularly the adductory phase, coincides with a period of fast velar movement, so that reliable differences in maximum glottal opening and maximum opening and closing speed cannot be assumed.
The timing of glottal activity can be approached with more confidence, however. There are some clear differences, but also some interesting similarities between HC and NH. Firstly, it should come as no surprise that the interval from peak glottal opening (PGO) to formation of the occlusion for the plosive is completely different for the HC and NH sounds. The data for all 18 tokens of HC and NH gives:

$\bar{x}$ = 85.4 ms; sd = 13.5 ms for the nasals
$\bar{x}$ = 10.4 ms; sd = 8.6 ms for the preaspirates

This is a natural consequence of the broad similarity in the glottal gesture and the different structuring of the oral articulation exemplified in Table 1.
Regarding similarities in timing the first point to be made is that there is no difference in the overall duration of the gesture as estimated by the interval from instant of maximum opening speed to instant of maximum closing speed.
Support for the hypothesis developed on the basis of the earlier German material was also found; while it unfortunately proved impossible to measure the beginning of glottal abduction reliably, nonetheless the interval from vowel onset to the moment of maximal abduction velocity, and from vowel onset to PGO were not significantly different in the two classes of sounds.

Velar Activity

Regarding the nasals as a group it has already been pointed out that the voiceless nasals as measured from the audio signal were clearly longer than the voiced counterparts. This result was highly significant. There was, however, no reciprocal adjustment in the length of the following plosive, nor differences in the length of the preceding vowel. This replicates results previously presented by Pétursson /9/ quite closely, although he did find some reciprocal adjustment in the length of the plosive.
Regarding the velar gesture itself there were no significant differences in maximum raising or lowering speed or in overall excursion. Taking the length of the interval between the positions of maximum lowering and raising speed as a measure of the length of the velar gesture the values for "dempi" were significantly greater than those for "dembi" in a straight t-test. For "henti"/"hendi" this comparison would be meaningless because of the considerable anticipatory coarticulation (see Fig. 1).
A further motivation for examining velar activity in these groups of sounds was that in his cineradiographic recordings Pétursson /7/ had observed slight velar lowering in the preaspirates as well as in pre-vocalic /h/ and interpreted this as additional evidence for regarding preaspiration phonemically as /h/. We were unable to replicate this finding in our material; the pairs "seppi"/"sepi" and "hitti"/"hiti" clearly do not differ in velar activity (See Fig. 1). In both of these pairs the velum is lower in the vowels than in the adjacent occlusives, but this simply reflects the well-known differences in intrinsic velar height for these sounds /2/. In pre-vocalic /h/ (e.g "hitti", "henti") the velum can clearly take on a high or low position depending on the surrounding sounds. In fact, this dependence on the environment seems so complete that one could doubt whether /h/ has any intrinsic height specification at all.

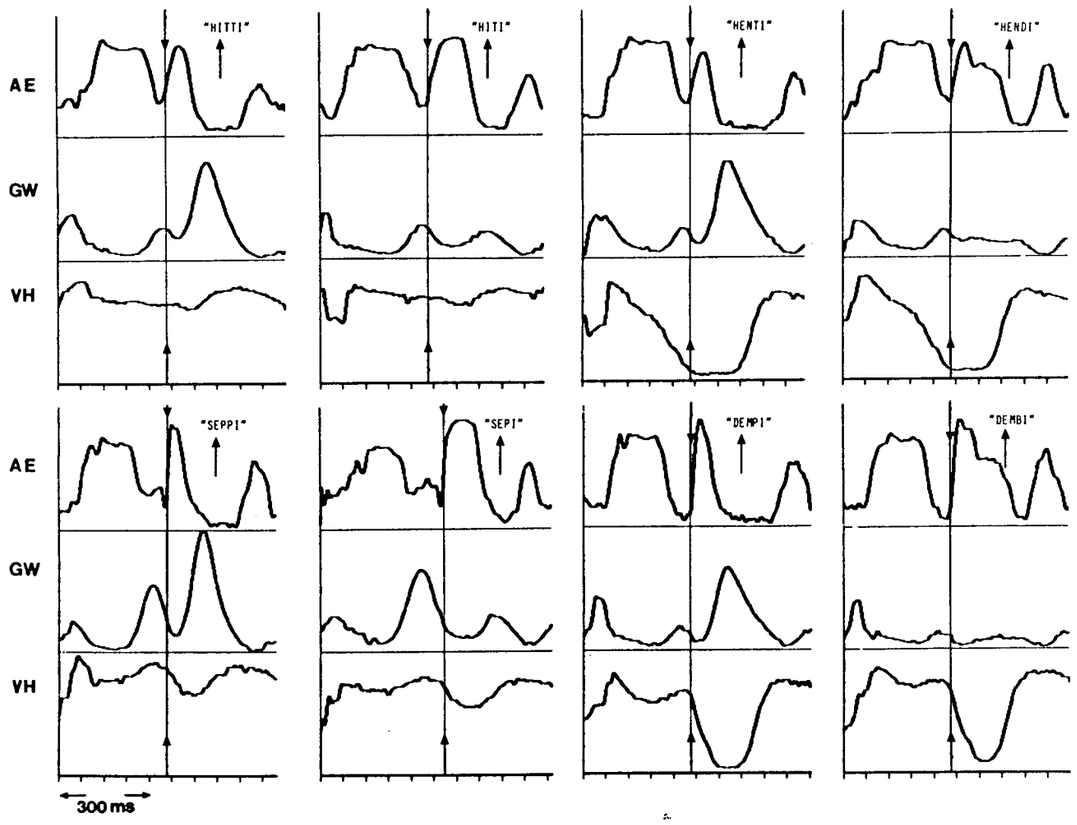


Fig.1 Averaged audio envelopes (AE), transillumination (GW), and velograph signals (VH), lined-up at onset of arrowed vowel.

DISCUSSION

The first question to consider here is the segmental status of preaspiration. Since we were unable to reproduce Pétursson's observations of velar behaviour in the preaspirates the main point is how much weight should be attached to comparisons of the length of preaspiration and other kinds of segment. In our material HC seems to be behaving like a single consonant since 1) preaspiration was shorter than the nasal segments in the nasal-plosive clusters, 2) the occlusion for preaspirated and unaspirated plosives is the same length, 3) plosive occlusions in the nasal-plosive clusters are much shorter than in the simple preaspirated or unaspirated plosives, 4) the shortening effect on the plosive of introducing a preceding nasal is virtually identical in the hypothetical preaspirated case and in the unaspirated case.

There is still perhaps an extremely cogent reason for retaining the segmental status of preaspiration; as Árnason points out /1/, words with preaspiration are given contrastive stress by lengthening the preaspiration phase, which in this respect behaves completely independently of the following plosive. It would be interesting to examine this phenomenon glottographically.

Turning to the glottal activity one could argue that the similarities for HC and NH suggest that two voiceless segments are present in both cases. However one might equally well point to the great difference between pre-vocalic and putative pre-consonantal /h/.

Thus this kind of argument does not get us very far, the same being true to a lesser extent for the segment-length arguments.

I believe that it is more fruitful to consider the aerodynamic and physiological constraints within which the language's contrasts must be produced.

Thus it is, for example, by no means clear that preaspiration and postaspiration of equal length are equally perceptually prominent, the one being superimposed on a closing, and the other on an opening movement of the vocal tract. Moreover, reliable devoicing with an unconstricted vocal tract, as at the beginning of the preaspiration phase, requires a large-amplitude glottal opening. This may explain why preaspiration does tend to be longer than postaspiration and perhaps also a less common phenomenon. Yet the fact that they are not perfect mirror-images of each other clearly need not mean that fundamentally different types of segment are involved.

The question of perceptibility may, as Pétursson has pointed out /9/, also explain why voiceless nasals are longer than voiced ones: these voiceless segments must be distinguishable from the quite large number of other voiceless segments that Icelandic can allow in this position.

We thus believe that up to this point there are no compelling reasons for regarding preaspiration as a separate segment, whether analyzed as an /h/ phoneme or as an auto-segment that has moved from the plosive where it originated.

This opens the way for viewing the existence of the voiceless nasals from the point of view of the speech motor system as a simple coarticulatory phenomenon. We would suggest that the voicelessness of the nasals is essentially a mirror-image of the voicelessness of the /l/ segment in English "plea". It was shown here that, as in the earlier German example, HC and NH share a very similar glottal gesture, but with reorganisation of the supra-glottal articulation. We would hypothesize that pairs such as "plea" and "pea" in English also superimpose a constant glottal gesture on reorganised oral articulation. While we believe that our results allow a coherent description of these sounds from a motor-speech perspective the restricted nature of the material examined naturally leaves a number of questions unanswered. In particular it would be desirable to include a more comprehensive range of voiced and voiceless continuants plus plosive in the description and to examine the realization of these sounds under contrastive stress.

REFERENCES

/1/ Árnason, K. (1986): "The segmental and suprasegmental status of preaspiration in modern Icelandic". Nordic J. of Linguistics 9:1-23

/2/ Bell-Berti, F. (1980): "Velopharyngeal function: A spatio-temporal model. Haskins SR 63/4: 41ff.

/3/ Garnes, S. (1973): "Phonetic evidence supporting a phonological analysis". J. Phonetics 1:273-283

/4/ Hoole, P. A., Pompino-Marschall, B. and Dames, M. (1984): "Glottal timing in German voiceless occlusives". Proc. 10th. Int. Congress of Phonetic Sciences, pp. 399-403. Van den Broecke M.P.R. and Cohen A. (eds.)

/5/ Künzel,H.J (1979). "Röntgenvideographische Evaluierung eines photoelektrischen Verfahrens zur Registrierung der Velumhöhe beim Sprechen". Folia Phoniatrica 31:153-166

/6/ Löfqvist, A. and Yoshioka, H. (1980): "Laryngeal activity in Icelandic obstruent production". Haskins SR 63/4: 275ff

/7/ Pétursson, M. (1972): "La préaspiration en islandais moderne. Examen de sa réalisation phonétique chez deux sujets". Studia Linguistica 26:61-80

/8/ Pétursson, M. (1973b): "L'origine des phonèmes nasals et liquides sourds et du [h] postvocalique de l'islandais moderne". Orbis 22:335-342.

/9/ Pétursson, M. (1975): "Examen expérimental de la réalisation phonétique des nasales sourdes de l'islandais moderne". In: K.-H. Dahlstedt (ed.), The Nordic Languages and Modern Linguistics 2:634-653.

/10/ Pétursson, M. (1976): "Aspiration et activité glottale. Examen expérimental à partir de consonnes islandaises". Phonetica 33:169-198.

CONTRIBUTION DU CRICOTHYROIDIEN ET DES MUSCLES SOUS-HYOIDIENS AUX VARIATIONS DE LA FREQUENCE FONDAMENTALE EN FRANCAIS : APPROCHE ELECTROMYOGRAPHIQUE

☆Denis AUTESSERRE, Bernard ROUBEAU, ☆Albert DI CRISTO, Claude CHEVRIE-MULLER, ☆ Daniel HIRST, Jean LACAU et Bernard MATON

☆ Institut de Phonétique, UP 1, UA 261 CNRS, Aix-en-Provence, France
U 2 INSERM, Hôpital de la Salpêtrière, Paris, France.

RESUME

Les signaux électromyographiques du cricothyroïdien (CT) et de trois muscles sous-hyoïdiens sont recueillis chez un locuteur français réalisant des configurations mélodiques Moyen-Haut-Bas (MHB) et Haut-Bas-Haut (HBH). Les résultats obtenus lors de ce premier enregistrement sont précisés par une deuxième étude, restreinte aux degrés d'activité du CT pendant la prononciation d'un ensemble de phrases, représentatif des patrons intonatifs du français.
Nous analysons plus particulièrement les relations entre les pics d'activité de ces muscles, d'une part, et d'autre part la distribution des tons mélodiques dans la phrase et le contexte segmental (fréquence intrinsèque et voisement). Les maxima d'amplitude des pics de CT coïncident avec le point d'inflexion de la courbe ascendante de Fo, comme s'il s'agissait de la première dérivée de cette courbe.

INTRODUCTION

Il semble établi, à la suite de nombreuses études, que les variations de la fréquence fondamentale (Fo) dans la parole dépendent de deux composantes interactives majeures : l'abaissement graduel de la pression sous-glottique (PS), responsable de la ligne de déclinaison, et l'activité des muscles du larynx qui contrôlent les inflexions locales de Fo (montées et chutes) superposées à cette ligne de déclinaison /1/, /2/, /3/.
La plupart des spécialistes, et en particulier /4/ et /5/, attribuent à trois paires de muscles laryngés le contrôle des variations positives de Fo : le cricothyroïdien (CT), le vocalis (V) et le crico-aryténoïdien latéral (CAL). Toutefois le V et le CAL, dont la fonction primaire est d'assurer l'adduction des cordes vocales, semblent moins directement associés aux variations locales de Fo que ne l'est le CT. Le contrôle des inflexions négatives de Fo est moins bien connu /6/, /7/. Il n'est pas exclu que le CT joue, cette fois encore, un rôle fondamental, mais dans ce cas, par sa relaxation. Celle-ci s'accompagne certainement d'une participation de certains muscles sous-hyoïdiens tels que le sterno-hyoïdien /7/, le sterno-thyroïdien /8/, et peut-être le sterno-cleido-hyoïdien.

Nous nous proposons d'étudier les relations entre les variations de la fréquence fondamentale et les phases d'activité de trois muscles sous-hyoïdiens et du CT dans des conditions expérimentales qui seront précisées ci-après.

1. - Corpus et enregistrements.

Le corpus enregistré par un locuteur français, est constitué, d'une part, de configurations mélodiques complexes, montantes-descendantes, Moyen-Haut-Bas (MHB) et descendantes-montantes, Haut-Bas-Haut (HBH); d'autre part, de phrases choisies afin d'illustrer la réalisation des principaux patrons intonatifs du français : continuation majeure et mineure, questions totale ou partielle, implication, appel, parenthèses.

2. - Technique électromyographique (EMG).

Nous avons recueilli simultanément l'activité EMG de quatre muscles : le cricothyroïdien (CT) et trois muscles sous-hyoïdiens : le thyro-hyoïdien (TH), le sterno-thyroïdien (ST) et le sterno-cleido-hyoïdien (SCH). Pour l'étude des configurations mélodiques complexes (MHB) et (HBH) nous avons considéré l'ensemble de ces muscles. En revanche pour l'analyse des phrases, nous avons limité l'investigation au seul cricothyroïdien. La technique choisie est celle des électrodes bipolaires (2 fils de platine) décrite par (10). Afin de vérifier si les électrodes ont bien été implantées dans le muscle choisi, on a procédé à plusieurs épreuves de contrôle : respiration calme puis profonde, déglutition, résistance à l'ouverture de la machoire.
Les signaux EMG et la parole sont enregistrés sur un magnétophone Ampex (7 pistes) et reproduits sur papier à l'aide d'un enregistreur Gould ES 1000. Les signaux de parole sont captés à la fois par un microphone et par un électroglottographe (Frokjaer-Jensen, type EG 830) relié à un détecteur de mélodie. Le locuteur et les différents capteurs sont placés dans une chambre insonorisée et les enregistreurs disposés en dehors de la cabine.

3. - Les paramètres enregistrés et leurs mesures.

L'analyse porte sur les tracés du phonogramme, du glottogramme, de la courbe mélodique extraite du signal glottographique, des deux types de représentation des signaux EMG, brut et redressé.
Les divers types de tracés correspondant à des étapes différentes de la contraction musculaire (ébauchée, franche ou forte) ou à des épisodes de relâchement ont été identifiés qualitativement par ré-

férence au signal acoustique. Cette comparaison est complétée par une mise en correspondance, à partir du signal EMG redressé, des pics d'activité des bouffées EMG les plus importantes et des sommets des configurations de Fo sur la courbe mélodique.

RESULTATS

1. Comparaison des activités électromyographiques du cricothyroïdien et des muscles sous-hyoïdiens dans la réalisation des configurations mélodiques Moyen-Haut-Bas (MHB) et Haut-Bas-Haut (HBH)

- On observe de façon systématique une anticipation de l'activité EMG de tous les muscles considérés, par rapport au début de la phonation (Fig. 1 A et 1 B). Ceci confère à ces muscles un rôle important dans la préparation de l'activité vibratoire du larynx (attaque).
- Les muscles sous-hyoïdiens explorés fonctionnent de manière synergique (Fig. 1 A).
- A l'attaque de la configuration MHB, on remarque une activité plus importante des sous-hyoïdiens, ce qui n'est pas le cas lors de l'attaque de HBH, où l'activité du CT l'emporte.
- L'activité du CT se produit avec une anticipation plus grande à l'attaque qu'au cours de l'émission d'un ton haut à l'intérieur de la configuration MHB ou qu'en finale de HBH.
- D'autre part, l'activité des sous-hyoïdiens est fortement corrélée à la réalisation des tons bas, ce qui semble conférer à ces muscles un rôle important dans le contrôle des variations descendantes de Fo (7). Leur fonctionnement apparaît alors de type antagoniste par rapport à celui du CT. Ceci n'est plus vrai dans le cas d'un ton haut final (Fig. 1 B) pour lequel l'ensemble des muscles fonctionne de manière agoniste (Tableau 1) : ce traitement particulier en position finale impliquerait l'existence d'un processus de contrôle mis en œuvre pour éviter un dépassement de la cible de Fo à atteindre /11/.

2. Comparaison des activités du cricothyroïde dans la réalisation des configurations mélodiques des principaux patrons intonatifs du français.

Si les effets de la contraction du CT apparaissent bien connus, les signaux EMG de notre locuteur nous conduisent à mettre en évidence des faits moins souvent signalés de relâchement de ce muscle (prenant même une allure d'inhibition) qui paraissent ici particulièrement significatifs.

2.1. La présence ou l'absence de ce relâchement est liée à deux facteurs :

- La nature du ton à l'initiale de la phrase : le relâchement se produit à l'initiale des phrases commençant par un ton moyen ou un ton bas (phrases déclaratives et questions totales). Il est absent au début des questions partielles (du type "qu'est-ce que tu lui as dit ?") qui commencent par un ton haut (Fig. 2).
- La nature de l'unité phonique à l'initiale : dans les phrases débutant par un ton haut moyen ou un ton bas, l'organisation temporelle de ces phases de relâchement semble être influencée par le caractère voisé ou non voisé de la première unité

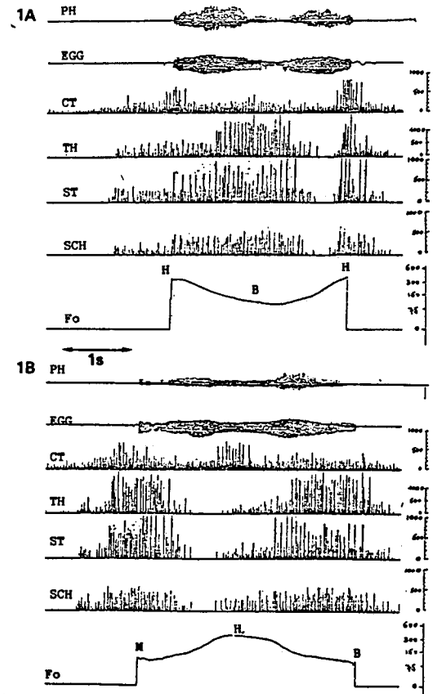


Fig. 1. - Phonogramme (Ph), Glottogramme (EGG) et signaux EMG redressés du cricothyroïdien (CT), du thyro-hyoïdien (TH), du sterno-thyroïdien (ST) et du sterno-cleido-thyroïdien (SCH) et courbe de fréquence fondamentale (Fo), correspondant aux configurations MHB (1 A) et HBH (1 B) réalisées à partir de la voyelle /o/.

TONS	M	H	B
FO (Hz)	110	356	94
CT (µV)	1307	1693	564
ST (µV)	1107	457	821

TONS	H	B	H
FO (Hz)	337	93	377
CT (µV)	1500	366	2216
ST (µV)	550	1150	1083

Tableau 1

Valeurs moyennes de la fréquence fondamentale (Fo), de l'amplitude des pics d'activité EMG du cricothyroïdien (CT) et du sterno-thyroïdien (ST) correspondant aux points cibles (Medium,M,Haut, et Bas,B) des configurations mélodiques montante-descendante (1 A) et descendante-montante (1 B).

phonique. Ces moments de relâchement précèdent la réalisation d'une voyelle ou d'une consonne voisée initiales, alors qu'ils se produisent pendant la tenue d'une consonne non voisée initiale (Fig. 3 et 4).
En finale d'énoncé, le relâchement est en relation directe avec la chute de Fo pour la finalité, la parenthèse basse (Fig. 3) et l'appel ("Jean Paul !", Fig. 5) dont le contour mélodique comporte une montée puis une chute de la voix.

2.2. Les phases d'activité.

Les pics d'activité du CT sont systématiquement associés aux sommets (tons hauts) des contours mélodiques de continuation mineure et majeure et de la question totale. Dans ces cas, la hiérarchie des pics d'activité du CT reflète celle des sommets de Fo (Fig. 6).
Pour les contours continuatifs, la phase d'activité la plus importante commence de manière anticipée sur la consonne de la syllabe accentuée portant le ton haut (Fig. 7).
En revanche, l'anticipation du pic EMG est moins marquée dans les questions totales où elle se situe vers le milieu ou dans le dernier tiers de la voyelle accentuée, qui porte le ton haut (Fig. 8).
D'autre part et en première approximation, il semblerait qu'il y ait une corrélation positive entre l'amplitude du pic d'activité du CT et la fréquence intrinsèque des voyelles en syllabe accentuée, dans la mesure où les valeurs de ces pics reflètent la hiérarchie des valeurs de fréquence intrinsèque observées par /12/ et /13/.

DISCUSSION

Les résultats de cette première investigation sur l'activité des muscles du larynx pendant la réalisation des configurations mélodiques du français nous permettent de formuler les remarques suivantes :

1. Les muscles sous-hyoïdiens, retenus pour cette expérience fonctionnent, dans la plupart des cas, en synergie, en relation avec les inflexions négatives de Fo.
2. Les relations entre ces muscles et le CT sont de type antagoniste, dans le cas d'un ton initial ou médian et agoniste, lors de la réalisation d'un ton final.
3. L'activité du CT est systématiquement associée, en français /14/ à la production d'un ton haut conformément aux observations déjà effectuées dans d'autres langues.
4. Nos résultats tendraient à mettre en évidence l'existence d'une véritable pré-programmation, prenant en compte, à l'attaque de l'énoncé, la présence du voisement et la réalisation d'un ton haut.
5. La comparaison du chronométrage de l'activité de CT pour les questions totales et les continuations semble justifier la distinction établie par /15/ entre configuration concave et configuration convexe, ou encore, d'après notre analyse, entre montée retardée (question) et montée anticipée (continuation).
6. Les résultats obtenus à partir d'un premier sondage effectué sur quelques syllabes (Tableau 2),

	Voyelles orales			Voyelles nasales
	Voyelles basses	Voyelles hautes		
	[ɛ, a]	[i]	[u]	[ɛ̃]
Série 1	866	1033	1066	933
Série 2	779	841	957	985
Moyenne	822	937	1011	959

Tableau 2

Valeurs de l'amplitude du pic d'activité du CT (en uV).

tendraient à montrer que la fréquence intrinsèque n'est pas le résultat d'une contrainte physiologique mais correspond bien à un processus actif /16/.
7. La mise en correspondance systématique des pics d'activité EMG et des courbes mélodiques nous conduit à considérer la courbe de Fo comme l'intégrale de l'activité du CT : le maximum d'amplitude de ces pics coïncide avec le point d'inflexion de la courbe ascendante de Fo (Fig. 9), comme s'il s'agissait de la première dérivée de cette courbe.

BIBLIOGRAPHIE

/1/ COLLIER, R., "Laryngeal muscle activity, subglottal air pressure and the control of pitch in speech", **Haskins Lab. Status Rep. Speech Res.**, SR - 39/40, pp. 137-170, 1974.
/2/ COLLIER, R., "Physiological correlates of intonation patterns", **J. Acoust. Soc. Am.**, vol. 58, n° 1, pp. 249-255, 1975.
/3/ LADD, D. R., "Declination : a review and some hypothesis", **Phonology Year Book**, n° 1, pp. 53-74, 1984.
/4/ HIRANO, M., OHALA, J. et VENNARD, W., "The function of laryngeal muscles in regulation of fundamental frequency and intensity of phonation", **J. Speech Hear. Res.**, 12 : pp. 616-628, 1969.
/5/ SAWASHIMA, N., "Laryngeal research in experimental phonetics", **Current trends in linguistics**, vol. 12, pp. 69-115, 1972.
/6/ LIEBERMAN, P., "A study of prosodic features", **Status Report on Speech Research** (Haskins Lab.), SR - 23, pp. 179-208, 1972.
/7/ OHALA, J., "How pitch is lowered ?", **Mim. Phonology Lab. Univ. of Berkeley**, pp. 25-30, 1972.
/8/ FUJIMURA, O., "Phonological functions of the larynx in phonetic control", **Invited Paper at the International Congress of Phonetic Sciences**, Miami, 1977.
/9/ MAEDA, S., "On the Fo control mechanisms of the larynx", **Séminaire Larynx et Parole**, Institut de Phonétique de Grenoble, GALF, pp. 243-257, 1979.

/10/ BASMAJIAN, J. V. et STECKO, G. A., "A new bipolar indwelling electrode for electromyography", **Journal of Applied Physiology**, 17, 849, 1962.
/11/ SUNDBERG, J., "Maximum speed of pitch changes in singers and untrained subjects", **Journal of Phonetics**, 7, 2, pp. 71-79, 1979.
/12/ DI CRISTO, A., "De la microprosodie à l'intonosyntaxe", éd. J. Lafitte, 854 p., 1985.
/13/ ROSSI, M. et AUTESSERRE, D., "Movements of the hyoid bone and the larynx and the intrinsic frequency of vowels", **J. of Phonetics**, 9, 233-249, 1981.
/14/ AUTESSERRE, D., DI CRISTO, A. et HIRST, D., "Approche phonologique des intonations de base du français : cricothyroïdien et fréquence fondamentale", **15es Journées d'Etude sur la Parole**, Institut de Phonétique d'Aix-en-Provence, 27-30 mai 1986, pp. 37-41.
/15/ DELATTRE, P., "Les dix intonations de base du français", **French Review**, 40, pp. 1-14, 1966.
/16/ DI CRISTO, A. et HIRST, D., "Modelling French micromelody", **Phonetica**, 43, 1, pp. 11-30, 1986.

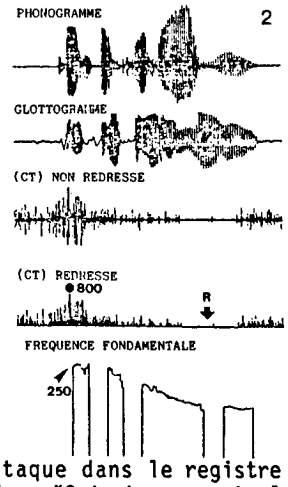


Fig. 2. - Attaque dans le registre haut. Question partielle : "Qu'est-ce que tu lui as dit ?".

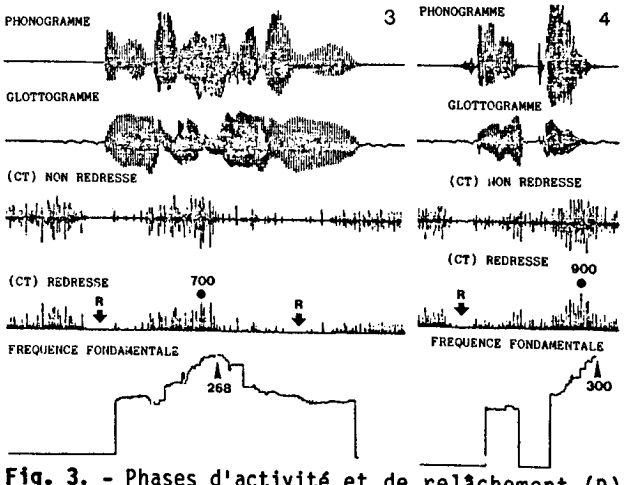


Fig. 3. - Phases d'activité et de relâchement (R) du CT dans l'énoncé "Evidemment ! Je l'ai vu" (implication + parenthèse basse).

Fig. 4. - Phases d'activité et de relâchement (R) du CT dans la phrase "Sur le quai ?" (question totale).

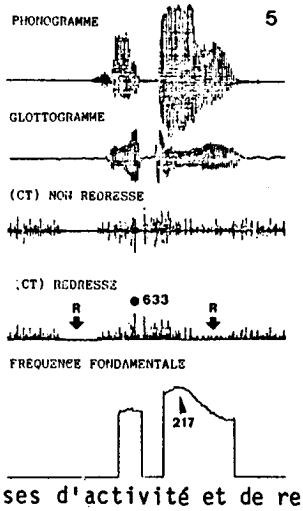


Fig. 5. - Phases d'activité et de relâchement (R) du CT dans le contour mélodique complexe de l'appel ("Jean-Paul !").

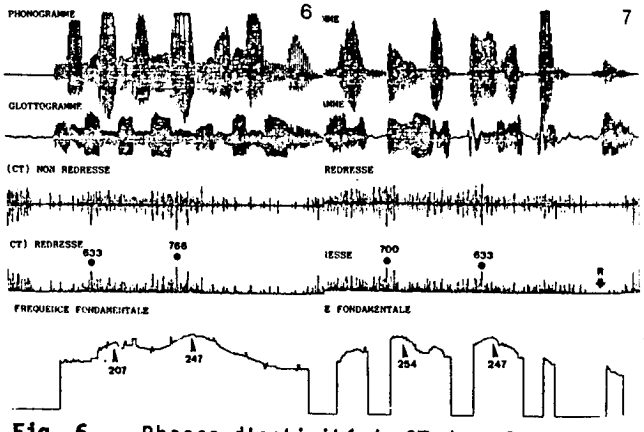


Fig. 6. - Phases d'activité du CT dans la réitération ("mamama...") de la phrase "Le fils de Paul est parti".

Fig. 7. - Phases d'activité du CT dans la phrase "Le fils de Paul est parti".

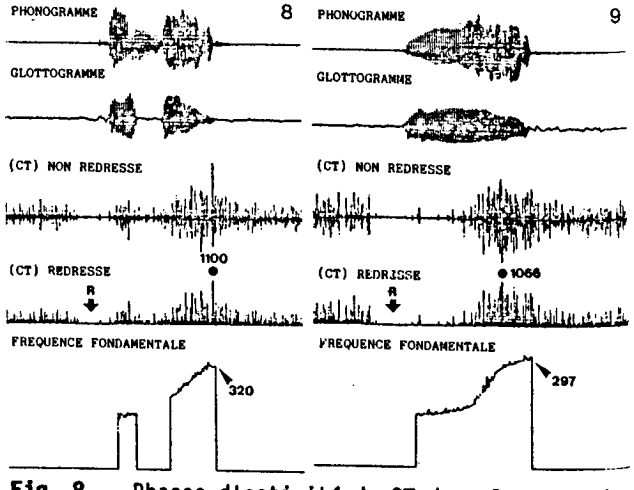


Fig. 8. - Phases d'activité du CT dans la question totale "Tu es sûr ?".

Fig. 9. - Correspondance entre le pic d'amplitude du CT et le point d'inflexion de la configuration de Fo (BH).

INTERSEGMENTAL (VC & VCC) AND INTRASEGMENTAL (VOT & VTT) PHASINGS IN FRENCH

Rudolph SOCK Liisa OLLILA Christine DELATTRE Corinne ZILLIOX

Institut de Phonétique de Grenoble
Institut de la Communication Parlée
38031 Grenoble Cédex, FRANCE

ABSTRACT

The main thrust of this study is to examine on the acoustic level and across rate conditions, intersegmental and intrasegmental phasings in close phonetic classes : VC vs VCC. The psychomotoric recognition pattern paradigm adopted here is a means of inferring motor programs, separate or generalized, from the different types of phasing structures observed. We should also be able to compare our acoustic data with related results on movement studies, especially in the intrasegmental domain, for which lesser data is available.

INTRODUCTION

Studies of timing in speech on the articulatory level postulate a relative constancy, across stress and rate for gestures associated with adjacent segments /1/ and also for gestures associated with events within one segment /2/. These results obtained for behaviours between oral articulators (intersegmentally) and between oral and laryngeal gestures (intrasegmentally) both suggest some motoric invariance.

In the light of this literature, we will investigate for French on the acoustic level intersegmental (VC & VCC) and intrasegmental (VOT & VTT) phasings across two rate conditions. The cyclic production activity in speech allowed us to determine, on the acoustic level, from reiterating events a release cycle and retain, within this cycle, four phases : VOT, Vocalic, VTT and Silence (cf. below). Our results will concentrate on the Vocalic, VTT and VOT ones. Adopting hence the psychomotoric recognition pattern paradigm borrowed from studies based on phasing, - e.g. those for human intralimb kinematics /3/, - we will try to show from our results that instead of simply advocating relative timing or phase invariance, we could search for clear-cut structural patterns (inferential of separate motor programs) and this for different - though linearly linked (as to cycle transformation) - linguistic tasks, like those found for the same gait pattern analysis in walking vs jogging /4/.

On the intrasegmental level, we will examine the behaviours of VOT /5/ and VTT (Voice Termination Time, see /6/ for measurements ; and /7/ for cue validation) in the VC domain since both are linked to the timing of the glottal gesture in relation to supraglottal release and closure respectively. If the intrasegmental timing for VOT is comparatively well known (for French see /8/), we do not dispose of enough information on the role that VTT plays in consonantal coordinations. Thus, we will attempt to situate our results in the perspective of glottal gesture control in relation to the two consonantal commands for closing and release responsible for producing the consonant hold. In recent research context, preoccupied with the phasing of these commands as revealing ultimate intersegmental motor programs /9/, this study should provide a comparison between the acoustic manifestations of our VTT with results obtained by /2/ examining the production of an intrasegmental coordination.

LINGUISTIC MATERIAL

The linguistic material consisted of pairs derived from the following French verbs : empatter / empâter, téter / éteter, coter / écôter, égoutter / goûter. In certain contexts, these verbs become real minimal pairs permitting to test, in the VC domain, the effect of consonantal gemination on vowel length (pairs of the type : "nous l'empattons ? / nous l'empâtons ? vs "nous l'empatt't-on ? / nous l'empât't-on ?". See /10/ for a detailed description of this corpus). 12 repetitions were obtained for each item, amounting to 768 items for each speaker in two rate conditions :

normal and fast. To account for inter/intrasegmental phasings between VC and VCC classes, we will limit here our data to the first and last pairs in the series above (vowels /a/, /a:/, /u/, /u:/) and to preliminary results on 5 speakers from a larger data base. Our results will focus on three of them : speaker J.P. respects quantity contrasts (rare in present French) ; speaker R.L. is an example of representative results obtained in this study and elsewhere ; and Speaker C.F. has markedly different phasing structures from others.

After recordings, speech signals were digitalized and items extracted from their carrier sentences, segmented manually into acoustic events with the help of a speech editor /11/. A total of 44,490 events were detected.

The following parameters were retained in the VC domain : vocalic phase (VOT plus DVOC = clear vocalic formant structure), closure duration, VOT, VTT (from the beginning of closure to the end of voicing), the silent phase and the release cycle. For details on measurement procedures, cf. /12/.

RESULTS AND DISCUSSIONS

1. INTERSEGMENTAL PHASINGS

Speaker J.P. (with vowel length)

In studying the effects of rate on the relative acoustic timing for our different phonetic classes, we observe in the release cycle a clear vocalic phase percentage difference (around 17%) between VC,V:C and VCC,V:CC classes. This distinction between the two categories, although highly significant (t=13.76) is not at all surprising since one is to expect a difference in ratio between a vowel followed by a single consonant and a vowel followed by a geminate one. What is interesting is that the vocalic phases for VC and VCC classes stay relatively constant in the release cycle, across quantity and rate conditions. The separation remains efficient in such a way that, for example, VCC classes in fast rate conditions are never confused with VC classes in normal rate conditions. We noted also that classes differed clearly along the cycle dimension. The outcome is two structural patterns for the two linguistic entities.

Speaker R.L.

Although this speaker has a different vowel system from speaker J.P., the overall consonantal phasing structure is the same. Figure 1 is a good illustration of the intersegmental strategy adopted also by the previous speaker. The relative stability across conditions evoked above is also evident here, with a phase percentage difference of around 20% (t=15.42). This mean phase percentage difference coupled with the obvious mean cycle difference (around 97 ms ; t=13.86) provides a maintenance of two well defined phasing patterns for different phonological tasks. Similar manoeuvres have been systematically reported by /13/ for 8 French vowels (/a/, /a:/, /E/, /E:/, /ɔ/, /O:/, /u/, /u:/) in an identical experimental paradigm for Savoie regional French.

Speaker C.F.

Examining this speaker's results, we noticed some sort of a linear transition between VC and VCC phases. Even though mean values are rather different between the two phonetic classes (around 10% ; t=6.50), the transition from VCC in fast speech to VC in normal speech is assumed without any striking phase rupture. This transition is due, to a great extent, to the phase change, within the VC class, from fast to normal speech. Phonetic classes are less distinct along the cycle dimension. It is interesting to note that this pattern is extended, for this speaker, to the other vowels, with the phase transition phenomenon being even more continuous in the case of the vowel /u/. This behaviour, deviating from those adopted by the other speakers, allows us to observe the notion of relative permanence in a difference between motor programs (which are very likely not based solely on phase relations). For a discussion on this phenomenon of phase transition, giving a revival to old experiences in motor phonetics /14/ in the new scope of synergetics, see /15/.

2. INTRASEGMENTAL PHASINGS

2.1. VOT PHASINGS IN THE RELEASE CYCLE

The part that VOT plays in the release cycle seems to be quite different from both the vocalic and the VTT phase (see below). Like for the former, it is not surprising that quite similar dispersions between the two classes in absolute values give different percentages in such different cycles as VC and VCC. Note (e.g. for speaker R.L., Figure 2) that mean percentage differences are statistically significant (t=6.45). A more detailed examination of the two classes shows that : at first, the VOT phase diminishes abruptly as the release cycle increases and then the regression slope suddenly becomes relatively stable (around 240 ms) with further release cycle increase. So it seems that programming for longer cycles tends best to stability.

2.2. VTT PHASINGS IN THE RELEASE CYCLE

The overall tendency is to reduce the VTT proportionnaly as the release cycle increases. Figure 3 plots the duration of this release cycle against the VTT phase for the productions of the same Speaker R.L.. Correlation coefficients are significantly negative for both simple and geminate consonants (r=-.83 and -.61 respectively ; r=0.51, p=0.10), the falling regression slope being significantly steeper for simple than for geminate consonants. Between the two segmental classes the mean phase difference is small (just significant, t=2.90). On a whole, we can posit from these results that, as far as the VTT phase and the release cycle are concerned, the parameterization for this intrasegmental phasing is similar for both phonetic classes, thus implying a generalized motor program.

2.3. VTT PHASINGS AND CLOSURE

Having in mind that /16/ recently reaffirmed the fact that the glottal opening phase remained relatively constant in relationship with consonantal closure/constriction across rate conditions, we examined the ratio of VTT to closure variations across normal and fast speaking rates. Our results are as follows : the VTT phase is negatively correlated with closure duration. An increase in closure duration for the geminate class, however, has less proportionnal reducing effects on the VTT phase, the regression slope for the single consonant class being generally more pronounced than for the geminate one. Figure 4 is an illustration of this strategy. It seems, in other words, that there is more tendency towards VTT phase constancy as closure duration increases. These results are similar to those obtained by /17/ in an interspeaker related study. Note that the VTT phase tends to increase rapidly as closure is reduced, and notwithstanding the fact that we are still far from complete voicing (with a maximum of 40% VTT for about 65 ms closure), the drift promises to be swift towards flaps for which prevoicing no longer participates in categorizing the feature.

CONCLUSION

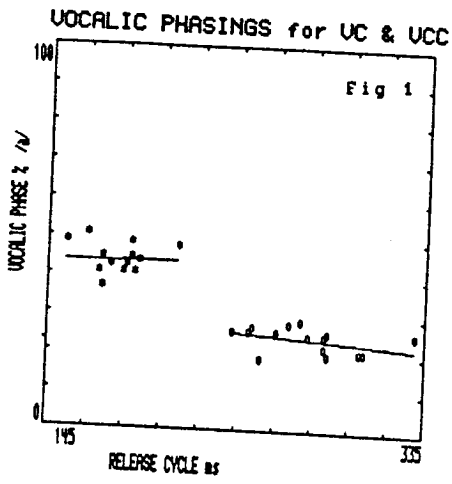
The very first question we raised in the beginning of this paper was in fact a reformulation of the invariance issue. Giving up research on relative timing invariance for a single phase in a cycle /18/, which appears to be quite problematic /19/, we looked at differences in multiphase patternings like those found for gait analysis /4/. Such data show clearly that re-patterning is a result of different types of behaviour within each phase. Some are constant (like E2 in the Philippson Cycle) ; some are slowly and smoothly changing (like E1) ; finally, a few give true discontinuities (like F). Likewise in speech, we found out that the only phase with such a decisive contribution to differences in patterning, related to close linguistic tasks, was the so called vocalic one. Contributions of VOT is of the E1 type, like for the silent phase (nearest to E3). This is true for our acoustic measurements. But it should be pointed out here that data in intrasegmental timing, considering the few available movement studies, will lead to similar results. Concerning VTT in particular, it is worth noting that /16/'s data could be situated in such research paradigms. Relative invariance of the peak glottal opening occurence in the closure/constriction time span is of course different from the acoustic VTT output, whose percentage varies inversely with closure increase (Figure 4). Movement results, when replaced in the total release cycle will give more differences than those found for our acoustic VTT : values around 50 to 60% for the peak glottal opening phase in closure will lead to quite different values in VC and VCC cycles (say 25% vs 40%). It seems therefore that this articulatory phasing is more discontinuous than for VTT, thus giving a more substantial contribution to differences in patterning. This remains to be quantified so as to evaluate the participation of intrasegmental phasings in speech cycles. The main question to ask now is : how can the dependency of laryngeal timing in its coordination with supralaryngeal gestures as a single programming be more evidently revealed ?

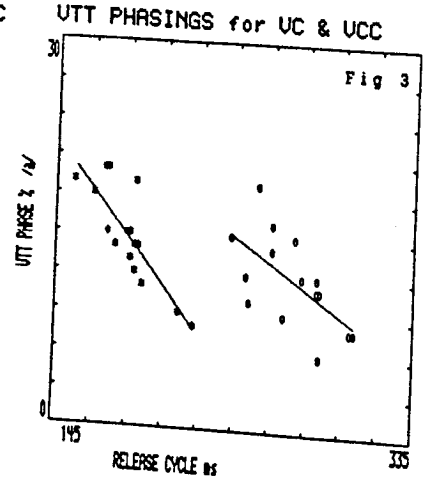
ACKNOWLEDGEMENTS : Special thanks to Christian ABRY, who followed this work closely, for his comments.

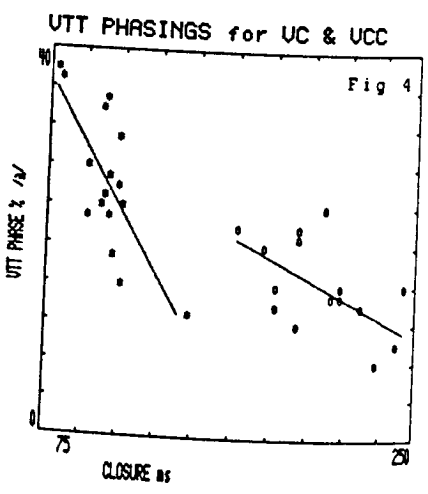
REFERENCES

/1/ TULLER B. KELSO J.A.S. & HARRIS K.S. (1982) Interarticulator Phasing as an Index of Temporal Regularity in Speech. J. Exp. Psychol. HPP 8, 460-472.
/2/ LOFQVIST A. & YOSHIOKA H. (1981) Interarticulator Programming in Obstruent Production. Phonetica 38, 21-34.
/3/ PHILIPPSON M. (1905) L'Autonomie et la Centralisation dans le Système des Animaux.Trav. Inst. Solvay 7, 1-208.
/4/ SHAPIRO D.C. ZERNICKE R.F. GREGOR R.J. & DIESTEL J.D. (1981) Evidence for Generalized Motor Programs using Gait Pattern Analysis. J. Motor Behav. 13, 33-47.
/5/ KLATT D.H. (1975) Voice Onset Time, Frication and Aspiration in Word-Initial Consonant Clusters. J. Speech Hearing Res.18, 686-706.
/6/ AGNELLO J.G. (1975) Measurement and Analysis of Visible Speech. in : Measurement Procedures in Speech, Hearing and Language. S.SINGH ed. 379-397.

/7/ BERG van den R.J.H. (1986) The Effect of Varying Voice and Noise Parameters on the Perception of Voicing in Dutch Two-Obstruent Sequences. Speech Com. 5, 355-367.
/8/ WAJSKOP M. (1979) Segmental Durations of French Intervocalic Plosives. in : Frontiers of Speech Communication Research, LINDBLOM & OHMAN eds. 109-123.
/9/ FOWLER C. RUBIN P. REMEZ R.E. & M.T. TURVEY (1980) Implications for Speech Production of a General Theory of Action. in : Language Production 1, B. BUTTERWORTH ed., 373-420.
/10/ ABRY C. SOCK R. BOE L.J. OLILLA L. DOUBLIER D. DELATTRE C. & ZILLIOX C.(1986) L'Organisation Temporelle des Voyelles et des Consonnes du Français. Durée Phonologique et Vitesse d'Elocution. Rapport CNET LANNION.
/11/ BENOIT C. (1984) EDISIG : Encore un Editeur de Signal ?! 13èmes JEP du GCP du GALF, 211-213.
/12/ ABRY C. BENOIT C. BOE L.J. & SOCK R. (1985) Un Choix d'Evénements pour l'Organisation Temporelle du Signal de Parole. 14èmes JEP du GCP du GALF, 133-137.
/13/ DOUBLIER D. (1986) La Résistivité de l'Organisation Temporelle des Oppositions de Quantité dans le Français de la Chapelle d'Abondance (Hte-Savoie) face aux Variations de la Vitesse d'Elocution. TER de Sci. du Lang. Grenoble III, 87p. (dir. C. ABRY).
/14/ STETSON R.H. (1951)Motor Phonetics : a Study of Speech Movements in Action. Amsterdam:North Holland.
/15/ KELSO J.A.S. SALTZMAN E.L. & TULLER B. (1986)The Dynamical Perspective on Speech Production : Data and Theory. J. of Phonetics 14, 29-59.
/16/ LOFQVIST A. & YOSHIOKA H. (1984) Intrasegmental Timing:Laryngeal-Oral Coordination in Voiceless Consonant Production. Speech Com. 3, 279-289.
/17/ SOCK R. & BENOIT C. (1986) VOTs et VTT en Français.15èmes JEP du GCP du GALF, 307-310.
/18/ TULLER B. & KELSO J.A.S. (1984) The Timing of Articulatory Gestures: Evidence for Relational Invariants. J. Acoust. Soc. Am. 76, 1030-1036.
/19/ BENOIT C. & ABRY C. (1986) Vowel-Consonant Timing across Speakers. 12th Int. Congr. Acoust. A6-1.







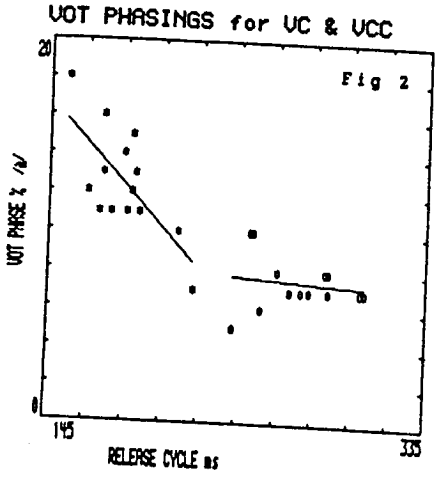


Fig. 1-3 Vocalic, VOT and VTT phasings in VC (*) and VCC (0) release cycles for "nous l'empattons ?" vs "nous l'empatt't-on ?" spoken at two rates, normal and fast (French Speaker R.L. without vocalic length contrast).

Fig. 4 : For the same utterances the VTT phase in the closure hold.

EFFECT OF PHONETIC CONTEXT AND TIMING ON THE F-PATTERN OF THE VOWELS IN CONTINUOUS SPEECH

Jacqueline Vaissière

Departement RCP
Centre National d'Etudes des Télécommunications
Lannion (France)

ABSTRACT

Very importantly for automatic phonetic decoding of continuous speech in speech recognition systems (SRS), the acoustic contrast (in terms of the observed formant -F- values) between phonemically distinct phonemes can be enhanced, or reduced or even neutralized depending on specific contexts. For a given vowel V, the F-pattern (i.e. F-values and F-temporal evolution), is not only a function of ideal F-target values for each V, but also of (a) the articulatory contrast (i.e. distance) in terms of the tongue position among the phonemes in consonant-V-consonant-vowel sequence and (b) the timing between the required opposite movements of the tongue. The first part of this paper illustrates the observations done on the three most open oral vowels in French, extracted from 1040 sentences spoken by five male speakers. It is concluded that temporal constraints (TC) on the motion of articulators become a prominent factor in continuous speech. The second part concerns the consequences of such observations for SRS: (a) the TC are effective in explaining many of the confusions appearing in the actual SRS, where decision about V identity is mainly based on measuring a "rate- and context-insensitive" acoustic distance between the segment to be recognized and a set of reference templates; (b) some suggestions are provided on how to introduce the TC in the design of the future SRS.

INTRODUCTION

In automatic phonetic transcription of speech, the decision about the identity of the vowels is generally based on the measurement of an acoustic distance between the spectra sampled in the vowel to be recognized and a set of stored (speaker-dependant) reference templates (one or more template for each vowel) is often erroneous. The error rate is known to vary greatly depending on the speaking rate, on the particular speaker and of the number of distinctive vowels in the language. For a language which has a fairly rich vocalic system as French, 50 % of errors is very common. The inadequacy of the speech parametrisation and a poor distance metrics may be partly in fault, but clearly others facts are in cause. As often noticed in the literature, phonemically distinct vowels are not always entirely separated in the F1/F2 plane, when extracted from continuous speech even for a single speaker. This paper explores the formant (F) values of the French three most open vowels in continuous speech, in a fairly large number of sentences.

EXPERIMENTAL PROCEDURES

Two-hundred eight sentences spoken by 7 speakers (5 males and 2 females) were recorded in a sound-proof room at CNET (France), Firstly, high quality digital spectrograms for the 1456 sentences were calculated. Secondly, about 200 selected sentences were submitted to the acoustic-phonetic analyzer developped at CNET [1], to automatically transcribe them into a set of phonemic hypotheses. Thirdly, the sentences were further analyzed using the Spire facilities at the Speech Group at MIT (USA). The sentences were automatically segmented [2] and the F-values were calculated from LPC spectra (See Fig.1 for an example of the material used). The present study only concerns the 1040 sentences sopken by the male speakers: the data obtained for the female speakers were disregarded because of problem in F1 detection.

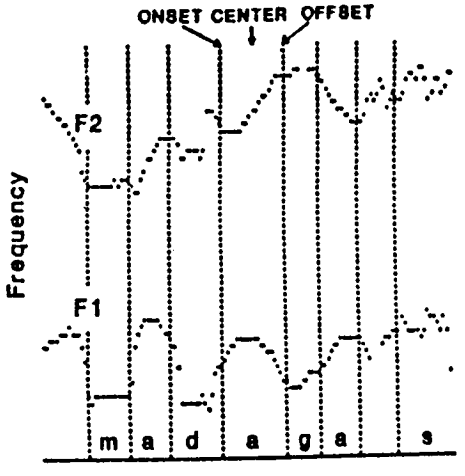


Fig. 1: MATERIAL USED
Display of the superimposed results of the segmentation program (vertical lines) and of the two first formants (F1 and F2) calculated from LPC spectra, every 5 msec.

I. RESULTS

I.1 TYPICAL PLACES OF ARTICULATION

The temporal course and the values of the F's in a V is influenced by the place of articulation of the surrounding phonemes [3]. The 16 French consonants and the three semi-vowels can be regrouped into the four traditionnal classes [4], depending on their places of articulation (See Fig. 2): (1) the bilabial and labio-dental, involving both lips or the lower lip in their production, (2) the dental, involving the tip or the fore part of the blade of the tongue, (3) the prepalatal, mediopalatal and velar, requiring an hightening movement of the dorsum of the tongue toward the hard palate and the velum, and (4) the uvular /r/ (backing of the tongue).

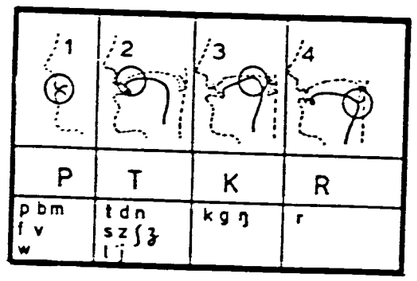


Fig. 2: THE FOUR CLASSES OF CONSONANTS
Classification of the 19 French consonants and semi-consonants into 4 articulatory classes (see text).

The letters P, T, K and R in this paper designate all the labial, dental, velar and /r/ consonants, respectively. A word like /Madagascar/ is thus considered equivalent (for what concerns the effect of the consonants on the vowel) to the sequence PaTaKaTKaR; the French words /femme/, /pape/, /bave/ to PaP, etc... V designates one of the three open vowels.

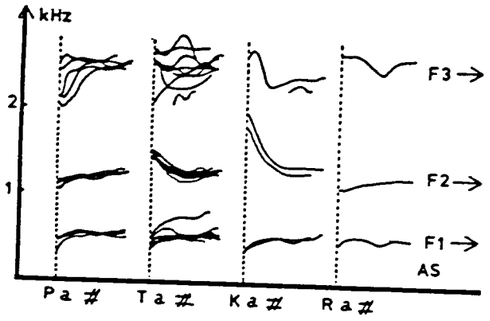


Fig.3: INVARIANT TARGETS
Display of the temporal course of the three first formants, in the vowel /a/, preceded by a labial, dental, velar or uvular consonant and followed by a pause, for the speaker AS. The data have been time-aligned on the vowel onset, found by the automatic segmentation program. The target F-values derived by visual inspection are indicated by arrows on the right part of the figure.

I.2 IDEAL TARGET IN LONG VOWELS

There is a well-known tendency, at least in controlled studies for non-sense sequences of speech, for the values of F1 and F2 of the vowels to be directed asymptotically toward the same target values. Such an ideal target is assumed to exist for each V, which is independant of consonantal context and thus can be rergarded as an invariant attribute of the vowel [5]. This tendency has been confirmed in our study of continuous speech, for vowels with relatively long duration (See Fig. 3 for illustration on /a/ lengthened because of prepausal lengthening).

I.3 THE EFFECT OF SHORTER DURATION

The F1 and F2 patterns for vowels with shorter duration exhibit the following characteristics.

a) Firstly, there is almost no reduction phenomena for V in labial context, even in short or very short occurences of V. In PVP context, the amplitude of the transitional F-movements from C to V and from V to C is small and the observed F-values all along V are closed to the F-target values. Compare the F-values at the F2 turning point in the sequences PaPa with the F-target values indicated by an arrow on Fig.4. Such an observation is expected by the fact that the lips movements for the production of the consonants does not interfere with the tongue movement for the production of the vowel.

b) The amount of reduction (in terms of difference between the F-target values and the observed values) can be predicted from articulatory considerations. When V and the surrounding consonants share about the same place of articulation (i.e. either anterior/front/dental or posterior/back/velar-uvular), the magnitude of F-transitions are reduced. In this case, the F-values observed at the F2 turning point are closed to the F-target values. It is the case for the front (anterior) /ɛ/ in dental (anterior) context (TɛT) and for the back (posterior) /ɔ/ in velar and uvular (posterior) context (KɔK and RɔR).

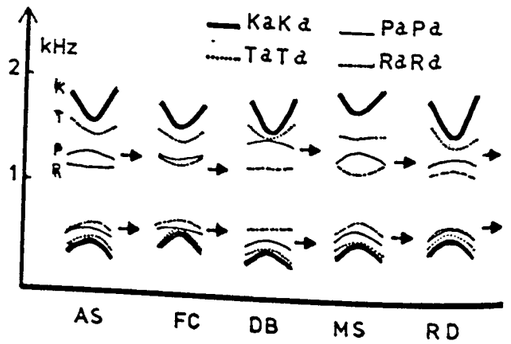


Fig.4: SYMETRICAL CONSONANTAL ENVIRONMENT
Range of the observed of F1, F2 and F3 variations observed in /PVPa/, /TVTa/, /KVKa/ and /rVra/ context for the 5 speakers. For all speakers, F2 is the highest and F1 the lowest in the velar context, and F2 the lowest and F1 the highest in the uvular context.

c) The F-target values are not reached in other cases. Great reduction in terms of differences between the observed F-values and F-target values has been observed for anterior vowel in posterior context (KɛK/ and posterior vowel in anterior context (TɔT). The shorter the vowel, the greater the amount of reduction. In very short occurences of TaT, for example, the values F2 at the vowel center are very closed to the values at the vowel onset and offset. (the "V" shape of F2 temporal course with appears in vowels with longer duration is almost reduced to a straigth line)

I.4 OPENNESS OF THE FOLLOWING VOWEL

The effects described above are not entirely sufficient to account for F-trajectories and values. For example, in a sequence like PaPi, where the closed vowel /i/ is following the open vowel /a/, and the intervocalic consonant is labial, two observations have been made. Firstly, there is always (for all speakers) a rising F2 (as the consequence of fronting of the tongue) Such a rise leads to higher F2 values in the mid region of the vowel, as compared to observed values in a sequence like PaPa. Secondly, there is some tendency of lowering of F1 (as a consequence of a lesser opening of the vocal tract). When the intervocalic consonant is dental (PaTi), the rising of F2 in /a/ is accentued in comparison with /a/ in PaTa.
The higher F2 values in /a/ at center and at offset when the following vowel is /i/ can be interpreted in two traditional ways; first, in terms of a palatalisation of the intervocalic consonant, secondly, as an anticipatory FRONTING-CLOSING due to vowel-to-vowel coarticulation [6].
More data are needed to check in a systematic way an eventual effect of anticipatory labialisation and velarization, on the three open vowels.

II. MODELLING THE INFLUENCES OF CONTEXT

Figure 5 schematizes our observations. The F-values observed at the onset, the mid region and the offset of a given vowel is the a function of the place of articulation of the adjacent consonants and the next vowel. This model is valid for the five speakers (the exact values of the F's are speaker-dependent).

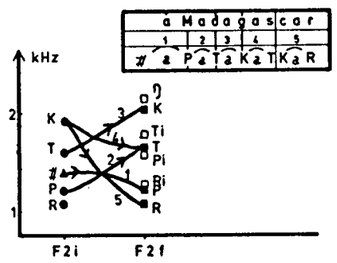


Fig. 5: MODELLING THE DIFFERENT INFLUENCES OF THE CONTEXT
F2 temporal course predicted for F2 during the vowel /a/, from F2 initial (F2i) to F2 final (F2f). For example, in the acoustic realisation of the sequence "à Madagascar" which is, on the articulatory equivalent to /aPaTaKaTkar/, F2 is sharply falling in kar (5), falling in gas (4), slightly falling in am (1), rising in mad (2) and dag (3), independantly of the speaker. When the following vowel is /i/ (as indicated by opened squares on the figure), F2f is higher in comparison with the cases where the following vowel is not /i/, as indicated by filled squares.

III. CONSEQUENCE FOR ASR

III.1 EXPLAINING PARTLY THE ACTUAL ERRORS

We have examined in an informal way the results obtained by the phonetic decoder of the KEAL system on a part of this sentences.and of a When the system is trained for the three V in their least reduced form (long V and PVPV context), good results are obtained in the test corpus for the V in similar contexts. Confusions often arise in different contexts, or in the cases the duration is shortened, in an expected way (See Fig.6). For example, the /a/ in a fast version of the word /quatre/ is generally identified as /ɛ/.

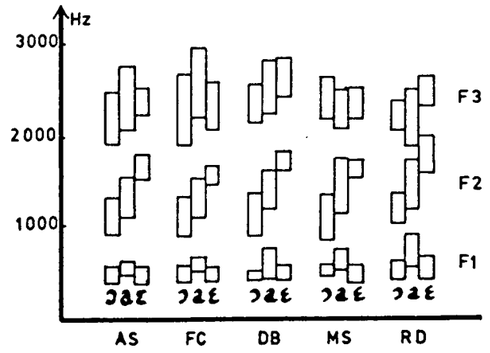


Fig. 6: RANGE OF F1, F2 AND F3 VARIATIONS
The vowels /a/, /ɔ/ and /ɛ/ are extracted from continuous speech. F1 tends to be higher for /a/, and F2 lower for /ɔ/ and higher for /ɛ/. There are, however, cases where the distinction between /a/ and /ɔ/ or /a/ and /ɛ/ is difficult to establish, at least from the formant values calculated on the spectrum sampled in the mid portion of the vowel.

In other terms, the results of the system depends directly on the particular allophones included in the training set, and of the duration of the vowels.

III.2 TRAINING THE SYSTEM

It is possible to create contexts maximizing or minimizing the coarticulatory effects [7,8]; The use of a very limited number of words for training allows to estimate, for each vowel and for each speaker, the range of possible variations for the F-values, at the center, but also at its onset and offest (See also [9]).

II.2 VERIFICATION RULES

The knowledge of physiological constraints is best to be put into the calculation of the fitness score between the output of an automatic decoding and the description of each word in the lexicon. Both the exact observed F-values and the description of the temporal course of the F during V (V shape, rising, falling, ...) are useful to verify the validity of an hypothesized CVCV phonetic sequences. For example, F2 in a sequence like /pak/ is obligatorily rising, independantly of the speaker (See Fig. 7): if the final consonant is not released, there is no burst and the F2 temporal course during the value (or the F2 value at the vowel offset) is the only way to distinguish between /pap/, /pat/ or /pak/. The acoustic distance between the vowel in the sequence /kak/, on one side, and the F-target values for the vowel /a/ and /ɛ/ should be expressed as a function of V duration: if V is less than 150 msec, it should be acoustically more closed to /ɛ/ than /a/.

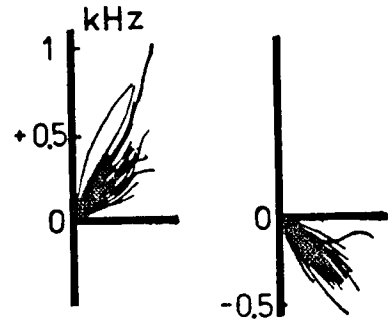


Fig. 7: OBSERVED F2 TEMPORAL COURSE DURING /a/
F2 temporal course during the vowel /a/ (five speakers plotted). On the right: PaK, PaT and TaK context, and on the left: TaP[-i], TaR[-i], and KaT[-i] context. The observed temporal course of F2 is in accordance with the model schematized in Fig. 5. The data have been aligned on the F2 value detected at the vowel onset.

CONCLUSION

First, for a general phonetic point of view, the study has confirmed for French and the usefulness of the notion of an invariant target for each vowel [5], the importance of the influence of the adjacent consonants on vowel formants [3], the existence of a vowel-to-vowel coarticulation phenomenum [6], and the importance of the time-factor for interpretating the formant patterns. Understanding of coarticulatory phenomena are essential in explaining the apparent variability in the acoustic realisations of a given vowel in continuous speech. It rises also questions as how the perceptual apparatus can deal with the distinction between phonemically distinct vowels which, through constraints due to the production apparatus, have become acoustically very closed in terms of F-pattern. The second conclusion concerns automatic speech recognition. Since there is exists a reasonably systematic relation between the F-pattern and the articulatory activity involved in the production, the systematic study of the F-pattern of the vowels in continuous speech provides a valuable tool for interpreting the limits of our actual phonetic decoders in terms of articulatory-acoustic relationship. Whether vowel identification is based on formant or not, the system has to deal in a proper way with the coarticulation phenomena in continuous speech.
Thirdly, it coins the usefulness of articulatory data concerning continuous speech. The possibility of obtaining articulatory data on large quantity of data and speakers [10] and the availability of a sound theory of speech production to interpret the link between articulatory data and the resulting acoustic signal [11] should lead to a proper theoritical framework for describing the range of possible acoustic variability for each sound. Such a framework is also needed to reduce the amount of training data necessary to adapt the systems to a new vocabulary or to new speakers.
Before such a concrete framework for describing coarticulatory phenomena is found, the knowledge can be only exploited in a rather ad-hoc manner, at the training level and/or at the level of phonetic acoustic decoder and lexical retrieval for a verification.

REFERENCES

[1] Mercier, G. and al, (1979), "The Keal speech understanding system", Proceedings of the NATO Advanced Study Institute, Simon J.C ed., 525-543.

[2] Leung, H. C., (1985), "A procedure for automatic alignement of phonetic transcriptions with continuous speech", Master of Sciences, Massachusetts Institute of Technology.

[3] House, A.S. and Stevens, K.N., (1963), "Perturbations of Vowel Articulations by Consonantal Context: An Acoustical Study", JSHR, 6, 111-128.

[4] Heffner, R-M, S., (1950), GENERAL PHONETICS, The University of Winconsin Press.

[5] Lindblom, B., (1963), "Spectographic Study of Vowel Reduction", JASA, Vol. 35, No. 11, 1773-1781.

[6] Ohman, S.E.G., (1966), "Coarticulation in VCV Utterances: Spectrographic measurements", JASA, Vol. 39, No. 1, 151-168.

[7] Abry, Ch. and Boe, L-J, (1984), "[i,a,u] Pas si fou? ou les lèvres des consonnes maximisent-elles l'espace vocalique des voyelles", 13èmes Journées d'Etudes du groupe de la communication parlée, Bruxelles, 28-30 Mai 1984, pp. 205-208.

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

[9] Vaissière, J., (1985), "The use of allophonic variations of /a/ in automatic cotinuous speech recognition of French", 108Th Meeting of the Acoustical Society of America, Austin, Texas.

[10] Kiritani, S., Itoh, K. and Fujimura, O., (1975), "Tongue-pellet tracking by a computer-controlled x-ray microbeam system", JASA, Vol. 57, No. 6, 1516-1520.

[11] Fant, G., (1960), ACOUSTIC THEORY OF SPEECH PRODUCTION, The Hague: Mouton.

ACOUSTIC CORRELATES OF REDUNDANCY AND INTELLIGIBILITY

SHERI HUNNICUTT

DEPARTMENT OF SPEECH COMMUNICATION AND MUSIC ACOUSTICS
ROYAL INSTITUTE OF TECHNOLOGY, BOX 70014 S-10044 STOCKHOLM

ABSTRACT

The relationship between context redundancy and key-word intelligibility was earlier examined in both high- and low-redundancy contexts. Word pairs were placed in similar positions in two sets of sentences: text-type sentence pairs and adages together with spoken-type sentences. With the text-type sentence pairs, there was an intelligibility advantage for the words in lower-redundancy contexts. For the adage and spoken-sentence pairs, there was no intelligibility advantage for words in either context.

An acoustic study has now been undertaken to determine whether differences can be measured in the production of the word pairs. The peak dB level of word pairs, their durations and the duration of various sub-word components, F0 maximum, range, excursion and contour complexity were considered as possible correlates. It was found that the correlation between any one of these factors and word intelligibility was quite low. However, it was also observed that in a pairwise comparison, differences in these factors occurred in the expected direction in a majority of cases.

INTRODUCTION

The purpose of this study has been to determine to what extent context redundancy and key-word intelligibility are related, and to examine possible acoustic correlates of these factors. It was intended that the results should help to answer some general questions about information control by talkers. For example, Lindblom /1/ has hypothesized that talkers naturally control the amount of information they wish to give to a listener. An ideal talker might attempt to increase the intelligibility of words that he expects his listener to have fewer "higher level" cues about. How, then, is this higher intelligibility brought about? There are several possible avenues: grammatical inversions, phonetic precision, prosodic stress. And, further, do these adjustments actually improve intelligibility for the listener?

A study which has been cited for many years as support for an inverse relationship between context redundancy and key-word intelligibility is one by Philip Lieberman /2/. In his study, Lieberman investigated the intelligibility of word pairs and examined the correlation of percent identification in these pairs with word duration, VU meter reading and relative peak amplitude. Since Lieberman's study was a small one, analyzing only seven words from eleven sentences, and since the redundancies in some of the word pairs were not well opposed, it was decided to replicate this work, considering more word pairs with more extreme redundancies.

INTELLIGIBILITY-REDUNDANCY STUDY

Test Materials

The initial study, in which subjects attempted to identify words extracted from sentence contexts, was described earlier (Hunnicutt, /3/). The words were taken from similar high- and low-redundancy sentence contexts containing matched pairs of words. One set of high-redundancy sentences were adages. The corresponding set of low-redundancy contexts were from grammatically similar sentences that might be spoken. The two other corresponding sets of sentences were rather long, grammatically standardized sentences which one might find in a text. All sentences were read by one speaker.

There were 76 text-type sentences recorded. These were taken from a set of Swedish test sentences developed for speech perception tests by Rolf Lindgren /4/. The test words in these sentences are all common words beginning with a stop and are in sentence-object position, far enough along in the sentence for a specific context to be built up. Redundancies for the sentences were established by having subjects fill in the blanks left by removing the test words.

The 36 Swedish adages were collected with the criterion that they contain a non-initial noun, preferably also non-final and of two syllables. The companion spoken-type sentences were constructed to have similar grammatical and syllable structure to those of the adages, with the test word in the same position. These sentences were constructed to give as little information as possible about the key-word, i.e., they were constructed as low-redundancy contexts.

Half of each type of the sentences were used as actual test material and half as fillers.

Speech Processing

The sentences were copied from tape onto a large computer disk where the test words were edited out using an interactive speech processing program /5/. A 2.1-second period of speech-like noise was added to cover each word, and the noise level chosen so that the average signal-to-noise ratio over the set of all words was 4 dB. The noise was increased from zero at the onset and attenuated at the offset over a period of 100

milliseconds to avoid an abrupt, and possibly distracting or confusing presentation. These words were then recorded on tape again with 4 seconds between words and 10 seconds after each set of 15 words. An open intelligibility test was then administered to 10 subjects.

Results

For the 19 text-type sentence pairs, there was a clear intelligibility advantage for the words in lower-redundancy contexts. There were 10 sentence pairs in which the low-redundancy words were more intelligible, 7 in which the low- and high-redundancy words were of the same intelligibility, and only 2 sentences in which the high-redundancy words were more intelligible. The mean number of correct answers for words in the low-redundancy contexts was found to be significantly greater using a paired-comparison test /6/.

There was no such advantage for words in lower-redundancy contexts for the 21 adage/spoken-sentence pairs, however. The number of sentence pairs in which the low-redundancy words were more intelligible was only 5, the number of sentences where the word pairs were of equal intelligibility was 9, and in 7 sentence pairs, the high-redundancy words were more intelligible. The mean number of correct answers were quite close, and the difference not statistically significant.

ACOUSTIC CORRELATE STUDY

An acoustic study has recently been undertaken to determine whether differences can be measured in the production of the word pairs, and whether observed differences could account for the previous results in word intelligibility. It was hypothesized that the more intelligible word of the pair should have some combination of the following attributes: higher intensity, longer duration, larger F0 excursion or more complex F0 pattern. A "perfect example" is shown for the word bollen below. This example, however, is not a norm, as will be seen in the text. ("Stress syll dur" is the heading for stressed syllable duration.)

BOLLEN

Redundancy	No. Correct	dB level	Stress syll dur	Word dur (sec)	F0 excursion (Hz)	F0 complexity
1.00	7	10.6	.31	.41	50	3
.24	10	11.6	.34	.52	117	3

Intensity

A comparison of peak dB levels for the word pairs in the text-type sentences revealed a tendency towards a lower dB level for words in the high-redundancy contexts, as expected. Of the 19 word pairs, 13 words in high-redundancy contexts had a lower dB value. One additional pair had equal dB-levels. A comparison of the adage/spoken-sentence pairs shows a lesser tendency in the same direction. Of the 21 word pairs, 12 words in high-redundancy contexts (57%) had a lower dB value. An additional three pairs had equal values.

A t-Test for related measures shows a significant difference in intensity of key-words in high- and low-redundancy contexts in both types of sentence pairs. Correlation between words correctly identified and dB-level is low, however, being .07 for the words in more redundant contexts and .18 for the words in less redundant contexts.

Duration

Duration measurements of the word pairs were made on the whole word, the stressed syllable (and stressed open syllables plus following consonants), the stressed vowel and the consonants(s) preceding the stressed vowel. It was found that the duration of the stressed vowel and the duration of the preceding consonant were often complementary. The two measurements were therefore combined into a prestress consonant plus stressed vowel duration. This measurement is the same as that for the stressed syllable in open syllables, of course. There were three cases in which one word of a word pair was longer in its stressed open syllable and the other word of the pair longer when the duration of the following consonant(s) were added, so these two measurements were both retained. This division effectively gives two sub-word measurements: one for the initial consonant(s) and vowel of the stressed syllable, and the second for these plus the following consonants, regardless of syllable division.

For the text-type sentences, there was a slight tendency towards shorter durations for words in the high-redundancy contexts, as expected. Of the 19 word pairs, 12 words in high-redundancy contexts had shorter initial consonant plus vowel, and 2 were of equal length. Eleven words in high-redundancy contexts had shorter initial consonant plus vowel plus following consonant(s). Only 10 entire words were shorter, with 2 of equal length. All t-Tests of durational differences show small differences of rather low significance.

For the adage/spoken-sentence pairs, the results were the opposite: there was a slight tendency towards longer durations for words in the high-redundancy contexts, contrary to expectations. Of the 21 word pairs, only 8 words in high-redundancy contexts had shorter initial consonant plus vowel, and 1 was of equal length; 12 words had longer initial consonant plus vowel. Seven words in high-redundancy contexts had shorter initial consonant plus vowel plus following consonant(s), 3 were of equal length and 11 were longer. Only 9 entire words were shorter, 2 were of equal length, and 10 longer. Except for a nonsignificant difference in word duration, all t-Tests of duration differences show small differences of rather low significance. Correlations between the number of words correctly identified and each of the duration measures is fairly low for both types of sentence pairs. The highest correlations were with word duration in the adage/spoken-sentence pairs (.40) and with stressed syllable duration in the text-type sentence pairs (.33).

F0 movement

Four measures of F0 movement have been employed: maximum F0 value, F0 range and total F0 excursion during a word, and F0 contour complexity. Total excursion is taken to be the sum of rises and falls (of at least 6 Hertz each), and contour complexity is expressed as the number of such rises and falls. Whereas intensity and duration measurements showed greater differences in the expected direction in the text-type sentences, F0 measurements show greater expected differences in the adage/spoken-sentence pairs. All four F0 measures show decided greater values for spoken-type low-redundancy context sentences compared to adages, as expected. Contrary to expectation, however, measurements were often smaller for words in low-redundancy context text-type sentences compared to high-redundancy contexts.

For text-type sentences, maximum F0 value is larger for words in low-redundancy contexts in 11 of the 19 cases, and total F0 excursion is larger in 10 of 19 cases. However, F0 range is larger for words in low-redundancy contexts in only 7 cases, and the number of rises and falls is larger in only 3 cases, 10 words having an equal number.

In the adage/spoken-type sentence pairs, maximum F0 value is larger for words in the low-redundancy spoken-type sentences in 13 of 21 cases, F0 range is greater in 17 cases, total excursion greater in 18 cases and number of rises and falls greater in 8 cases, 9 words having an equal number.

GENERAL COMPARISON

In Table 1, shown below, an attempt has been made to condense the preceding information and to give a comparative overview. A check mark in the table indicates that the data supports the expected result of more correct identifications of a word or more marked prosody for a word in a low-redundancy context; an "x" indicates the opposite. An equals sign (=) indicates equivalent data and a percent sign (%), ambivalent data. A check mark appears in the first column if the word in the less redundant context was identified better (as expected), in the second column if it had a higher dB level, in the third column if it had a longer duration, and in the fourth column if it exhibited greater F0 movement (indicating that at least 3 of the 4 F0 measurements are greater).

Summing the check marks in the table for text-type sentences, we see that the word in the low-redundancy context was identified better in 53% of the cases. We also see that 53% of the possible prosody cues are present to support identification (i.e., of the 57 possible check marks, 30 are present). Summing x's, we also see that there were 26% contradictory cues, i.e., lower dB-level, shorter duration or less F0 movement in the low-redundancy context. Only 11% of high-redundancy context words were better identified.

The number of prosody cues for the adage/spoken-sentence pairs is quite similar to the text-type sentences, even though identification was not, as previously reported. Words in low-redundancy contexts were better identified in only 24% of the cases, and words in high-redundancy contexts, 38%. However, we see that 54% of the possible prosody cues are present to support better identification of words in low-redundancy contexts. Contradictory cues were present in 24% of the cases.

	<Redund. >Correct	<Redund. >dB	<Redund. >Dur	<Redund. >F0 Mvmt
TEXT-TYPE SENTENCES				
BOKEN	√	√	√	√
BOLLEN	√	√	√	√
BORDEN	√	√	√	√
DÖRREN	X	√	√	%
GOLVET	=	√	√	X
KAMMEN	√	=	%	√
KARTAN	√	X	X	X
KLOCKAN	X	X	X	X
KNIVEN	√	√	%	√
PENNAN	=	√	%	%
TAKEN	√	X	%	%
TAVLAN	=	X	√	X
BREVEN	√	√	√	√
DUKEN	√	X	%	X
HANDDUKEN	=	√	X	%
SKÅPET	=	√	√	√
TRAPPAN	√	√	%	X
ÄDELSTENEN	=	√	√	X
BILEN	=	√	√	%
ADAGE/SPOKEN-TYPE SENTENCES				
FÅGEL	√	√	√	√
HANDEN	√	√	X	√
SKOGEN	X	√	%	X
BÄCKEN	=	X	%	√
SOMMAR	=	√	%	√
KORVEN	=	√	√	X
BÖRJAN-1	X	√	√	√
BÖRJAN-2	=	X	%	√
KVARNEN	X	X	X	√
TJUREN	X	√	√	√
GRYTOR	√	√	√	√
KVASTAR	=	√	X	√
HUNDEN	√	X	X	%
VAGNEN	X	=	X	√
RÄVEN	X	X	√	%
KATTEN	X	√	X	X
BORDET	=	√	%	√
SKINNET	√	√	%	√
BJÖRNEN	=	=	%	√
DÅRAR	=	=	%	√
VATTEN	X	X	X	√

TABLE 1. A GENERAL PROSODIC COMPARISON

Combining prosody cues for all 120 words, it can be seen that expected cues appear in 53% of the cases, contrary cues in about half so many, 26%. One observation supportive to the hypothesis is that at least one contrary cue appears in 8 of the 10 words in which identification was also contrary to expectation. An examination of the recordings and spectrograms of the two remaining word pairs reveals nothing of a segmental quality which could have caused these identification results. It should also be noted that there are 3 word pairs in which the word in the low-redundancy context was better identified even though 2 or 3 contrary prosodic cues were present. No particular segmental effects have been noted for these words either.

CONCLUSIONS

An earlier study investigated the relationship of a word's intelligibility to the redundancy of its context. This study found an intelligibility advantage for words in lower-redundancy contexts in text-type sentences. For adage and spoken-type sentence pairs, however, there was no such advantage for words in either the low-redundancy spoken-type sentences or the high-redundancy adages. It was conjectured that the metaphorical nature of adages and the influence of the social conditions in which they are used may have prevented the realization of the intelligibility-redundancy relationship observed in the text-type sentences.

The current study examined possible prosodic correlates of redundancy and intelligibility. The peak dB level of word pairs, their durations and the duration of various sub-word components, F0 maximum, range, excursion and contour complexity were considered as likely correlates. It was found that the correlation between any one of these factors and word intelligibility was quite low. However, it was also observed that differences in word intensity, duration and F0 movement in a pairwise comparison occurred in the expected directions in a majority (53%) of cases. It was also observed, however, that in 26% of cases, differences occurred in the direction opposite to that expected. An examination of particular cases revealed that 8 of the 10 word pairs in which intelligibility was contrary to expectation also exhibited at least one contrary prosidic cue. In addition, there were 3 word pairs in which intelligibility was as expected even though 2 or 3 contrary prosodic cues were present. Segmental quality did not seem to be a factor.

We might say that if prosody indeed correlates with intelligibility, that it correlates as a whole, being expressed in various combinations of higher intensity, longer durations and more lively F0 contour. Correlation of these prosodic cues with lower context redundancy was somewhat more robust, especially in the case of F0 movement. This suggests that a talker is not always successful in improving word intelligibility for his listener in a low-redundancy context, but that subtle differences in performance intended to bring about this effect are often present.

REFERENCES

/1/ Lindblom, B., The Interdisciplinary challenge of speech motor control. In S. Grillner, B. Lindblom, J. Lubker and A. Persson (eds.), Speech Motor Control, Oxford, pp. 3-18, 1982.

/2/ Lieberman,P., Some effects of semantic and grammatical context on the production and perception of speech. Language and Speech,6,172-187, 1963.

/3/ Hunnicutt,S., Intelligibility versus Redundancy - Conditions of Dependency. Language and Speech, 28,47-56, 1985.

/4/ Lindgren, R., Informationsteoretiskt viktade testsatser för talperceptionstest. Joint publication of Karolinska Institutet and Stockholms Universitet, Stockholm, 1982.

/5/ Rolf Carlson, MIX, Dept. of Speech Communication and Music Acoustics, Stockholm, 1984.

/6/ Box, G., Hunter, W. and Hunter, J., Statistics for Experimenters, New York, 1978.

A PROBLEM OF ASSIMILATION BETWEEN NASAL VOWEL AND PRECEDING NASAL CONSONANT, A PERCEPTUAL EXPERIMENT

PIETER VAN REENEN

Dept. of Language
Free University, Amsterdam
The Netherlands

HANS VAN DEN BERG

Dept. of Language
Free University, Amsterdam
The Netherlands

ABSTRACT

On the basis of mainly articulatory data it was concluded in Van Reenen [1] that a nasal vowel consists of a nasal part preceded by an oral part. The results of the experiment reported below show that the increase in nasality from the oral to the nasal part is relevant for the perception of the vowel as nasal, even when the vowel is preceded by a nasal consonant. The result explains a paradox with respect to assimilation.

THE PROBLEM

Assimilation

When two phonemes (or underlying segments) are realized adjacently, assimilation may occur, i.e. the realizations of these phonemes (or segments) may become more like each other. When the underlying sequence of phonemes (or segments) /na/ is perceived as [nã], the /a/ has been assimilated to the /n/ with respect to nasality, a case of progressive assimilation. However, if the sequence /na/ is perceived as [na], there is no (or hardly any) assimilation. This is essentially the view of assimilation in structural and generative phonology.

In diachronical studies it is commonly claimed that if [na] > [nã] (whatever the underlying segments) there has been progressive assimilation. However, this kind of change is not common. Much more common is the change [an] > [ãn] > [ã] or [nan] > [nãn] > [nã]. The formation of the [ã] is usually the consequence of regressive assimilation by the following nasal consonant. However, no explanation is provided for the asymmetry in behaviour of the vowel preceded or followed by nasal consonant.

In both views is it assumed (although often implicitly) that assimilation concerns both coarticulation and coperception. If the [ã] in [nã] is more like the [n] than the [a] in [na] the resemblance concerns both the position of the velum and the nasal quality of the two sounds. We will refer to these views as the Assimilation Hypothesis.

In the following we will provide an explanation why there is an asymmetry between the nasalisation of a vowel preceded and a vowel followed by nasal consonant. In particular we claim that with respect to articulation the [ã] in [nã] has not become more like the [n] than the [a] in [na], although in both forms some coarticulation occurs.

The articulatory structure of a nasal vowel

In Van Reenen [1] it was found that vowel nasality is better defined in terms of the articulatory notions nose coupling and mouth coupling than in terms of nose coupling alone. The amount of nose coupling N was defined as the opening in mm^2 of the nasal port measured in a cross-section perpendicular to the airstream at the point of greatest constriction between the velum and the pharyngeal wall. The amount of mouth coupling MC was defined as the opening in mm^2 of the mouth passage, measured in a cross-section perpendicular to the airstream at the point of greatest constriction in the mouth. Articulatory evidence showed that rather than in terms of N the nasality of vowels may be expressed in terms of the proportional relationship N% between N and MC, as in formula (I):

$$N\% = N/(N+MC) \cdot 100\% \qquad (I)$$

It follows from (I) that an increase in N% is produced by means of an increase in N and/or a decrease in MC.

A second finding in Van Reenen [1] was that there is an increase in N% during the production of a nasal vowel and that this increase in N% is an intrinsic property of this vowel. This conclusion was mainly based upon the fact that the lack of N% in the first part of the vowel appeared to be not simply a case of coarticulation with a preceding nonnasal consonant, since the increase in N% was present in nasal vowels preceded by a nasal consonant as well. Between the central phase of the nasal consonant - during which MC=0 and N% is 100%, see formula (I) - and the central phase of the nasal vowel, a dip in the amount of N% occurred. In terms of N this dip was sometimes present as well. In other cases there was an increase of N from the end of the central phase of the nasal consonant to the central phase of the vowel. This increase in N may correspond to a dip in N% as well, on the assumption that the MC of the vowel is arrived at early in the vowel and is more or less steady until its central phase. We will refer to this conclusion as the Increase in N% Hypothesis.

A paradoxical result of a perceptual experiment

A perceptual experiment carried out by Linthorst [2] provided some indirect evidence that the increase of N% during the nasal vowel was perceptually relevant. In the experiment, Linthorst manipulated the vowel

in French même "self" [mɛm] to which we will refer as [e_n], since it may be assumed that it was produced with a considerable amount of NZ from the beginning to the end, and the vowel in words like baie "bay" [be] to which we will refer as [e_o], since it may be assumed that it was produced with NZ being 0. Linthorst found that speakers of French perceived the vowel [e_n] as nonnasal. In one speech sample offered to the listeners the second [m] had been cut off as in [me_n] and in another the vowel [e_n] had been added once again as in [me_ne_n]. However, when the same vowel [e_n] was added to [be_o] as in [be_oe_n] the speech samples were perceived as containing nasal vowels, although in the first part of [e_oe_n] it may be assumed that NZ was 0.

In order to account for the results found by Linthorst, the following explanation was proposed in Van Reenen[1]. On the one hand, the increase in the amount of NZ during the [e_oe_n] preceded by **oral** consonant made the listeners perceive the vowel as nasal. On the other hand, the more or less constant amount of NZ in [e_ne_n] and [e_n] preceded by **nasal** consonant, made the listeners perceive [e_ne_n] as nonnasal. In other words, the listeners interpreted implicitly the off-glide of the [m] as still present during the articulation of the vowels [e_n] and even [e_ne_n].

A first experiment
In order to test this explanation a preliminary experiment was carried out in which a direct link was laid between articulatory properties of nasal vowels and their perception. Artificial nasal vowels were produced by means of a plexiglass model in which NZ could be calculated on the basis of the articulatory properties N and MC. Listeners judged vowels in which NZ increased as being more nasal than vowels in which NZ was equally distributed over the vowel for the same amount of NZ. Since the result showed that the increase in NZ is more important for the perception than the total amount of NZ, the increase in NZ hypothesis was confirmed. The details of the first experiment are presented in Van Reenen and Groen [3].

Implication
The explanation in Van Reenen [1] of the experiment in Linthorst [2] has an implication which can be tested. If it is correct, it follows that a nasal vowel preceded by a nasal consonant should contain an increase in NZ in order to be perceived as nasal. A perceptual experiment was devised in which this implication was examined in order to make a choice on this point between the Assimilation Hypothesis and the Increase in NZ Hypothesis.

Figures 1 through 4 illustrate the two hypotheses to be tested from an articulatory point of view. In these figures C=oral consonant, N=nasal consonant, V=oral vowel, $\tilde{v}$=nasal vowel. The amount of NZ is represented from the end of the central phase of the consonant until the end of the central phase of the vowel.

With respect to perception the two hypotheses predict that [$\tilde{v}$] in [C$\tilde{v}$] will be perceived as nasal and [V] in [NV] as oral (see figures 1 and 2). Figures 3 and 4 represent forms of coarticulation. Instead of following down the broken lines, NZ stays high in figure 3. Figure 4 represents another, less outspoken form of coarticulation. Here NZ decreases until it crosses the line of increase of the vowel. According to the Assimilation Hypothesis figure 3 represents [N$\tilde{v}$], whereas figure 4 is nonexistant. According to the Increase in NZ Hypothesis figure 3 represents [NV] and figure 4 [N$\tilde{v}$]. The experiment reported below shows that figure 4 represents [N$\tilde{v}$], thus providing confirmation of the Increase in NZ Hypothesis.

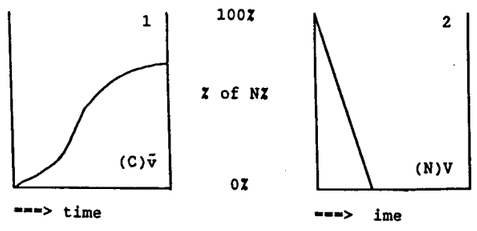


Fig. 1 and 2. Schematic representation of the amount of NZ on the basis of articulatory data.

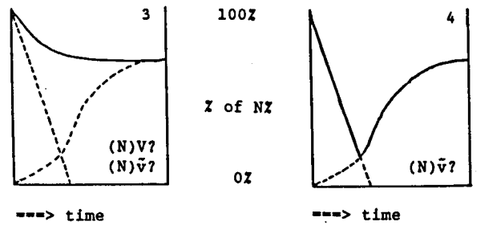


Fig. 3 and 4. Schematic representation of the amount of NZ during the vowel on the basis of articulatory data.

THE EXPERIMENT

Test words
In the experiment the properties of the vowels of four Dutch words of type [Cvs] and [N$\tilde{v}$s] were tested perceptually. The words are:

[pɑ̃s, pɪ̃s, mɑ̃s, mɪ̃s]

They occur in low standard Dutch pronunciation, in which the [n] - present in the standard language forms - is absent and the vowel is nasal. The words with [ɑ̃] are possible French words as well.

Speech samples
The series of four words was pronounced five times by the first author and registered on Revox A77 by the second author. The words were pronounced slowly and clearly, as much as possible at the same frequency (150 Hz) and equally loudly. The most regular set was selected for further treatment.

The [p], [m] and [s] were cut off the words by means of the computerprogram SESAM. As the borderline between [p] and vowel we took the transition from non-periodic to periodic. As the borderline between [m] and vowel the change in amplitude. In the case of [p] and [m] these borderlines correspond almost exactly to the end of the central part of the consonant during which the mouth canal is blocked. As the borderline between the vowel and [s] we took the transition from periodic to non-periodic. The vowel length of the two [ɑ̃] sounds was 420 ms, of the [ɪ̃] in [pɪ̃] 368 ms, in [mɪ̃] 394 ms.

For the [p$\tilde{v}$] words we expected a perceptual score corresponding to figure 1. These items served as a control. The scores with respect to the vowels in [mɪ̃s] and [mɑ̃s] made it possible to check the reality of figure 4.

Five times two periods were selected from each of the four vowels. The two periods were chosen at 0, 50, 100, 150 and 200 ms from the beginning of the vowel. Since a period has a length of 8 ms the selected vowel periods were situated at 0-16 ms, 50-66 ms, 100-116 ms 150-166 ms and 200-216 ms. We may assume that the nasality during each selected vowel fragment of two periods was (almost) constant.

Each fragment of two periods was lengthened to a vowel of 196 ms by means of a computer program (GHEVU). These vowels will be called: 1 to 5, where 1 refers to the first, and 5 to the last selected vowel segment. We refer to the artificial vowels as (p)V_3, (m)i_5 i.e. the third lengthened vowel segment in the word starting with [p], and the fifth segment of [i] preceded by [m].

The ten artificial [ɑ̃] sounds were grouped into pairs and the pairs were put on tape in random order. Each pair was compared to itself and to the other [ɑ̃] sounds, just as the ten artificial [ɪ̃] sounds. The test is of the type A-B and B-A. So, for instance, [(p)i_1] formed a pair with [(p)i_4] and [(p)i_4] formed a pair with [(p)i_1]. The test started with ten supplementary pairs which did not count in the score.

Subjects
Altogether 33 subjects listened to the tape, 18 of whom were known to be trained listeners. They were asked to score which of the two vowels was more nasal. The results have been calculated for the two groups separately and together. Since we were not interested in the question whether Dutch listeners are able to hear vowel nasality, but in the question whether listeners who are able to perceive vowel nasality judged our artificially lengthened vowels, we consider the results of the trained group as more relevant and report them in detail. The data of the untrained listeners - showing the same trends but less clearly - are presented in an Appendix.

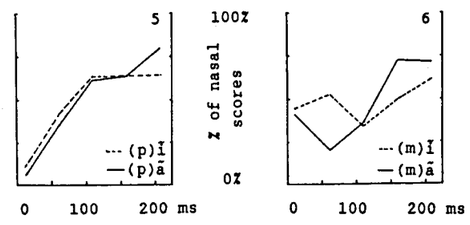


Fig. 5 and 6. Nasal scores by trained listeners.

Results
The results of the trained listeners are reproduced in figures 5 and 6 as percentages of more nasal scores. They show that trained listeners hear the vowels preceded by [p] as consisting of a minimal nasal part, followed by increasingly more nasal parts. In the case of [(p)i_3], [(p)i_4] and [(p)i_5] no increase in nasality was perceived. Apparently the maximal nasality is reached early in this vowel, earlier than in the [ɑ̃].

The [(m)$\tilde{v}$] vowels offer a different picture. During the central part of the nasal consonant NZ is 100%. As soon as the vowel starts (in terms of periodic waves or increased amplitude) the perceptual correlate of NZ decreases sharply until the second vowel to increase sharply again until the 5th. In the case of [ɪ̃] there are two increases at the beginning of the vowel. The second and most serious increase is situated later in the vowel than in the three other words.

Agreement of scores
In order to check the agreement of the scores of the trained listeners, the Friedman rank analysis was applied. The higher the X^2, the better the agreement among the listeners. As appears from table 1 (for the untrained listeners see Appendix), the agreement of the scores of the trained listeners was high. It was always significant at the .01 level, and, except in one case, at the .001 level. We may conclude that the scores of the trained listeners are highly reliable.

	X^2	Df	a=.05 at	a=.01 at	a=.001 at
(p)ɑ̃	51.19	4	9.49	13.28	18.47
(m)ɑ̃	33.30	4	9.49	13.28	18.47
(p)ɑ̃+(m)ɑ̃	94.06	9	16.92	21.67	27.88
(p)ɪ̃	38.52	4	9.49	13.28	18.47
(m)ɪ̃	14.74	4	9.49	13.28	18.47
(p)ɪ̃+(m)ɪ̃	95.34	9	16.92	21.67	27.88

Table 1. Agreement of the scores of trained listeners in terms of the Friedman rank analysis. Agreement was always significant at the .01 level, and, except in one case, at the .001 level.

DISCUSSION

Confirmation of the Increase in NZ Hypothesis
Although we have not examined the question whether the speech samples were articulated with an increase in NZ during the nasal vowel, the way the listeners perceived the succeeding parts of the original nasal

vowels suggest that such an increase was present. The perceptual results represented in figure 5 closely parallel the articulatory representation in figure 1. Apparently the increase in NZ brought about by the amount of N and MC is perceived as an increase in nasality. The perceptual results in figure 6 closely parallel the articulatory representation in figure 4. Apparently, the dip in NZ between the nasal consonant and the central part of the vowel is perceived as a relative lack of nasality. It would be difficult to imagine how these perceptual results correspond to an articulation without such a dip. Therefore we consider the perceptual results represented in figure 6 as evidence in favour of the Increase in NZ Hypothesis. It follows that both figure 2 and figure 3 represent [NV].

It may be the case that the results found do not even represent the dip in its most outspoken form. As we have seen above, the vowels 1 through 5 were produced on the basis of 16 ms samples from the original vowel. In between every two succeeding samples fragments of 50-16=34 ms were not examined. It is possible that if the two periods had been chosen from these 34 ms, the NZ and the perceived nasality might have been even (slightly) less than what has been found. This would have made the results even more convincing.

Differences concerning [Ĩ] and [ã]

There is less agreement among listeners concerning [(m)Ĩ] than concerning [(m)ã] and on an average the increase in NZ in the [Ĩ] is less clearly perceived than in the case of [ã]. Apparently, the role of the dip is less important in the case of [(m)Ĩ]. We have have seen that in the case of [(m)Ĩ] the dip falls later in the vowel than in the case of [(m)ã]. Does the amount of nose coupling N during the first part of the [Ĩ] ([(m)i_1]) increase slightly? The palatoglossus - which is responsible for the closing movement of the velum - is a relatively slowly reacting sphincter and since MC is relatively low the influence of such a slow reaction may be noticeable where it is not in the case of [ã].

Assimilation, coarticulation and coperception

A nasal vowel preceded by a nasal consonant will not easily become nasal. The NZ will be interpreted as part of the off-glide of the nasal consonant. Even if a nasal vowel is formed by means of a dip between the central phase of the nasal consonant and the central phase of the nasal vowel, this vowel is not as easily perceptible as a nasal vowel in isolation, since the increase will be less outspoken.

Coarticulation between [m] and [V] implies that the vowel is produced with a considerable amount of NZ during its first part, which may go on during its central phase. The result is perceived as an oral vowel. In order to perceive such a vowel as nasal this coarticulation should be avoided and a dip in NZ should be created.

If assimilation is considered to concern both coarticulation and coperception, we may conclude that the succession of a nasal consonant and a nasal vowel is characterized by coperception but not by coarticulation. Coarticulation between nasal consonant and following nasal vowel is not more outspoken than between nasal consonant followed by an oral vowel.

If this conclusion is accepted there is no specific coarticulation between a nasal consonant and a following nasal vowel. This would explain why in language change progressive assimilation of a nasal vowel is less common than regressive assimilation.

APPENDIX

Data concerning the group of untrained listeners.

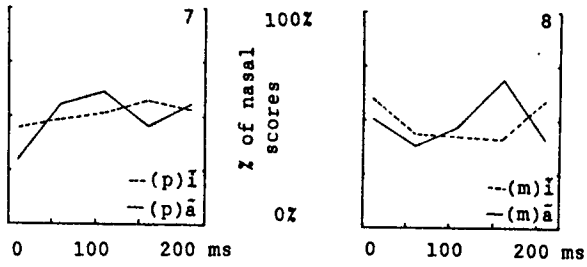


Fig. 7 and 8. Nasal scores by untrained listeners.

	X^2	Df	a=.05 at	a=.01 at
(p)ã	15.56	4	9.49	13.28
(m)ã	9.67	4	9.49	13.28
(p)ã+(m)ã	33.89	9	16.92	21.67
(p)Ĩ	4.04	4	9.49	13.28
(m)Ĩ	13.48	4	9.49	13.28
(p)Ĩ+(m)Ĩ	15.63	9	16.92	21.67

Table 2. Agreement of the scores of untrained listeners in terms of the Friedman rank analysis. The agreement among untrained listeners was less than among trained listeners, although, except one case, significant at the .05 level and, except two cases, at the .01 level.

ACKNOWLEDGEMENTS

For the experiment we have used the equipment of the Institute of Phonetic Sciences, Amsterdam. We thank Daan Broeder, Laurens Damen, Emiel Kappner, Florien Koopmans and Louis Pols for their help in setting up and carrying out the experiment.

REFERENCES

[1] Reenen, P. Th. van, **Phonetic feature definitions, Their integration into phonology and their relation to speech, A case study of the feature NASAL**, Dordrecht/Cinnaminson: Foris, 1982.

[2] Linthorst, P., **Les voyelles nasales du français, Etude phonétique et phonologique**, Thesis University of Utrecht, 1973.

[3] Reenen, P. Th. van and M. Groen, **The relation between the articulatory properties of nasal vowels and their perception**, in prep.

HIERARCHY OF LEVELS IN SPEECH PERCEPTION

V.B. KASEVICH, Y.M. SHABELNIKOVA

Dept. of Oriental Studies
University of Leningrad
Leningrad, USSR 199034

ABSTRACT

Speech perception is argued to be essentially a top-down process coming down stepwise from higher levels to lower ones, the higher levels being characterizable, from the phonological point of view, in terms of their prosodic features.

In Donald Norman's words, "it is usually thought that the analysis of speech requires levels of abstraction. For example, speech sounds are transformed into phonemes, and phonemes into words"/9, p.388/. The analysis of this kind, nicknamed 'bottom-up', until very recently was accepted almost universally. The only alternative was presented by the one-time influential analysis-by-synthesis model which typically did not make use of the notion 'level of abstraction'. In other words, the predominant views link the very notion of levels to the more traditional bottom-up model, while its top-down counterpart, the analysis-by-synthesis model, is not thought to require the notion. The reason seems to be rooted in a rather narrow understanding of what levels of abstraction may be like: it is believed by many that at the outset of the process of speech perception man deals with the phonetically 'richest', i.e. the least abstract, characteristics of the incoming signal, the latter being step by step 'refined' so as to dispose of communicatively irrelevant details.

Yet the psychology of vision /II/ as well as the still earlier ideas of N.A. Bernstein /I/ suggest a valuable hint to the contrary. It is argued that at the first stages of visual perception man deals with highly generalized, and therefore abstract, features of the object to be perceived. Such features are sufficient to get a 'rough idea' of what is being seen, i.e. to assign the object to a very broad class of entities. If the actual setting is informative enough to provide ground for (subliminal) choice within the class, the object is recognized with all its relevant details without further analyzing its actual physical characteristics. If not, its lower-level features, less abstract and more specific, are to be brought into consideration until the information is made sufficient to identify the object.

It may be seen from the above very sketchy exposition that visual perception exhibits a clearly top-down character. On the other hand, it is hardly in line with the analysis-by-synthesis model -- at least not beyond the anticipation routines common to all perception strategies. It seems to be of primary importance that the perceptual process evolves as a stepwise progressing from a more abstract representation of an object to a more specific (concrete) one. That means, at the same time, that the process is hierarchical in nature. Levels of abstraction are also levels of control where the output of a higher level largely constrains and, for that matter, controls the functioning of the lower one.

If auditory perception is presumed to be essentially parallel to visual perception, then we may accordingly seek similar stages in processing the sound information. One of the crucial problems is singling out particular sound features to be assigned to higher levels of speech perception. Since in processing the speech signal the listener aims from the very outset at grasping its meaning, the features sought should be applicable to as big speech chunks as possible. Clearly such features are most likely to be p r o s o d i c (suprasegmental), i. e. pertaining to intonation, stress (accent) or tone.

One possible method of investigating the relative role of suprasegmentals and segmentals (syllables, vowels, consonants) in speech perception is artificial distortion of certain acoustic parameters responsible for realization of particular segmentals or suprasegmentals, which gives an opportunity to see their contribution to the process. For instance, Price and Levitt /IO/ report that insufficient prosodic information makes the /š/ - /č/ distinction unstable. They speak of pro-

sody as of "an aid in initial parsing of a sentence" (p. 302). In our view, such data point to an hierarchically higher role of prosody as compared to that of segmentals.
Bosshardt /3/ has observed that if parts of test sentences are interchanged (cf. Der Student schreibt seine Arbeit in den kalten Dachstube → In den kalten Dachstube der Student schreibt seine Arbeit → Seine Arbeit der Student schreibt in den kalten Dachstube) the changes affect comprehension because of the "perturbations in the suprasegmental information" (p.193).
Krulee et al. /8/ argue that the prosodic system is "responsible for the segmentation of continuons speech into sentences, phrases, and words. It attempts to establish a context within which a second system, responsible for the processing of words and syllables, can operate"(p.531). This, again, is tantamount to recognizing an hierarchy of levels in speech perception, albeit somewhat narrowly understood, for the interaction of prosody and "a second system" is reduced to segmentation and identification respectively.
In what follows we sought to bring to light the relative role of suprasegmentals and segmentals by experimentally changing the acoustic parameters responsible for their natural relationship.
The present writers' previous findings obtained for Chinese and Vietnamese /6; 7/ have shown that syllable tones and segmentals behave differently in the process of speech perception. When 'monotonized' by means of vocoder technique, i.e. deprived of the fundamental frequency differences responsible for the identity of lexical tones, the syllables are recognized considerably worse. The confusion of our 'toneless' syllables taken at its face value would be a trivial fact, for tones are known to be unalienable features of each syllable in Chinese or Vietnamese. Yet, less trivial is the quite consistent confusion of the test syllables' s e g m e n t a l components whose acoustic characteristics were kept intact in our experiments. On the other hand, white-noise masking(signal/noise ratio 0 dB), while resulting in a drop in recognition scores for segmentals, practically does not impair those for tones. The latter results, althogh quite natural, seem to be of crucial importence: in their absense one might be tempted to argue that any interference with the speech signal, whatever its nature, may lead to a poor recognition of the signal in its totality. However, our data seem to indicate that the recognition of tones is a prerequisite to that of segmentals -- but not the other way round. In other words, there must be an hierarchy in the processing of segmentals and tones which consists in the following: syllables are classified into tonal categories and the identification of segmentals is carried out by the listener within thus previously determined tonal classes of the syllables. When this natural process is made impossible because of inavailability of the tonal information, the listener finds himself at a loss facing the necessity of discriminating between all the possible syllables instead of using the overlearned strategy of operating within a tonally delineated class. Hence the multiple confusions reflected in our data.
The general aim of the experiments reported in this paper is to find out whether the process of speech perception in stress languages such as Russian can be visualized as structurally similar to what has been observed for tone languages. If the hierarchical relationships of perceptual strategies for stress languages do parallel those observed for tone ones, then the place of syllable tones of the latter should be taken up by word stress patterns of the former, for functions of the two are alike: in tone languages a tone marks prosodically a syllable-morpheme, the basic operative unit of the languages /5/, whereas in Russian-type languages it is the word prosodically patterned by means of stress that is the basic unit of a functionally similar type.
For obvious reasons, it is more difficult to artificially deprive syllables of a stress-language word of their prosodic features than to make Chinese or Vietnamese syllables 'toneless'. That is why we resorted to a different experimental technique, viz.: the final and the initial unstressed.CV syllables of two conjoined two-syllable words spoken without a pause were cut out so as to make one pseudo-word with both its syllables unstressed, e.g. X'IMU out of the sentence /duX'I MUžej n'e pugal'i/ 'The ghosts did not frighten our husbands'. As a result, we obtained pseudo-words composed of two stressless syllables. Similar pseudo-words, this time trisyllables, were obtained by means of mutual (crosswise) transplanting of the syllables, e.g. out of the two sentences: /a vot japonskaja p'evica cús'ita/ and /a vot japonskaja p'evica cus'íta/ 'Here is Tsushita, the Japanese singer', a third one was constructed, where all the constituent syllables of the word /cus'ita/ were made unstressed. All in all, about 20 isolated disyllables and the same number of trisyllables in carrier sentences were presented to 20 phonetically untrained subjects. The subjects were asked to write down what they heard without leaving blanks. In both experiments (with di- and trisyllables) recognition scores for vowels and consonants in 'stressless'words have been found to drop noticeably down to 60-70 and 76-83 per cent in di- and trisyllables respectively.
A word of caution would be in order here. The expression 'stressless', 'unstressed' when applied to words, should not be taken without some reservation. The reason is that the syllables of our pseudo-words are really stressless only with respect to their 'original' words. Within the artificial test words, the syllables, from a perceptual point of view, are to form a new syntagmatic hierarchy of their own, because for a Russian speaker/hearer independent 'stressless'words simply do not exist. In short, stress patterns of the pseudo-words are not absent but rather distorted. The same can be said of the 'monotonized' Chinese and Vietnamese syllables referred to above.
Turning back to the recognizability data, we must admit that the low intelligibility of the test words might be partly attributable to less pronounced syllable contrasts which is characteristic of any unstressed syllable /2/. Yet, there seems to be every reason to believe that what has most affected the performance of our subjects is the distortion of the stress pattern, the identification of the latter thus being an important precondition for the recognition of segmentals -- vowels and consonants. In other words, the situation appears to be similar to that observed for tone languages. It was more than once suggested in literature, that prosodic information is processed independently of that concerned with vowels and consonants. What is no less important, prosodic information seems to be processed p r i o r to segmental information, which means, again, that perception procedures are hierarchical in nature: words appear to be classified first according to their stress patterns, and only then the identification of the words, including the segmental composition, is carried out within the previously determined stress-pattern classes. Both in tone and stress languages the interrelationship of such hierarchical processes exhibits the top-down direction, if we take suprasegmentals as belonging to higher levels of the phonological component and segmentals to lower ones.
In experiments with trisyllables it has been also observed that the effect of stress pattern distortion on recognition of vowels and consonants is the least if the 'de-stressed' word is to be found in final position where sentence stress is typically located. The data are reminiscent of the findings for measuring reaction times in phoneme monitoring experiments as reported by Foss et al. /4/. According to the authors, the targets on words bearing sentence stress are responded to more rapidly than targets on words outside the sentence stress. All such results seem to indicate that the perception of segmentals is dependent not only on prior identification of lexical stress pattern but also on that of sentence intonational pattern (cf. Bosshard's experiment referred to above).
In order to further probe into the nature of the hypothesised relationship, we designed one more experiment where distorted was the sentence intonational pattern. Out of two sentences of the type /p'et'a igrajet na g'itar'e/ and /p'et'a igrajet xarašo/ the third was constructed by means of interchanging and transplanting the words, namely, /na g'itar'e p'et'a igrajet xarašo/ 'The guitar, Pete plays well'. Similar test sentences were constructed with meaningless words, the words and the sentences being modeled after their meaningful prototypes, cp. /na g'itar'e p'et'a igrajet xaraso/ and (meaningless) /na d'ikal'e t'ep'i udlor'it šalaso/. 45 such sentences were presented to the seme team of subjects, first I5 meaningless test sentences, then their I5 meaningful counterparts, and, last, control original sentences.
The results show that the intonation pattern distortion equally impairs recognizability of vowels and consonants. The effect is especially significant for trials with the meaningless sentences which show 60-70 per cent recognition for segmentals.
Our data do not provide a sufficient basis for determining whether recognition of segmentals is directly dependent on identification of intonation contours or segmentals and intonational patterns are associated via lexical stress (the latter option seems preferable).
As can be seen, the 'top' -- i.e. the higher level -- is not identified here directly with semantics. At the same time, it is precisely because of the immediate association with communicatively relevant meaningful units such as sentences, phrases, words, that prosody takes up the role of starting point in speech recognition processes.In fact,prosodic description,i.e.a description in terms of suprasegmentals, serves to provide a first-approximation abstract representation of the meaningful unit to be recognized. The representation is open either to direct semantic interpretation (if the context is highly suggestive) or to further elaboration and transformation by means of bringing into play additional low-level information about segmentals.
We are not going to argue that speech perception is a unidirectional, strictly serial process. Highly plausible is the

existence of modules operating in parallel. Besides, if the initial hypothesis about a word or, say, phrase is rejected as violating some regularities of mapping prosodic structures onto segmental ones the process is started anew -- thus acquiring a shuttle-like character in its functioning.

REFERENCES

I. N.A.Bernstein, The Coordination and Regulation of Movements. London: Pergamon Press, I967.
2. L.V.Bondarko, The syllable structure of speech and distinctive features of phonemes . - Phonetica, I969, vol.20, N I, p.I-40.
3. H.-G.Bosshardt, Suprasegmental structure and sentence perception. - In: H.W.Dechert and M.Raupach (eds.),Temporal Variables in Speech. Studies in Honour of Frieda Goldman-Eisler. The Hague-Paris-New York, I980,p.I9I-I98.
4. D.Foss, D.A.Harwood, and M.A.Blank, Deciphering decoding decisions: Data and devices. - In: R.A.Cole (ed.), Perception and Production of Fluent Speech. Hillsdale: Lawrence Erlbaum Associates, I980, p.I65-I99.
5. V.B.Kasevich, Phonological Problems in General and Oriental Linguistics. Moskow: Nauka Publishers, I983 (in Russian).
6. V.B.Kasevich and E.M.Shabelnikova, Segmentals and suprasegmentals in speech perception. - In: Proceedings of the 9th International Congress of Phonetic Sciences. Copenhagen, I979.
7. V.B.Kasevich and E.M.Shabelnikova, Tempo, rhythm, and the choice of strategy in speech perception. - In: Abstracts of the I0th International Congress of Phonetic Sciences. Dordrecht, I983.
8. G.K.Krulee, D.K.Tondo, and F.L.Wightman, Speech perception as a multi-level processing system. - Journal of Psycholinguistic Research, I983, vol. I2, N 6, p.53I-554.
9. D.Norman, Copy-cat science or does the mind really work by table look-up? - In: R.A.Cole (ed.), Perception and Production of Fluent Speech, p. 38I-395.
I0. P.J.Price and A.G.Levitt, The relative role of syntax and prosody in the perception of the /š/ - /č/ distinction. - Language and Speech, I983, vol.26, pt.3, p.29I-303.
II. I.Rock, An Introduction to Perception. New York - London: Macmillan Publishing Co., Inc. - Collier Macmillan Publisher, I975.

AUDIOVISUALLY PERCEIVED "FUSIONS" WITHIN DIFFERENT VOWEL CONTEXTS

JÖRG SCHORRADT HANS G. PIROTH HANS G. TILLMANN

Institut für Phonetik und sprachliche Kommunikation
der Universität München
Schellingstr. 3, 8000 München 40, F.R.G.

ABSTRACT

This paper describes two experiments with video dubbing. Subjects had to identify a talker's utterances of CV-syllables within a sentence frame. The test syllables had conflicting bimodal information about place of articulation. By desynchronisation we wanted to examine the influence of timing phenomena with regard to different vocalic contexts. The results show a main effect of "fused" answers only in /a/-environment. A missing visual consonantal articulation as in the context of /u/ leads under certain conditions to a systematic elision of the initial acoustic stop consonant.

INTRODUCTION

H. McGurk and J. MacDonald 1976 [3] first described the effects which result from presenting utterances of CV-syllables with conflicting visual and acoustic initial consonants, namely that subjects will hear the consonantal part of the syllable as a function of the distribution of the consonantal information over the two modalities. For example a visual ‹ba› combined with an acoustic [ga] will be heard as a syllable containing both consonants /b/ and /g/ but a visual ‹ga› combined with an acoustic [ba] will be heard as a /da/. The first effect was called "combination", the second effect was called "fusion". Both these effects still work even if subjects know how they are achieved and they work after many repetitions as well.
By systematic temporal desynchronisation of the visual and the acoustic component of the stimulus Tillmann et al. 1984 [7] looked for the characteristic temporal limitations of the effects.
In accordance with earlier findings of audio-visual desynchronisation (Dixon & Spitz 1980 [1]) it could be shown again, that a desynchronisation with the acoustic stimulus leading the visual stimulus was generally more sensitive for being perceived as non-contradictorilly, than a desynchronisation with the visual stimulus leading. This greater tolerance for integrating conflicting information in the latter order as a manifestation of a general principle was refined by further experiments leading to so called "phonological fusions", for example visual ‹ba› and acoustic [la] are heard as /bla/. Here it could be shown that perceived phonotactically regular combinations are much more sensitive to desynchronisation than "fusions" are.
The experiments presented in this paper are concerned with the influence of desynchronization in different vocalic environments.

EXPERIMENTS

In two experiments on audio visual fusions we wished to test a 10-step-desynchronisation continuum from 0 ms to 270 ms delay of the acoustic component following the visual component of the stimulus and the influence of the vowel contexts /i, a, u/. We used three acoustic realisations of each vocalic context for dubbing.

Subjects

39 untrained subjects participated in experiment I and 30 untrained subjects in experiment II.

Stimuli

a) Video recording. The recordings of a male speaker were done in a sound-treated studio using a Panasonic VHS system. Head and shoulders were visible on the monitor screen in a straight front picture. The talker was instructed not to move during recording. A 1000 Hz sinusoidal reference signal of 300 ms duration was generated periodically every 10 s by a PDP 11/50. Every time the speaker heard the signal, he had to utter the following sentence with one of the test syllables: "Ich habe /ba, bi, bu, ga, gi, gu/ gesagt". To avoid misleading information about the closing of the following stops, he was instructed not to close his lips after he had said "ich habe". Reference signal and utterance were recorded on the first soundtrack of the video tape.
b) The visual component of the stimuli. To obtain fusions we only used visual ‹ga, gi, gu› and acoustic [ba, bi, bu] utterances.
For the visual components of the stimuli we used only one realisation of each vocalic context. Cutting the tape was performed on Panasonic source and editing recorders with an editing controller unit

in such a way, that one visual sequence for each vocalic context was arranged according to the randomization plan.
c) The acoustic component of the stimuli. Three realizations of /ba/, /bi/, /bu/ were taken for dubbing the pictures. They were recorded on Revox and digitalized and stored on a PDP 11/50 for later processing. Spectrographic measurements and auditory comparisons showed no perceptual differences. Therefore the three acoustic realizations could be counted as repetitions in the statistical evaluation.
d) Audio-visual dubbing of the stimuli. Editing the acoustic signals for dubbing was controlled on the PDP 11/50. The programs developed especially for experiments in the McGurk paradigm allowed the second soundtrack of the video tape to be dubbed with exact desynchronization using the first period of the sinosoidal reference signal and different segment files providing the necessary information. During the dubbing process the output of the reference signal was suppressed so that the test tape received the sentence frames, the test syllables and the pauses in relation to the visible articulation and the randomizing plan.

PROCEDURE

In experiment I we had 90 stimuli randomized over 10 steps of desynchronisation (0 ms - 270 ms), three vocalic contexts (/a/, /u/, /i/) and three acoustic realisations of each context. Vocalic context and desynchronisation counted as factors, the realizations as repetitions in a two-factorial design.
Subjects were tested singly in the speaker room of the recording studio. They sat at a distance of 3 m from a Sony colour monitor with a tube of 68 cm in diameter. Only one of the monitor's loudspeaker was used and was placed on top of the monitor. Therefore we must take into account about 10 ms delay of acoustic information relative to visual information due to the velocity of sound.
The subjects read an instruction-sheet before the tests started. They were asked to make a binary forced-choice decision about the identity of the initial consonant. On an answering sheet they could choose to mark a /d/ answer, which we interpreted as "fusion" or to mark a /b,g/ answer which we interpreted as no "fusion"
It was pointed out to the subjects, that a correct fullfillment of the experimental task was entirely dependent on the fact that they looked at the speaker's lips and answered which syllable they heard.
In a first demonstration the subjects had to respond to 18 syllables. The three acoustic realisations were presented with a 30-ms- and a 270-ms-desynchronisation for each vowel context, first /a/, then /i/, and finally /u/. After a short break which could be used for questions the test was started.
With a small pause after each sentence and a larger pause after each block of 20 sentences the total duration of the test was 15 minutes.
At the end of this test-run subjects were asked in a short interview to describe their impressions according to the second twofold category.
Following the interview the subjects' purely auditory perception was checked by presenting 18 sentences for demonstration but without vision.

In Experiment II we presented the same stimuli and used the same procedure as in experiment I, but only one vocalic context (/a/, /u/, or /i/) was presented in each of three tests and no demonstration preceded the test. Here the total duration of one test was 5 minutes.

RESULTS

Fig. 1, 2, and 3 show the percentage of "fused" answers, that is /d/, for each of the ten steps of desynchronisation, for each of three acoustic realisations and for each vocalic context averaged about subjects.

Fig. 1 <ga>[ba]

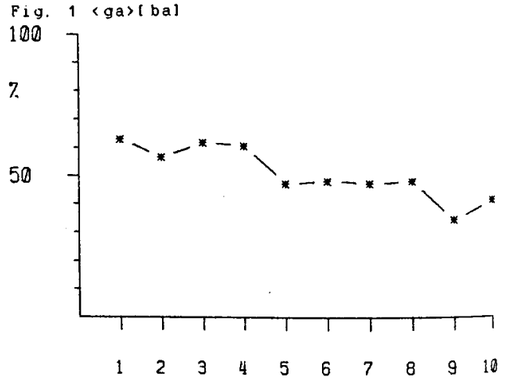


Fig. 2 <gu>[bu]

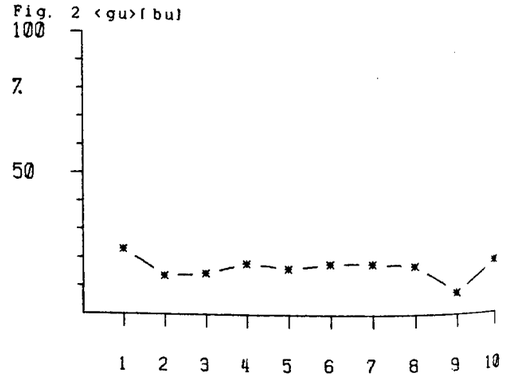
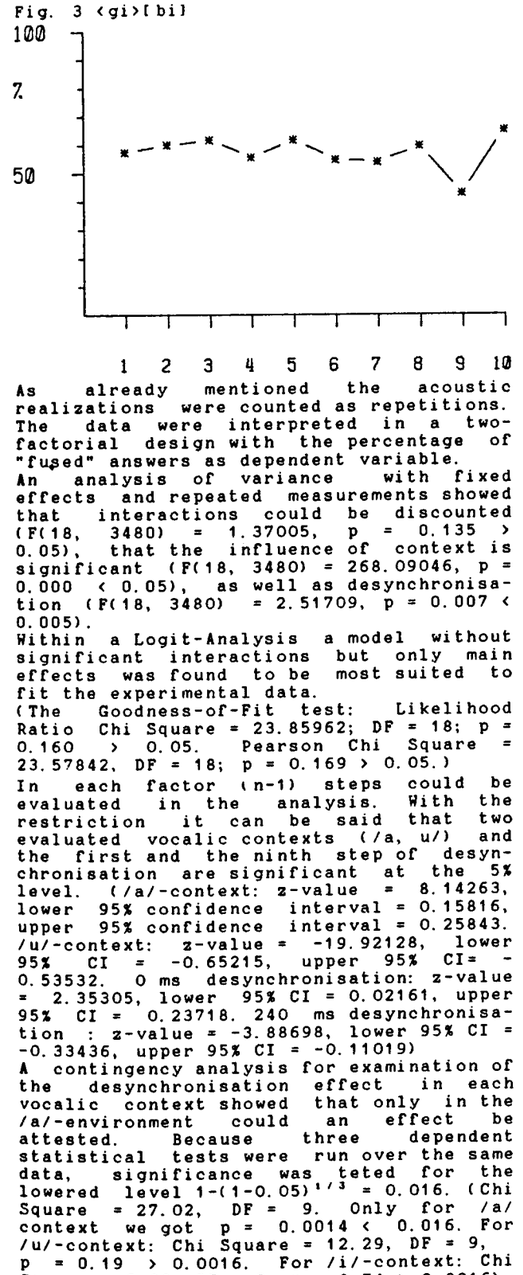


Fig. 3 <gi>[bi]



As already mentioned the acoustic realizations were counted as repetitions. The data were interpreted in a two-factorial design with the percentage of "fused" answers as dependent variable.
An analysis of variance with fixed effects and repeated measurements showed that interactions could be discounted ($F(18, 3480) = 1.37005$, $p = 0.135 > 0.05$), that the influence of context is significant ($F(18, 3480) = 268.09046$, $p = 0.000 < 0.05$), as well as desynchronisation ($F(18, 3480) = 2.51709$, $p = 0.007 < 0.005$).
Within a Logit-Analysis a model without significant interactions but only main effects was found to be most suited to fit the experimental data.
(The Goodness-of-Fit test: Likelihood Ratio Chi Square = 23.85962; DF = 18; p = 0.160 > 0.05. Pearson Chi Square = 23.57842, DF = 18; p = 0.169 > 0.05.)
In each factor (n-1) steps could be evaluated in the analysis. With the restriction it can be said that two evaluated vocalic contexts (/a, u/) and the first and the ninth step of desynchronisation are significant at the 5% level. (/a/-context: z-value = 8.14263, lower 95% confidence interval = 0.15816, upper 95% confidence interval = 0.25843. /u/-context: z-value = -19.92128, lower 95% CI = -0.65215, upper 95% CI= -0.53532. 0 ms desynchronisation: z-value = 2.35305, lower 95% CI = 0.02161, upper 95% CI = 0.23718. 240 ms desynchronisation : z-value = -3.88698, lower 95% CI = -0.33436, upper 95% CI = -0.11019)
A contingency analysis for examination of the desynchronisation effect in each vocalic context showed that only in the /a/-environment could an effect be attested. Because three dependent statistical tests were run over the same data, significance was teted for the lowered level $1-(1-0.05)^{1/3} = 0.016$. (Chi Square = 27.02, DF = 9. Only for /a/ context we got $p = 0.0014 < 0.016$. For /u/-context: Chi Square = 12.29, DF = 9, $p = 0.19 > 0.0016$. For /i/-context: Chi Square = 6.00, DF = 9, $p = 0.74 > 0.0016$) This explains the missing interactions between desynchronisation and context.
The Logit-Analysis of desynchronisation was confirmed by a different analysis using orthogonal contrasts. The differences between the average of the "fused" responses between step 1 and all the other steps is significant ($F(1, 3480) = 4.76418$, $p = 0.019 < 0.05$) as well as between step 9 and 10 ($F(1, 3480) = 6.39973$, $p = 0.007 < 0.05$).
Summing up these results it can be stated, that the variable "fusion" is dependent on the variability within the first and ninth step of the variable desynchronisation and on the variability within all three steps of the variable context. The influence of temporal shifts between visual and acoustic information is only relevant for the /a/-context.
After the test 33 subjects were interviewed. They were asked whether they had always heard an initial consonant, and which alternative of the second response category they preferred in each vocalic context.
Only three subjects heard mostly /b/ in /u/-environment, no subjects heard /g/, but 30 subjects heard with very few exeptions only the vowel /u/; some subjects mentioned a hard glottal attack.
The vowel alone was heard almost exclusively in /u/-context but never in /i/-context.
Experiment II was a control experiment with regard to the mixed vowel condition in experiment I. Fig. 4, 5, and 6 show

Fig. 4 <ga>[ba]

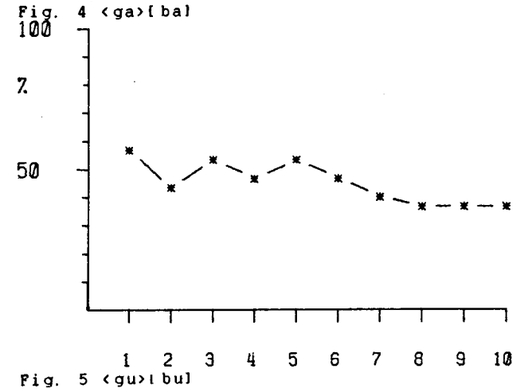


Fig. 5 <gu>[bu]

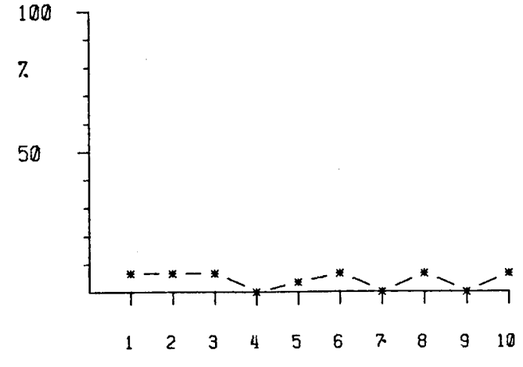
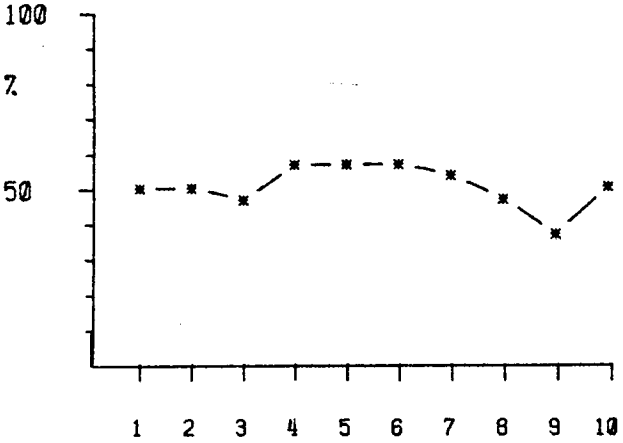


Fig. 6 <gi>[bi]



the percentage of "fused" answers of 30 subjects (10 subjects for each test)

The results agree broadly with the earlier test. But based on the following interview only one subject had heard only the vowel /u/, whereas 9 subjects heard /bu/ No differences occured concerning the /i/- and /a/-contexts.

DISCUSSION

In the /a/-context subjects could observe a downward shift of the tongue body, which was clearly different from tongue movement in the /i/-environment ending at the teeth and therefore enforcing information about a dental place of articulation, which led to highly "fused" responses. As the statistical evaluation shows, the desynchronisation effect is only significant in the /a/-context. This might be explained by the specific tongue movement being clearly visible.
In the /u/-context the protrusion of the lips totally masked all visible consonantal information. Therefore "fusions" did not occur. In the first experiment the /a/- and /i/-contexts provided visual consonantal information for fusions. In comparison to these stimuli the visual information in the /u/-context had a special effect: subjects heard no oral consonant at all. We call this effect "elision".
In the second experiment the different result of the interview concerning the /u/-context indicates that the visual component of the stimulus looses its influence, if no consonantal articulation can be seen. Here integration of the bimodal information did not happen and subjects realized the contradiction of the information presented by both modalities.
The effects of fusion and of elision in the /i/- and /u/-context have in common that they are not affected by desynchronization within the temporal domain tested in the experiments. The effect of desynchronization is specific only to the /a/-environment. Further experiments are necessary to investigate the bimodal temporal relationships and the special effect of elision.

REFERENCES

[1] N. F. Dixon, L. Spitz, "The detection of auditory visual desynchrony", Perception, 9, p. 719 - 721, 1980.

[2] M. McGrath, Q. Summerfield, "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults", J. Acoust. Soc. Am. 77 (2), p. 678 - 685, 1985.

[3] H. McGurk, J. MacDonald, "Hearing lips and seeing voices", Nature, 264, p. 746 - 748, 1976.

[4] J. MacDonald, H. McGurk, "Visual influences on speech perception processes", Perc. & Psyphy., 24 (3) p. 253 - 257, 1978.

[5] B. H. Repp, S. Y. Manuel, A. M. Liberman, M. Studdert-Kennedy, "Exploring the 'McGurk Effect'", paper presented at the 24th annual meeting of the Psychonomic Society, San Diego, 1983.

[6] Q. Summerfield, "Use of visual information for phonetic perception", Phonetica Vol. 36 No. 4 - 5, S. 314 - 331, 1979.

[7] H. G. Tillmann, B. Pompino-Marschall, U. Porzig, "The effects of visually presented speech movements on the perception of acoustically encoded speech articulation as a function of acoustic desynchronization", Proc. of the 10th ICPhS, p. 496 - 473, 1984.

ACKNOWLEDGEMENTS

Statistical advice was given by Dr. Alexander Yassouridis, Departement of Biostatistics, Max-Planck-Institute for Psychiatry, Munich F.R.G.

INTERACTION BETWEEN PHONETIC AND LITERAL COMPONENTS IN PERCEPTION AND PRODUCTION OF JAPANESE SPEECH

SHIZUO HIKI

School of Human Sciences, Waseda Universirty
Tokorozawa-shi, Saitama-ken, 359 Japan

ABSTRACT

In order to investigate sources which cause errors in the performance of language, the mechanisms of processing speech sound and letter pattern are schematized based on the results of analysis of the errors made by adults having normal competence in various modes of performance of Japanese language. First, all the steps of processing involved in the primary and secondary functions in listening, speaking, repeating, silent reading, writing and copying are formulated. Then, the process of reading aloud and stenographing is formulated by combining the processes of silent reading and speaking or those of listening and writing, respectively, through the steps of phoneme/letter conversion. Through this modelling, it has become possible to simulate systematically the sources of errors with special regard to the interaction between phonetic and literal components in perception and production of Japanese speech.

INTRODUCTION

In the previous study on errors in performance of the Japanese language made by adults having normal competence (Reference), it was shown that many of the errors, not only in listening, speaking and reading aloud but also in silent reading, writing and copying, that could be explained either in the syntactic, semantic or orthographic level were found to be related to the phonetic level too. It was also shown that the errors that could be explained in the orthographic level occured even in speaking and reading aloud, although they occured mostly in silent reading, writing and copying.

The data were about 1,100 samples of errors, in total, in the six modes of performance. The errors which were related to the phonetic level were 70% of the total errors. Among those, the percentage for listening was 100%, and those for speaking and reading aloud were 75% and 85%, respectively. Even for silent reading, writing and copying, the errors in the phonetic level were found to be 25%, 55% and 65%, respectively. Percentage of errors which were explained in the orthographic level were 35% for silent reading, 70% for writing and also 70% for copying. For speaking and reading aloud, the errors in the orthographic level were found to be a few percent.

In order to investigate in sources which caused those errors, the mechanisms of processing speech sound and letter pattern are schematized based on the results of analysis of the data. Modes of performance of language included in the modelling are;

listening,
speaking,
repeating,
silent reading,
writing,
copying,
reading aloud and
stenographing.

MECHANISMS OF PROCESSING SPEECH AND LETTER

All the steps of processing speech sound involved in the primary functions in listening and speaking are schematized (Figure 1a). The secondary functions such as prediction in the process of listening and readjustment in the process of speaking are formulated by connecting some parts of the steps involved in the complementary primary function.

The primary function of extracting concept from speech sound as auditory input in listening is composed of the following seven steps;

acoustical analysis of cue (L1),
phonetic perception of feature (L2),
phonemic judgement (L3),
word identification (L4),
syntactic analysis (L5),
semantic analysis (L6) and
concept analysis (L7).

That of generating speech sound as oral output from concept in speaking is composed of

semantic generation (S7),
syntactic assignment (S6),
lexical compilation (S5),
phonemic conversion (S4),
phonetic conversion (S3),
motor command generation (S2) and

acoustical synthesis of speech (S1). Each of the steps in the process of speaking corresponds to one of those in the process of listening, except for that the direction of flow from input to output is reversed.

The outputs of the steps of extracting phonetic feature, phoneme, word, syntax, meaning and concept in the process of listening are transfered via interconnections (shown in dotted lines) to the corresponding steps in the process of speaking as inputs, and the outputs of the predictive trial in those steps are compared (in CLS3, CLS4, CLS5 and CLS6) with the outputs of the previous steps in the process of listening, so that the operations of the steps are modified according to the detected difference between them. The inputs and outputs are stored in temporal memories (ML1, ML2, ML3, ML4, ML5, ML6 and ML7, and MS1, MS2, MS3, MS4, MS5, MS6 and MS7) during the operations. Those interconnections are utilized also for the readjustment in the process of speaking.

The mechanism of processing letter patterns in silent reading and writing is schematized in an analogous way to that of processing speech sound (Figure 1b).

INTERACTION BETWEEN SPEECH AND LETTER

The process of reading aloud or stenographing is formulated by combining the processes of reading and speaking or those of listening and writing, respectively, through the steps of phoneme/letter conversion. This provides the scheme which causes the interconnection between the mechanisms of processing speech sound and letter pattern.

In the process of reading aloud, letter extracted as the output of letter judgement (R3) in the process of silent reading is transfered to the input of phonetic conversion (S3) in the process of speaking (as shown in dotted line) through the step of phonological conversion (COP). In the same way, in the process of stenographing, phoneme is converted into letter (as shown also in dotted line) through the step of orthographic conversion (CPO) and transfered from the process of listening to that of writing.

By assuming all these mechanisms, possible sources of errors in performance of Japanese language, especially effects of letter pattern on processing speech sound and those of speech sound on processing letter pattern, can be simulated systemetically.

The errors that were explained best in the orthographic level but occured in processing speech sound through the auditory organs can be simulated as follows: Any of the inputs or outputs of the steps involved in the mechanism of processing speech sound such as word, syntax, meaning, concept and phonetic feature can be transformed into phoneme via the interconnections between the processes of listening and speaking, and converted into letter through the interconnection between the mechanisms of processing speech sound and letter pattern. Then, the letter pattern can be transformed into any of the inputs or outputs of the steps involved in the mechanism of processing letter pattern such as word, syntax, meaning, concept and orthographic feature via the interconnection between the processes of silent reading and writing. When converted again into phoneme, they are accompanyed by errors characteristic of the mechanism of processing letter pattern.

The errors that occured in processing letter pattern through the visual organs but were explained best in the phonetic level can be simulated in analoguous way to the above.

DISCUSSION

In the previous study on errors in performance of Japanese language, it was shown that many of the errors in silent reading, writing and copying that could be explained either in the syntactic, semantic or orthographic level were found to be related to the phonetic level too, while that the errors that could be explained in the orthographic level occured even in speaking and reading aloud. But, there has been no theoretical model which describe reasonably this essential aspect of language performance.

In this study, the mechanism of the interaction between phonetic and literal components was explained, both theoretically and experimentally, through modelling various modes of perception and production of speech and letter, based on the analysis of errors in the performance of Japanese language. The model can be applied not only to the basic study of speech perception and production processes but also to the application such as automatic speech recognition and understanding.

REFERENCE

Shizuo Hiki: "An analysis of errors in the performance of the Japanese language," Journal of the Acoustical Society of America, Vol.64, Supplement No.1, p.S94, Fall 1978 (Program of the Acoustical Society of America and the Acoustical Society of Japan Joint Meeting, 27 November-1 December 1987).

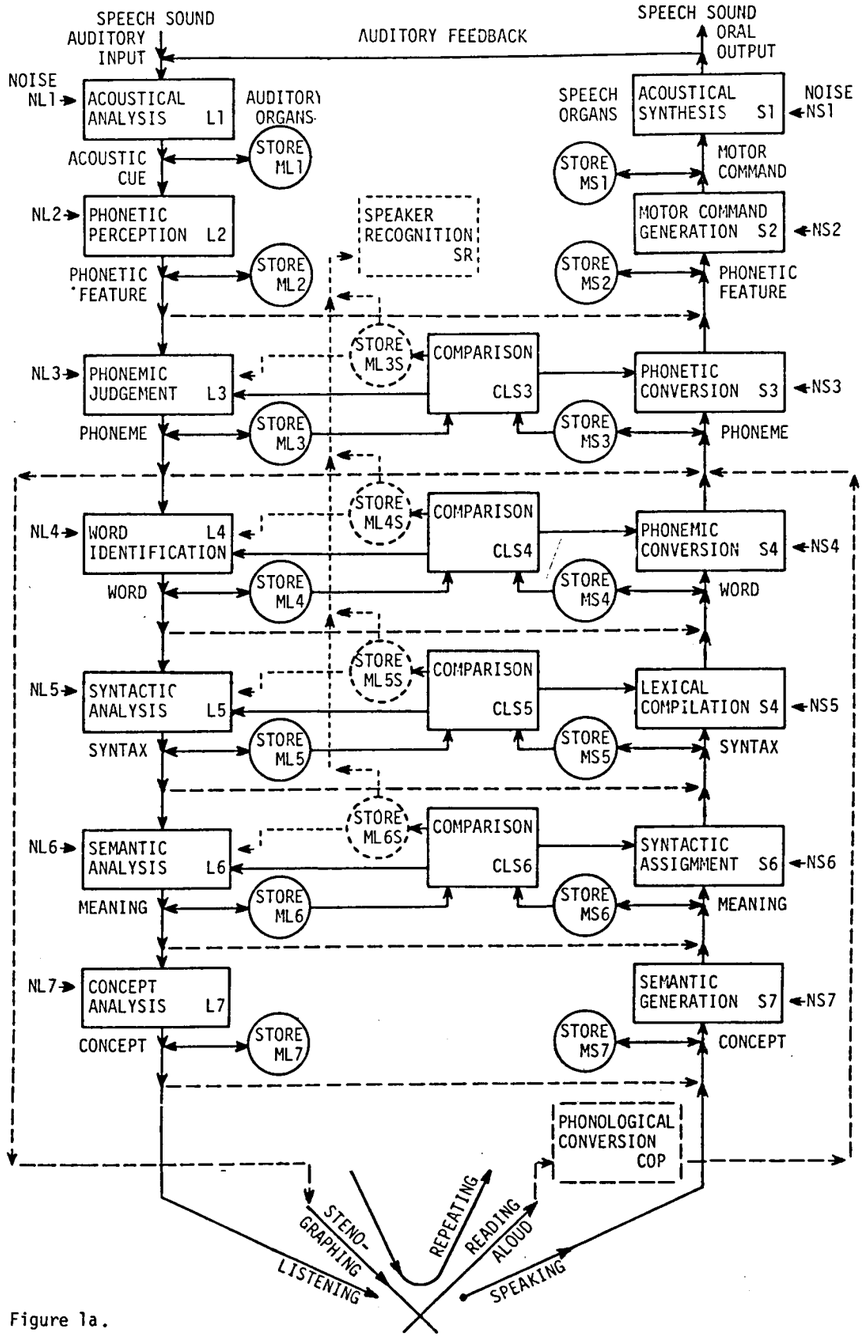
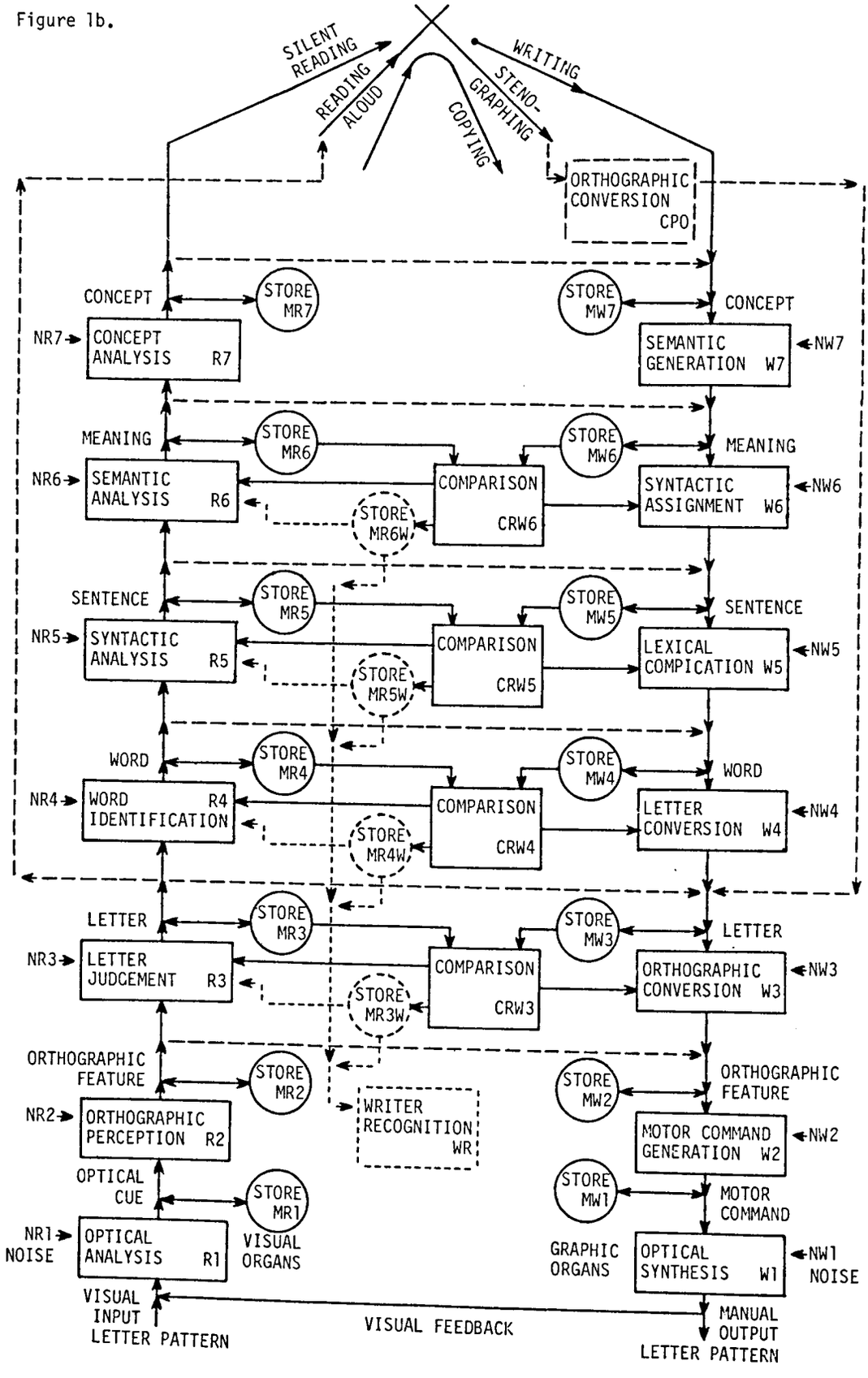


Figure 1a.

Figure 1b.


SILENT READING
READING ALOUD
COPYING
STENO-GRAPHING
WRITING
ORTHOGRAPHIC CONVERSION CPO
CONCEPT
STORE MR7
STORE MW7
CONCEPT
NR7
CONCEPT ANALYSIS R7
SEMANTIC GENERATION W7
NW7
MEANING
STORE MR6
STORE MW6
MEANING
NR6
SEMANTIC ANALYSIS R6
COMPARISON CRW6
SYNTACTIC ASSIGNMENT W6
NW6
STORE MR6W
SENTENCE
STORE MR5
STORE MW5
SENTENCE
NR5
SYNTACTIC ANALYSIS R5
COMPARISON CRW5
LEXICAL COMPICATION W5
NW5
STORE MR5W
WORD
STORE MR4
STORE MW4
WORD
NR4
WORD IDENTIFICATION R4
COMPARISON CRW4
LETTER CONVERSION W4
NW4
STORE MR4W
LETTER
STORE MR3
STORE MW3
LETTER
NR3
LETTER JUDGEMENT R3
COMPARISON CRW3
ORTHOGRAPHIC CONVERSION W3
NW3
STORE MR3W
ORTHOGRAPHIC FEATURE
STORE MR2
STORE MW2
ORTHOGRAPHIC FEATURE
NR2
ORTHOGRAPHIC PERCEPTION R2
WRITER RECOGNITION WR
MOTOR COMMAND GENERATION W2
NW2
OPTICAL CUE
STORE MR1
STORE MW1
MOTOR COMMAND
NR1
NOISE
OPTICAL ANALYSIS R1
VISUAL ORGANS
GRAPHIC ORGANS
OPTICAL SYNTHESIS W1
NW1
NOISE
VISUAL INPUT LETTER PATTERN
VISUAL FEEDBACK
MANUAL OUTPUT LETTER PATTERN

A MODEL FOR FILTERING AND ANALOG-TO-PULSE CONVERTION ON THE PERIPHERY OF AUDITORY PATHWAY

N.G. BIBIKOV, N.A. DUBROVSKY, G.A. IVANITSKY, L.K. RIMSKAYA-KORSAKOVA, V.N. TELEPNEV

N.N. Andreev Acoustical Institute, Moscow, USSR, 117036

ABSTRACT

An approach based on modelling the significant features of auditory processing, according to physiological evidence, provides a perspective opening into future speech analysis techniques. A model for information processing on the periphery of auditory pathway is presented. The model accomodates our knowledge of auditory nerve impulsation patterns and comprises a basilar membrane filtering, signal integration and rectification, and an analog-to-pulse convertion in first order auditory neurons which possesses refractoriness. Inserting the accumulating refractoriness in the model for auditory nerve fiber provides realistic representation of the short-term adaptation phenomena. The patterns of model reaction to tone bursts and stationary fragments of vowels are given.

The peripheral components of our model are represented by a bank of linear bandpass filters for simulation of the frequency selectivity of the inner ear[1], and a receptor-neuron model is to simulate generation of neural discharges. The impulse response of the minimum phase filter is as follows

$$h(t)=(\omega_0 t)^{\beta}\exp(-\lambda\omega_0 t)\sin(\omega_0 t)$$

Here ω_0 is the central frequency, and parameters λ and β define steepnes of the front and rear slopes of impulse response. The specific values of parameters λ and β were chosen according to known data on auditory pathway periphery impulse responses. Both data on basilar membrane response to short acoustical clicks [2] and data obtained by reverse-correlation technique [3,4] were used. The reverse correlation between spike activity of auditory nerve fiber and the noise stimulus at input yields an estimate of impulse response. The resultant quality of filters used is not dependent on central frequency being approximately 4.2 if measured for 10 dB SPL. We avoided the modelling of compressive non-linearity of basilar membrane mechanical oscillations and limited ourselves with studies of relatively weak signals. In order to take into account signal transformations in hair cells, a partial one-period rectification was included in the model. Amplitude of signal's negative half-wave was three times less than the positive half-wave, in accordance with physiological data of Sellick and Russel [5]. After rectification and filtering the signal was added to Gaussian noise which provided random fluctuations of neuron membrane potential and presense of auditory nerve fibers spontaneous activity. A sum of noise and determinated signal then passed through the low-pass filter with integration time-constant of 0.2 ms that corresponds to the known data on integrating properties of hair cell - spiral ganglion neuron dendrite system [6], and then fed the threshold circuit. We omitted modelling the phenomena of first synapse neuro-transmitter depletion since we could not find direct physiological evidence of this effect.

We paid much attention to modelling the postspike changes in the auditory nerve fiber. This changes, we consider, may play a significant role in coding variations in signal amplitude. We tried to evaluate refractoriness parameters from data obtained by Gaumond et al. [7]. These authors obtained hazard functions for auditory nerve fibers spontaneus activity using large statistical selections. We succeeded in making our model reproduce these

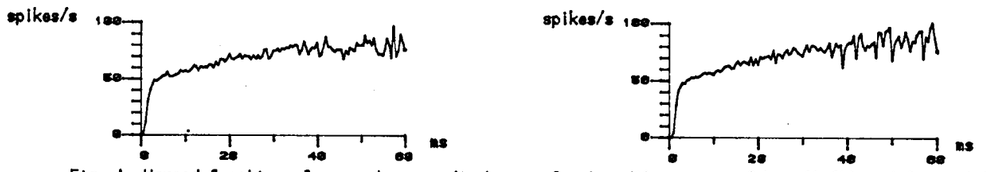


Fig. 1. Hazard functions for spontaneous discharge of cat auditory nerve fiber (left) and for model spontaneous impulsation (right).

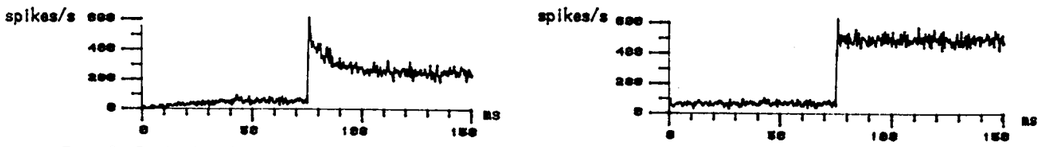


Fig. 2. Post-stimulus-time histograms of model reaction to tone bursts of 75 ms duration and 75 ms inter-burst gaps. Left: accumulation of refractory threshold changes is present; right: no such accumulation.

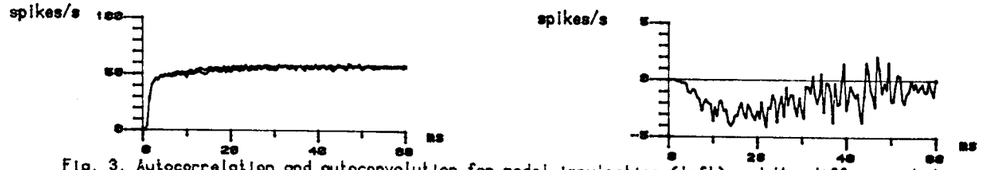


Fig. 3. Autocorrelation and autoconvolution for model impulsation (left) and the difference between these functions (right).

functions. The best agreement with experimental data was achieved when the model parameters were as follows: standard deviation of intrinsic noise - 0.6 of threshold at rest; absolute refractoriness - 0.5 ms; relative refractoriness comprises two "phases" - "fast" and "slow" with the values of time constants 0.35 ms and 20 ms respectively (fig. 1) [8]. The model, with parameters set to these values, was then exposed to high-frequency tone bursts. These bursts, after rectification and smoothing, become nearly rectangular depolarizing pulses of 50 ms duration each. We were interested of finding out whether or not such a model will show the intrinsic short-term adaptation properties, typical to the auditory nerve fibers [9]. It reveales that the shape of the model post-stimulus-time histograms depends crucially on the mode of summation of threshold elevations: namely, whether or not successive threshold elevations (relative to 20 ms "slow phase" refractoriness) are accumulated. If, at the end of absolute refractoriness, the threshold rises up to some fixed level, adaptation reveals to be very short and insignificant. But if, at the end of absolute refractoriness, the threshold increases for certain value beside the value it have had just before the spike occurrence, situation changes radically. In this latter case of accumulating refractory threshold changes the model shows the intrinsic property of short-term adaptation and decrease in ability for excitation succeeding the end of stimulus (fig. 2). Unfortunately we know few about the auditory

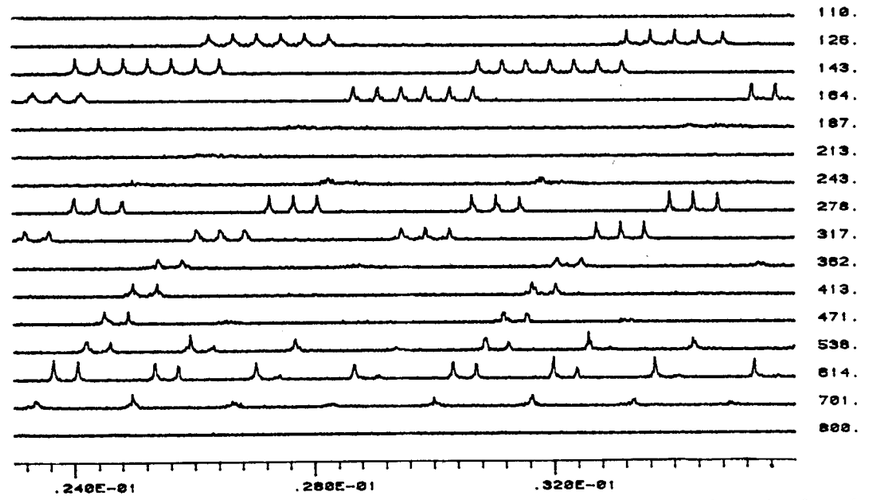


Fig. 4. Post-stimulus-time histograms of model's 16 channels reaction to stationary fragments of vowel. Channel's best frequency is on the right of any curve. Horizontal axis represents time in seconds.

nerve refractory patterns. There are data obtained by Gaumond et al. [7] showing successive inter-spike intervals negative correlation; this corresponds to accumulating refractoriness hypothesis. Some additional data were published by Lutkenhoner and Smith [10]. A simple method of detecting the presense and measuring the time-course of accumulating refractoriness is provided by a comparison of two functions: the autocorrelation of stationary spike train, and the autoconvolution derived from interspike intervals distribution obtained from the same spike sequence. If there is no accumulating refractoriness, and hence no interspike intervals correlation, these functions are identical. When accumulation is present, autocorrelation function drives lower then autoconvolution function and the difference between these functions represents the time course of accumulating refractoriness (fig. 3).
The model that includes all stages mentioned above was studied respective to some vowel-like signals. Fig. 4 shows the behaviour of the model with total of 16 frequency channels that cover 110 - 800 Hz band. There can be seen: 1)the ability to distinguish formants; 2)clear representation of main tone with appropriate segmentation; 3)time dispersion of low- frequency components; 4)distinction of waveform fine temporal peculiarities.
We do understand that the suggested model represents auditory periphery's properties roughly enough. Further development of the model implies involving non-linear small signal amplification and large signal compression mechanisms. The model, so modified, would become free of dynamic range restrictions. Our techniques may have much more of a physiological and speculative motivation then a mathematical one. But we hope that involving such approach may become useful when designing speech recognition systems with neuronal interactions.

REFERENCES

[1] Flanagan G.L. Speech analysis, synthesis and perception. Springer, Berlin - New-York, 1965.
[2] Robles L., Rhode W., Geisler C. Transient response of the basilar membrane measured in squirrel monkeys, using the Mossbauer effect. J.Acoust.Soc.Am., 59,926-939, 1976.
[3] Evans E., Palmer A. Dynamic range of cochlear nerve fibers to amplitude-modulated tones. J.Physiol., 928,33-34, 1984.
[4] Moller A. Frequency selectivity of single auditory nerve fibers in response to broad band stimuli. J.Acoust.Soc.Am., 62,135-142, 1977.
[5] Russel I., Sellick P. Low frequency characteristics of intracellulary recorded receptor potentials in the guinea pig cochlear hair cells. J.Physiol., 338,179-206, 1983.
[6] Palmer A., Russel I. Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair cells. Hear.Res., 23,1-15, 1986.
[7] Goumond R.P., Molnar C.E., Kim D.O. Stimulus and recovery dependence of cat cochlear nerve fiber spike discharge probability. J.Neurophysiol., 48,856-873, 1882.
[8] Bibikov N.G., Ivanitsky G.A. A model for spontaneous discharge and short-term adaptation in auditory nerve fibers. Biofizika, 29,141-145, 1985 (in Russian).
[9] Kiang N. Discharge patterns of single fibers in the cat's auditory nerve. M.I.T. Press, Mass., 1965.
[10] Lutkenhoner B., Smith R. Papid adaptation of auditory nerve fibers: fine strusture at high stimulus intensities. Hear. Res., 24,289-294, 1986.

МОДЕЛЬ ПЕРИФЕРИЧЕСКОГО СЛУХОВОГО АНАЛИЗА И АДАПТАЦИИ

В.С. ШУПЛЯКОВ Л.В. ЛЕСОГОР Ж.М. ДОЛМАЗОН

Институт физиологии им. И.П.Павлова АН СССР
Ленинград

Институт речевой связи
Гренобль, Франция

РЕЗЮМЕ

Обсуждаются частотно-избирательные свойства периферического отдела слухового анализатора и приводятся дополнительные доводы в пользу участия активного и нелинейного механизма в формировании частотной избирательности на уровне гидромеханики улитки внутреннего уха. На основе выдвинутой гипотезы об активном механическом раскачивании базилярной мембраны наружными волосковыми клетками предлагается модель элементарного звена и приводятся её характеристики.

Современные литературные данные свидетельствуют о том, что уже на уровне механических колебаний базилярной мембраны внутреннего уха частотная избирательность такая же, как и у всей слуховой системы [4]. Следовательно, необходимость во "втором фильтре", до последнего времени широко обсуждавшемся в литературе, отпадает. В таком случае возникает вопрос, каким образом формируются те частотно-избирательные характеристики, которые регистрируются как на различных уровнях периферической части слуховой системы: базилярной мембране, рецепторных клетках, волокнах слухового нерва, так и в психоакустических экспериментах. Перечислим основные их свойства: I. Частотно-избирательные кривые при сравнительно небольшой добротности (порядка I0) имеют весьма высокую крутизну спада как в сторону высоких, так и в сторону низких частот, достигающую нескольких сотен дБ на октаву. При этом коэффициент неравномерности кривых (в иностранной литературе используется термин "pik -tail parameter ") может достигать 70 дБ и более. Представляется, что такие по существу полосовые, характеристики могут быть получены лишь в системе с очень высокими добротностями элементарных звеньев. 2. Частотно-избирательные кривые нелинейны главным образом, в области частот резонанса; на удаленных от резонанса частотах система линейна. Другими словами, периферический отдел органа слуха обладает частотно-зависимой нелинейностью. 3. Нелинейность амплитудной характеристики периферического отдела слуха в наибольшей мере проявляется при малых и средних уровнях сигналов (от 20 до 70 дБ); при больших уровнях система линейна. Это находится в противоречии со свойствами пассивных цепей. 4. Форма частотных характеристик сильно зависит от условий, влияющих на обменные процессы в тканях (температуры, гипоксии, переутомления, при отравлении различными ядами и др.). 5. Частотно-избирательные свойства периферического отдела органа слуха таковы, что имеет место подавление одних составляющих спектра сложного сигнала другими, т.е. форма спектра на выходе анализатора в значительной мере зависит от соотношения амплитуд в спектре входного сигнала.

Перечисленные выше частотно-избирательные свойства периферического отдела орга-

на слуха говорят о сложности этой системы, о том, что она существенно отличается от известных технических систем спектрального анализа и что реакцию этого анализатора на конкретные речевые и другие сложные сигналы совершенно невозможно предсказать, если мы не имеем достаточно полной модели процессов, приводящих к формированию вышеперечисленных свойств. Предложенные до последнего времени модели гидродинамических процессов в улитке органа слуха исходят из того, что эти процессы пассивные и линейные. При всем многообразии моделей и допущений относительно свойств базилярной мембраны, ни одна из них не обеспечивает тех параметров частотно-избирательных характеристик, особенно значений коэффициента неравномерности, которые имеет слуховая система при малых уровнях сигналов [2]. В то же время соответствие свойств этих моделей с характеристиками, получаемыми при больших уровнях сигналов вполне удовлетворительное. Требуемые значения коэффициента неравномерности (до 70 дБ) принципиально невозможно получить с помощью классических пассивных моделей гидродинамики улитки [I], т.к. с ростом добротности элементарных звеньев (рис. I), увеличивается их шунтирующее действие, что ведет к понижению разностного давления, воздействующего на данную точку базилярной мембраны и, в итоге, к уменьшению коэффициента неравномерности частотно-избирательных кривых.

По-видимому, требуемые свойства могут быть получены с помощью активных моделей, у которых высокие значения добротности достигаются при больших значениях сопротивления потерь. Этот вывод хорошо согласуется с современными данными, в частности, с данными Кемпа [3] говорящими о существовании во внутреннем ухе активного процесса на уровне механических колебаний.

Этот процесс каким-то образом должен "выключаться" при больших уровнях сигналов с тем, чтобы свойства модели в этом случае приближались к свойствам классических линейных моделей.

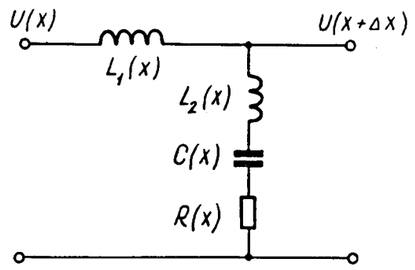


Рис. I. Электрический аналог элементарного звена классической линейной модели гидродинамики улитки органа слуха. $L(x)$, $C(x)$ и $R(x)$ – электрические аналоги удельной массы, податливости и сопротивления потерь при колебании базилярной мембраны внутреннего уха.

В настоящей работе предложена модель и описаны свойства элементарного звена модели гидромеханики улитки внутреннего уха с активной нелинейной обратной связью. Полная модель представляется в виде неоднородной структуры, распределенной по координате X, с рассматриваемыми элементарными звеньями.

ОПИСАНИЕ МОДЕЛИ

При создании модели мы исходили из выдвинутой нами ранее гипотезы [2], что наружные волосковые клетки Кортиева органа выполняют роль непосредственных усилителей механических колебаний базилярной мембраны. При этом собственно рецепторами, воспринимающими колебания и передающими информацию в центральные отделы слуха являются внутренние волосковые клетки. В пользу такой гипотезы говорят следующие факты: I. Афферентные волокна отходят, в основном, от внутренних волосковых клеток. Эфферентные волокна подходят, главным образом, к наружным волосковым клеткам. 2. Наружные волосковые клетки одним своим концом приходятся на срединную (самую чувствительную) частоту базилярной мембраны, другим – упираются в массивную текториальную мембрану, причем цилии наружных волосковых клеток расположены в виде достаточно жесткой на изгиб V-образной конструкции. Таком образом, любое изменение длины клетки под действием потенциала будет приводить к смещениям и базилярной мембраны. 3. В структуре наружных волосковых клеток найдены актин и миозин, необходимые части любой сократительной системы. Сократительные движения волосковых клеток под действием приложенного потенциала в настоящее время показаны экспериментально.

Смысл другого допущения, которое было положено в основу при создании модели, состоит в том, что активный механизм имеет ограниченный энергетический ресурс и амплитудная характеристика имеет линейный участок лишь при малых амплитудах; при больших амплитудах характеристика нелинейна, и имеет вид типа характеристики насыщения.

С учетом вышесказанного, структурная схема электрического аналога элементарного звена модели гидродинамики внутреннего уха с усилением сигнала в цепи положительной обратной связи представлена на рис. 2. Были исследованы два варианта: электрическая модель, выполненная в аналоговом виде, и математическая модель. В последнем случае принималось, что нелинейное звено имеет характеристику типа:

$$U_{вых} = K \operatorname{th}(\alpha U_{вх})$$

где α и K – параметры.
В электрической модели характеристика нелинейного звена была близка к логарифмической.

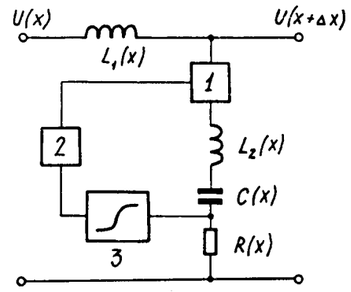


Рис. 2. Электрический аналог элементарного звена нелинейной модели гидродинамики улитки внутреннего уха.

I – сумматор; 2 – усилитель; 3 – нелинейное звено. Остальные обозначения как на рис. I.

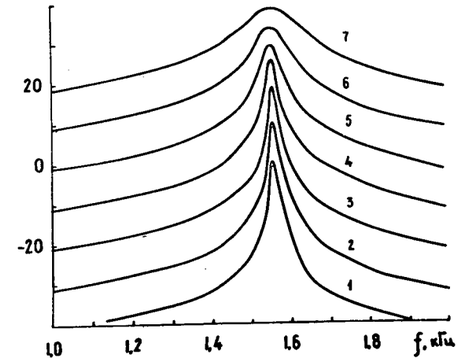


Рис. 3. Семейство амплитудно-частотных характеристик нелинейной модели.

По оси абсцисс – частота входного сигнала, в кГц; по оси ординат – уровень выходного сигнала , в дБ. Кривые I-6 получены при уровнях входного сигнала, отличающихся на I0 дБ; кривая 7 – без обратной связи.

Вследствие того, что в цепь положительной обратной связи включено нелинейное звено

с характеристикой типа насыщения, положительная обратная связь оказывается наиболее сильной при малых уровнях сигналов; с увеличением уровня её роль ослабевает и при больших уровнях система становится практически линейной.

Из рис. 3 можно видеть, что в рассматриваемом диапазоне уровней входных сигналов модель имеет различные амплитудные характеристики на различных частотах: в области резонанса она нелинейная, а на других частотах - линейна. В связи с этим можно отметить еще одно интересное свойство модели: выходной сигнал имеет малые искажения формы , несмотря на нелинейность модели. При малых и средних уровнях это происходит вследствие высокой добротности звена, которое отфильтровывает продукты нелинейных искажений, а при больших уровнях сами искажения сигнала невелики, вследствие линейности системы.

Исследование влияния вида нелинейности на реакции модели приводит к мысли о том, что если нелинейное звено имеет свойства, характерные для систем автоматической регулировки уровня (АРУ), с переходным процессом в момент резкого включения и выключения сигнала, то такая модель могла бы одновременно обладать и свойствами адаптации.

На рис. 4 показаны амплитудные характеристики модели с нелинейным звеном типа АРУ. Аналогичные характеристики получаются и в экспериментах по исследованию явления адаптации на уровне периферического отдела слуха.

Таким образом, предложенное в настоящей работе звено общей модели гидродинамических процессов в улитке внутреннего уха обладает рядом свойств, характерных для периферического отдела органа слуха.

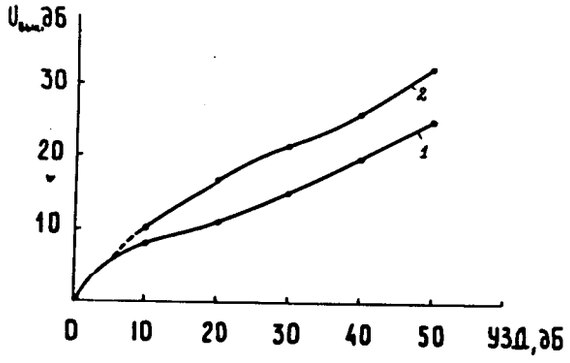


Рис. 4. Амплитудная характеристика элементарного звена с блоком АРУ в качестве нелинейного звена.

По оси абсцисс - уровень входного сигнала, в дБ; по оси ординат - уровень сигнала на выходе звена, в дБ. I- в момент включения посылки тона характеристической частоты; 2 - в установившемся режиме.

ЛИТЕРАТУРА

I. Шупляков В.С. Колебательные свойства структур улитки внутреннего уха. В сб.: "Анализ сигналов на периферии слуховой системы". Л., "Наука", I98I, стр.5-35.

2. Шупляков В.С.Математические модели гидродинамики улитки внутреннего уха. В сб.: "Сенсорные системы". Л.,"Наука", I982, стр. 3-I7.

3. Kemp D.T. Evidence of mechanical nonlinearity and frequency selective wave amplification in the cochlea. 1979, Arch.Otol.Rhinol.Laryngol. 224, p.37-45

4. Sellick P.M., R.Patuzzi, B.M.Johnstone. Measurement of basilar membrane motion in the guinea pig using the Mössbauer technique. 1982. J.Acoust.Soc.Amer., 72, p. 131-141.

LE SYSTEME DE SELECTION ET VALIDATION AUTO-ADAPTATIF VARAP POUR LA RECONNAISSANCE DE LA PAROLE CONTINUE

ANDREEWSKY A. DESI M. RINGOT P.

CNRS - LIMSI - ORSAY BP 30 91406 ORSAY CEDEX FRANCE

The VARAP system has been designed with the goal of being able to select references by using the representativity in a recognition task on a prelabelled data base. This base is analysed in selfconsistency with its own references. Several situations may occur according to whether the analysed references are recognised or not. The selection obtained improves recognition.

VARAP (Validation Autocohérente pour la Reconnaissance Automatique de la Parole Continue) est un système qui séléctionne dynamiquement les références d'un corpus préétiqueté automatiquement, sans utilisation de seuils, en faisant appel uniquement à la capacité de reconnaissance des dites références.

I. LE CORPUS.

Il comprend 700 phrases qui correspondent à 1500 mots différents. L'étiquetage effectué automatiquement est completé par des corrections manuelles qui portent sur environ 10% du corpus.

II. LA DISTANCE.

La comparaison des références (spectres étiquetés) se fait au sens d'une certaine distance que nous allons définir.

Soient deux spectres, l'un noté Oi et l'autre Xi, i variant de 1 a 16 (Oi, Xi sont les valeurs de chacun des spectres sur 16 canaux).

Formons $di = Oi - Xi$ $(i = 1...16)$ soit $Di = di+1 - di$, alors la distance entre deux spectres est donnée par la formule: $D = |Di| + ... + |D16|$

La distance entre deux références tient compte du fait que chaque référence est constituée de trois spectres consécutifs prélevés au voisinage de l'extremum étiqueté et choisis en minimisant leurs distances respectives, l'extremum faisant toujours parti de la référence ternaire. Cependant la disposition relative des spectres par rapport à l'extremum peut être paramétrée.

III. LE MODE DE SELECTION DU SYSTEME VARAP.

La sélection se fait en deux étapes en comparant le fichier CORPUS à un fichier dit VALID construit dynamiquement à partir de CORPUS. Toutefois, étant donné que les temps de calculs sont longs, le fichier CORPUS est partitionné en fractions de 100 phrases. De plus chaque lot de 100 phrases est à son tour partitionné en deux fichiers de références relatives respectivement aux maximums et aux minimums sélectionnés au cours de l'étiquetage automatique. La comparaison dont la description suit peut-être effectuée sur chacun des fichiers partitionnés et il est possible ensuite de concaténer les résultats des comparaisons et de recommencer la sélection sur le fichier concaténé.

III.1. La comparaison.

Dans ce qui suit, le fichier d'entrée sera appelé CORPUS, indépendemment du fait qu'il peut être relatif aux minimas ou aux maximas. On procède comme suit:

A. La première référence de CORPUS est mise dans VALID.

B. Les références suivantes de CORPUS sont alors comparées une à une à toutes les références de VALID. Dans l'expérience effectuée, on affiche les 15 premières références qui constituent ainsi un treillis. Cela n'est possible qu'à partir du moment où il y a déjà au moins 15 références dans VALID.

C. Sur chaque treillis, un scrutin majoritaire est effectué, qui tient compte du rang des références et de leur nombre dans le treillis. Les quatre premiers phonèmes majoritaires sont ainsi retenus.

D. Si le premier des quatre phonèmes retenus est de même valeur phonétique que la référence corpus analysée, alors une opération dite d'incrémentation est effectuée sur les références qui ont contribuées à dégager le premier phonème du scrutin majoritaire. Par exemple, on doit reconnaître le phonème /u/ n°1325, le treillis se présente comme suit:

Phonèmes : /u/ /y/ /u/ /u/ /o/ /y/
N° de ref: 270 323 719 227 520 250

Le résultat du scrutin donne: /u/ n°1325 reconnu 3 fois par /u/, 2 fois par /y/, etc.

Remarque 1.
Le /u/ a été reconnu 3 fois par des /u/ à l'aide de références validées du corpus et qui lui sont antérieures. Cependant des références /y/ par exemple peuvent apparaître par la suite et se révéler plus proches du /u/ n°1325 que les références ayant permis de dégager un scrutin majoritaire favorable. On a donc été contraint de faire un deuxième passage de VARAP afin de pallier à cet inconvénient.

Remarque 2.
Dans VARAP, chaque treillis comprend, outre la suite des références phonétiques, le context gauche et droit de chaque référence du treillis, la distance entre le candidat à reconnaître et les références du treillis, les compteurs d'incrémentation des références du treillis indiquant le nombre de fois où les-dites références ont participé à une bonne reconnaissance en scrutin majoritaire.

E. Si le premier des quatre phonèmes retenus est de valeur phonétique différente de la référence CORPUS analysée alors:
- ou bien la référence CORPUS à reconnaître ne se trouve pas dans le fichier des références validées, et alors elle est introduite dans le fichier VALID. Cela pourra se produire aussi bien pendant le premier ou pendant le second passage de VARAP.
- ou bien la référence CORPUS qu'il faut reconnaître se trouve déjà dans le fichier VALID, alors il y a confusion. Ce cas ne se produira que dans le second passage de VARAP.On effectue alors un nouveau scrutin majoritaire par souci de cohérence avec les incrémentations. Si après se dernier scrutin, la référence est reconnue par un phonème de même valeur phonétique, la validation est dite incrémentale, sinon il y a validation simple.

III.2. Les algorithmes de dépouillement du système VARAP.

Ils comprennent:

A. Un tableau donnant la compression du corpus des références phonétiques, phonème par phonème, après les deux passages de VARAP, ainsi que le nombre total d'incrémentations, de validations, de confusions. Exemple:

Phonème	/i/	/s/
avant	357	400
après	139	163
incrémentation	274	310
validation	32	17
confusion	51	73

B. Un tableau donnant la relation entre la confusion et l'incrémentation par ordre décroissant de cette dernière. Il montre en particulier qu'à partir d'une incrémentation égale à dix, les confusions deviennent très faibles: 12 confusions en tout pour l'ensemble du corpus, et à partir de 20 incrémentations, il y a rarement plus de une confusion.

C. Une matrice de confusion a été construite automatiquement et nous la commentons brièvement dans la conclusion.

D. Un fichier inverse a été construit automatiquement, permettant de savoir combien de fois chaque référence a participé à une bonne et à une mauvaise reconnaissance.

IV. BILAN GENERAL DU SYSTEME VARAP ET CONCLUSION.

Les expériences de validation par autocohèrence ont été faites sur 100, 300, 500 phrases. L'augmentation de la taille du corpus apporte une amélioration régulière. Nous donnons les résulats sur 500 phrases. Après le premier passage de VARAP, sur 9952 références du corpus, il y a eut 4810 incrémentations et 4002 validations. Après le second passage de VARAP, le nombre de références conservées est de 4375, il y a 5973 incrémentations. 644 nouvelles validations et 2195 confusions.

La matrice de confusions montre qu'en ce qui concerne les voyelles, si l'on écarte les confusions du type /e/, /ɛ/ ou encore /ə/, /ø/ ou encore /ɔ/, /o/ ou encore /a/, /ɑ̃/, alors les confusions sont peu importantes. Pour les consonnes, la matrice fait apparaître des confusions importantes entre /p/, /t/, /k/, d'une part et /b/, /v/, /d/ d'autre part. Cela tient en partie au mode d'obtention des références ternaires qui sont obtenues en regroupant les spectres les plus proches les uns des autres, c'est à dire en conservant la partie du minimum la plus stable au sens de la distance utilisée. Mais dans ce cas, les spectres de la référence sont regroupés autour du silence de l'occlusive. Pour pallier à cet inconvénient, un traitement complémentaire est prévu, permettant dans le cas des occlusives sourdes et sonores, de conserver les références ternaires en remontant vers les transitions, après ou avant le minimum.

Pour terminer, disons que le but du système VARAP est multiple:

- Vérifier la qualité de l'étiquetage automatique utilisé.

- Construire un système de sélection qui ne fasse pas appel à des seuils et qui teste l'aptitude à reconnaître des références étiquetées du corpus.

- Avoir un outil modulaire complet qui, entre la procédure et la constitution du dictionnaire de références, permet de paramétrer toutes les décisions intermédiaires. Entre autre l'adaptation au locuteur, l'utilisation de plusieurs types de distances différentes, enfin la conservation ou la mise à l'écart des références de l'apprentissage en fonction de leur aptitude à reconnaître.

BIBLIOGRAPHIE.

ANDREEWSKY A. DESI M. POIRIER F. "Le Système SHERPA - de l'étiquetage automatique à la reconnaissance par analyse ternaire", 5ème Congrès RFIA, 1985.

DESI M. POIRIER F. "Le système SHERPA: étiquetage et classification automatique par apprentissage pour le décodage automatique de la parole continue", Thèse en Sciences, Paris-Sud, Orsay, 1985.

DABOUZ M. "Transmission de la parole à faible débit par vocodeur à classification", Thèse de l'ENST, janvier 1984.

DILL A.M. "Extensions de la méthode des plus proches voisins". Thèse de l'Université Claude Bernard, Lyon, 1978.

MAC QUEEN J. "Some methods for classification and analysis of multivariate observations", The 5th Berkeley Symposium of Mathematical Statistic and Probability, 1967.

MICLET L. VICARD D. "Reconnaissance des parties stables de parole continue pour le décodage accoustico-phonétique", 15ème JEP, Aix en Provence, 27-30 mai 1986.

SAPHIR-1: SYSTEME MULTILOCUTEUR COMPRENANT LES PHRASES PARLEES CONTINUES

GRIGORY SLUTSKER

L'Institut d'agriculture d'URSS
B/A 8, Balachikha-9,
Moscou 143900, URSS

SERGUEI KRINOV

L'Institut des problèmes à
transmettre l'information
de l'A.S. d'URSS, GSP-4,
Moscou 101447, URSS

RESUME

Avec le SAPHIR-1, on ne doit pas dicter les mots à l'étape d'apprentissage. Ici, on a un programme de transcription automatique qui compose les étalons phonétiques des mots selon leur forme orthographique comprenant des sygnes d'accentuation nécessaires. Le programme peut être adapté à chaque langue nationale. C'est un expert qui fait l'adaptation, et il n'est pas obligatoire qu'il sache programmer. Les connaissances d'expert s'inscrivent déclarativement en forme des règles de transcription des textes à l'aide d'une métalangue inventée spécialement pour le système. Au régime de compréhension automatique des phrases parlées continues, on garantit un codage phonétique adéquat au niveau de segments.

INTRODUCTION

La plupart des systèmes à reconnaître la parole se fonde sur une procédure d'apprentissage, pendant laquelle le système forme les étalons des mots ou des phrases qui seront employés plus tard. Au cours d'apprentissage l'usager doit dicter tous les mots, et le nombre de prononciations de chaque mot varie de 1 à 5 pour les systèmes divers. Selon des donnés des firmes américaines produisant les appareils à reconnaître la parole, on ne peut assurer la reconnaissance que dans le cas où l'on prononce chaque mot pendant l'apprentissage 3 fois ou plus [1]. Selon les donnés, on sait aussi que la plupart des usagers refuse travailler avec le système si le temps d'apprentissage dépasse 30 min. Il est évident que ces difficultés retiennent l'application vaste des appareils à reconnaître la parole.

Bien des investigations des années dernières ont pour but l'élaboration des systèmes multilocuteurs. Le progrès dans le domaine dépend évidemment des possibilités d'accomplir une analyse phonétique de la parole indépendant de locuteur. Ça simplifie essentiellement le problème d'apprentissage, mais ne le résout pas. Car on doit tout de meme dicter préalablement le vocabulaire, ce qui conserve tous les défauts de l'apprentissage (les défauts principaux sont: des fautes phonétiques incontrolées et des bruits acoustiques qui peuvent déformer les étalons).

SAPHIR-1 permet se passer de microphone et de locuteur à l'étape d'apprentissage. Les mots du vocabulaire en forme orthographique munis des signes d'accentuation s'introduisent dans le système et un programme spécial les transforme aux étalons phonétiques. Pour représenter les étalons, on a choisi le niveau des segments phonétiques, ce qui réduit essentiellement le volume du mémoire nécessaire et le temps des calculs à l'étape de reconnaissance, en comparaison avec une représentation paramétrique. En outre, le procès de segmentation normalise le signal parlé dans le temps, car le nombre de segments et nombre de sons du mot sont approximativement les mêmes. La transformation automatique "orthogramme — conséquence de segments" se fait selon les règles reflétant les lois de prononciation de la langue en question. On tient aussi compte des charactéristiques techniques du segmenteur et du marqueur qui effectuent l'analyse acoustico-phonétique dans le système.

La méthode permet représenter le vocabulaire à apprendre en forme de morphèmes. La représentation par éléments est irréalisable aux systèmes exigeant l'enrégistration phonique préalable, car on ne peut pas prononcer certains éléments en isolation correctement (tels sont les morphèmes inaccentués). Le variant du vocabulaire réalisé en SAPHIR-1 présente plusieurs morphèmes inaccentués, par exemple, l'élément ЗАР qui a dans la grammaire le meme droit que le mot ЗА'РАБОТНАЯ en combinaison avec le mot ПЛА'ТА.

Dans le SAPHIR-1, c'est la programmation dynamique qui aide à choisir l'étalon le plus proche au signal d'entrée sur le niveau segmentatif. L'emploi de la programmation dynamique pour la reconnaissance automatique de la parole a été premièrement proposé en 1968 par un des auteurs du rapport [2].

En ce qui concerne les pricipes de l'analyse acoustico-phonétique accomplis dans le SAPHIR-1, on les a vérifiés aux systèmes multilocuteurs à reconnaître les mots isolés [3].

FORMATION AUTOMATIQUE DES ETALONS PHONETIQUES

Créant l'algorithme de transformation "orthogramme → conséquence de segments", on est tombé sur les difficultés suivantes. L'algorithme reflétant toutes les variations phonétiques de la langue en question et toutes les charactéristiques nécesaires de l'analyseur exige un programme au grand nombre d'opérateurs de transfert conditionnel. Les programmes de la sorte sont difficiles à mettre à point. Il est encore plus difficile de changer le programme. En réalité, le besoin de changement peut surgir, par exemple, dans le cas de mise en vue des lois phonétiques supplémentaires ou dans le cas d'emploi d'un analyseur acoustico-phonétique ayant un nouveau alphabet de segments. En outre, l'algorithme reflète l'organisation phonique d'une seule langue nationale.

Ce qui serait le mieux, c'est un algorithme souple qui permet modifier ou meme remplacer les règles de transformation, par exemple, au changement de la langue nationale en question. Pour atteindre ce résultat, on peut s'orienter aux connaissances d'expert prises comme un massif d'entrée pour le traitement des orthogrammes des mots. Un expert-phonéticien introduit ses connaissances en forme d'un ensemble ordonné des règles déclaratives de transformation. Il ne doit pas savoir programmer, car on emploie ici une métalangue adaptée pour l'inscription spectaculaire et l'interprétation automatique des règles. Donc, le programme à transformer les textes c'est un programme "comprenant" la métalangue [4].

Cette méthode augmente la souplesse du système et le programme meme devient universel. On entend l'universalité comme ça.

1. Comme l'alphabet de sortie, on peut prendre tout alphabet nécessaire décrit par les règles de transformation - seraient-ce les codes phoniques, les symboles de la transcription phonétique internationale, les symboles alphabétiques d'une autre langue nationale, etc. On peut employer la possibilité dans les manuels de conversation (la formation automatique de prononciation), pour la translittération, etc. On peut aussi user le programme aux systèmes de synthèse "texte — parole" (dans ce cas, l'alphabet de sortie consiste des codes à diriger le synthétiseur du sygnal parlé).

2. Les règles nécessaires étant données, on peut traiter un texte écrit en langues différentes.

METALANGUE POUR LES REGLES DE TRANSFORMATION

La métalangue à inscrire les règles de transformation suffit pour transformer un texte en forme phonétique et pourtant elle est très simple et évidente. Pour l'employer, on ne doit pas savoir programmer.

L'aspect général d'une règle de transformation est [A_B], où A est symbole ou chaîne de symboles, B est symbole ou chaîne de symboles qui est le résultat de transformation d'A. La chaîne A peut correspondre à un fragment d'orthogramme (fragment en code d'entrée) ou à un fragment d'une représentation intermédiaire. La chaîne B correspond à un fragment d'une représentation intermédiaire ou à un fragment en code de sortie.

Les représentations intermédiaires ne sont qu'auxiliaires; ce sont elles qui aident à donner une classification traditionnelle des symboles - par exemple, une classification selon la mode ou le point d'articulation; à l'étape on peut aussi introduire les marques spéciales pour les voyelles pré- et post-toniques, etc.Pour l'expert-phonéticien, la représentation intermédiaire de la sorte est facile à interprendre, car elle reflète les lois phonétiques de la langue en question. En cas de nécessité, les règles de transformation forment une image phonétique très détaillée - on peut p.e. décrire le charactère supposé des trajets de formants aux frontières "consonne-voyelle", etc. Dans le SAPHIR-1, ayant un analyseur acoustico-phonétique à un 16-segment-alphabet, les possibilités de la représentation intermédiaire sont limitées.

Les chaînes A et B des règles sont séparées par "_"; les règles sont toujours mis aux crochets. Tout ce qui est outre les crochets n'est qu'un commentaire. Les commentaires font le système des règles plus évident; l'ordre des commentaires n'est pas fixe. La mode de la mise en ordre des règles peut être choisi librement. P. e., on peut les donner en forme d'une table dont les lignes et les colonnes unissent les règles selon tels ou tels indices; les titres des lignes et des colonnes peuvent présenter les commentaires - comme les symboles des processus d'assimilation (d'assourdissement, d'avoisement, etc.).

Pour qu'on puisse écrire les règles en forme générale tenant compte d'analogie propre à certaines transformations,on a introduit des variables pour nommer les symboles de texte.

N'IMPORTE QUEL SYMBOLE. Le symbole © marque n'importe quel symbole - p.e., la règle [A©B_AB] élimine tout symbole placé

dans le texte entre A et B.

N'IMPORTE QUEL SYMBOLE D'UN ASSORTIMENT. La règle [<ABCDE>F_@Z] remplace le symbole F par le symbole Z si F succède à n'importe quel symbol de l'assortiment mis aux parenthèses angulaires. L'introduction d'une variable (marquée par parenthèses angulaires) permet généraliser les règles analogiques. On peut introduire dans une règle jusqu'à 10 variables; dans ce cas, on leur donne les index de 0 à 9. P.e., la règle [<AB>T<CDE>1F_@1@F] décrit les transformations éliminant le symbole T et transposant les symboles qui l'entourent.

N'IMPORTE QUEL SYMBOLE À L'EXCEPTION DES SYMBOLES NOMMÉS. Une variable de la sorte peut être désignée par ٦...> ; le signe recoit le nom des parenthèses avec la negation. La règle transforme le contexte sous la condition d'absence des symboles entre parenthèses avec la négation - p.e., [B٦UVW>_ZF@] ou [RA٦UVW>_RAW@] . Fixée par le contexte, la signification d' une variable peut faire partie d'une autre variable:[<JLNZ>@1<JLNZ>2?@1>3_@@1w@2@3].Les types des règles démontrés au-dessus permettent décrire toute transformation du texte en n'importte quelle langue.

CLASSIFICATION DES SONS DE LA PAROLE.

Pour qu'on puisse résoudre le problème de transformation ORTHOGRAMME →DESIGNATIONS PHONETIQUES, on doit paramétriser adéquatement la description des sons de la parole correspondant aux lettres. Tenant compte des possibilités de l'analyseur du SAPHIR-1, on a choisi une application de deux dimensions convertissant des symboles orthographiques à une représentation intermediaire.

Tous les symboles désignant les voyelles comprennent soit U (voyelle tonique) soit V (voyelle première prétonique ou atonique), soit W (voyelle réduite). Parfois, les réduites n'ont pas de correspondance avec les symboles orthographiques - ce sont d'habitude les intercalations réduites entre les consonnantes voisines.

Si la transcription automatique exige qu'on donne une charactéristique plus détaillée de la dépendence existant entre l' accent et les qualités des voyelles, on peut le faire à l'aide de la méthode proposée. On doit introduire les désignations pour les types des sons en question et composer les règles nécessaires.

Les lettres correspondant aux consonnes, on les remplace par une couple de symbole dont le premier signifie un des huit modes d'articulation possibles. L'autre montre un des quatre points d'articulation propres aux consonnes. Le codage de la sorte permet donner les règles dans une forme généralisée et évidente.

Dans la table ci-dessous, on peut voir les transformation faites srlon la méthode pour les consonnes russes. Les symboles de la table ont les significations

	F	D	A	*	G
Q	[П_QF]	[Т_QD]			[К_QG]
J	[Б_JF]	[Д_JD]			[Г_JG]
L	[В_LF]		[Л_LA]	[Й_L*]	
N	[М_NF]	[Н_ND]			
R			[Р_RA]		
Z		[З_ZD]	[Ж_ZA]		
S	[Ф_SF]	[С_SD]	[Ш_SA]	[Щ_S*]	[Х_SG]
C		[Ц_CD]		[Ч_C*]	

suivantes: Q - les occlusives sourdes, J - les occlusives voisées, L - les liquides, N - les nasales, R - les vibrantes, Z - les constrictives voisées, S - les constrictives sourdes, C - les affriquées;

F - les labiales, D - les dentales, A - les alvéolaires dures, * - les alvéolaires molles, G - les postpalatales.

Le groupe des règles, y compris les titres de la table, se trouve dans le massif d' entrée du programme de transformation. Le programme prend les titres comme commentaires et ne travaille pas. La table montre que les lettres russes correspondant aux consonnes n'épuisent pas toutes les positions. Une autre langue nationale peut donner une autre disposition des symboles dans la table par exemple, la " " ucrainienne doit être présentée comme "ZG" Pour d'autres langues, on peut proposer des tables variant le nombre des lignes et des colonnes.

Maintenant on peut citer les règles d'assourdissement et d'avoisement - par exemple, [J@<;;:. >1_Q@@1] (l'assourdissement des voisées à la fin du mot); [S@<JZ>1_Z@@1] (l'avoisement des constrictives sourdes devant les voisées - occlusives et constrictives).

MISE EN ORDRE DES REGLES TRANSFORMATIVES.

Toutes les règles de l'assortiment sont mises en ordre; on peut les diviser en quatre groupes.

Le premier groupe consiste des règles générales qui agissent tout d'abord,avant la transition à la représentation intermédiaire de deux dimensions, et des règles transformatives pour les voyelles - par exemple,[<шцж>И_@ы] (le remplacement d' "И" par "ы" après "ш","ц","ж" en russe).

Le seconde groupe comporte les transformations des consonnes; la plupart des règles est présentée dans la table. En outre, dans le groupe entrent les règles d'avoisement, d'assourdissement et de fusion des consonnes, les règles d'intercalation des réduites entre les consonnes voisées et d'intercalation des sourdes aspirées entre les consonnes sourdes, etc. Ce sont 80 règles environ qui font partie des deux groupes. Pour la langue en question, un expert-phonéticien introduit les règles à la fois: on n'a pas besoin de renouvellement de l'information.

Le troisième groupe permet la transition aux codes des segments phonétiques, qui peuvent être fixés par le segmenteur automatique du système.

Enfin, c'est le quatrième groupe des règles qui contrôle le fonctionnement du programme transformatif. Les règles font la transformation inverse qui donne à la sortie les orthogrammes initiales. On n'emploie pas les règles à former les étalons phonétiques des mots - elles ne servent qu'à contrôler.

D'AUTRES NIVEAUX DU SAPHIR-1.

Le SAPHIR-1 a 10 niveaux hiérarchiques: celui acoustique, celui paramétrique, segmento-phonétique, orthographo-morphologique, lexique, sémantique, syntaxique, phraséologique, dialogique, pragmatique. Donc, le niveau acoustique et celui pragmatique sont deux bouts inverses du système dont le principe de fonctionnément est suivant.

Etant donné les connaissances sur tous les niveaux obtenues à l'aide des experts-spécialistes, le SAPHIR-1 est prêt à comprendre les phrases en chaque situation de dialogue homme-machine avec n'importe quel locuteur.

D'une part, chaque phrase prononcée par celui-ci passe à travers les trois premiers niveaux de traitement et le signal parlé se transforme en une suite des segments phonétiques. Simultanément de son autre bout, le système engendre généralement parlant toutes les phrases (en meme forme phonétique) qui correspondent à la situation du dialogue.

Deux mouvements inverses se rencontrent au niveau phonétique à comparer les successions de segments. La comparaison s'effectue rapidement selon le graphe syntaxique à l'aide de la méthode DP [2] modifiée à reconnaitre la parole continue. Ainsi, le SAPHIR-1 trouve une phrase engendrée la plus proche à celle prononcée.

Au niveau lexique, chaque orthogramme s'accompagne par un code sémantique comprenant 4 symboles par correspondance au domaine thématique en question, la classe de notion, l'objet concret de la classe et les charactéristiques grammaticales de la forme. De cette façon, le vocabulaire du système consiste des couples "orthogramme-code sémantique". Cela permet différer entre eux les HOMONYMES coïncidants phonétiquement mais qui se distinguent par leurs codes sémantiques. Et au contraire, les SYNONYMES ont presque même code sémantique mais ils ont des formes divers.

Au niveau syntaxique à chaque situation du dialogue, l'usager peut exploiter un sous-ensemble des phrases qui est déterminé par la grammaire situative correspondante. Celle-ci est représentée par un graphe. Le noeud de sortie de celui-ci tient toutes les variantes des phrases de la grammaire. Après la fin du résonnement d'une phrase prononcée par l'usager, le SAPHIR-1 prend le résultat final au noeud de sortie et trouve la trajectoire la plus vraisemblable le long du graphe à reconstruire la suite des mots. Ce sont les orthogrammes des formes que l'on use pour l'extraction du texte à l'écran (ou pour la synthétisation de la parole) et la succession des codes sémantiques à calculer le sens de phrase.

Le Système Automatique à comprendre des PHrases contInues paRlées (SAPHIR-1) etait réalisé à la base de l'ordinateur du type PDP 11/40 [5] .

La sureté de reconnaissance de la parole continue pour n'importe quel locuteur n'est pas moins que 95%. Le vocabulaire d'exemple consiste de 200 formes. Son volume n'est pas restreint par les possibilités du système.

BIBLIOGRAPHIE.

1. Meng B. Speech recognition: not a typical engeneering problem. Digit Des., 1985, n6, p. 49 - 57.
2. Слуцкер Г.С. Нелинейный метод анализа речевых сигналов. Труды Научно-исследовательского института Радио (НИИР),М.,I968г. вып.2. с. 76 - 82.
3. Кринов С.Н., Савельев В.П., Цемель Г.И. Классификация сегментов при распознавании устных команд. Труды АРСО-I3,Новосибирск, Институт математики СО АН СССР,I984,с.IOI-IO3.
4. Кринов С.Н., Слуцкер Г.С. Автоматическое формирование звуковых эталонов слов по их орфографической записи. Труды АРСО-I4, ч.I, Каунас, I986, с.86.
5. Кринов С.Н., Слуцкер Г.С. Многоуровневая речевая диалоговая система "САПФИР". ТРУДЫ АРСО-I4, ч.I., Каунас, I986., с.92-94.

KEY FEATURES IN CONTINUOUS SPEECH

Mary O'Kane and David Mead

School of Information Sciences and Engineering
Canberra College of Advanced Education

ABSTRACT

'Key Features' is a term coined to indicate blocks of continuous speech which have been recognised with absolute certainty by computationally efficient techniques as having some particular acoustic-phonetic attribute. Key feature recognition is also characterised by no false positives although certain blocks having the features might be missed.

This paper describes the use of the concept of key features as a pivotal element in a project to build a dictation machine accepting continuous speech. A method of locating key features such as voiced speech, voiceless speech, stressed speech, nasals, liquids, plosive bursts, intervocalic voiced plosives and fricatives, voiceless fricatives and the phoneme /s/ is presented and the results of attempting to locate these key features in a passage read by several speakers are given.

INTRODUCTION

In this paper we address the notion that blocks of continuous speech which are recognised with absolute certainty as having some acoustic-phonetic attribute can be used as an integral part of the primary control mechanism of a dictation machine accepting continuous speech input. After describing the role of the key feature concept in an expert's reasoning about an unknown speech passage we go on to describe how reasoning using the key feature concept might be used in building a dictation machine. We then give one method of locating a set of key features and discuss the usefulness of this set of features for the dictation machine design.

A WAVEFORM-READING EXPERT

The aim of the FOPHO continuous speech recognition project [1] is to build a speech recognition system using expertise-capture techniques - in this case the expertise being captured is that of a phonetician transcribing a foreign language. While it is customary to think of the phonetician's expertise as being primarily auditory expertise, experiments by Cole, Rudnicky, Zue and Reddy [2] have demonstrated that certain phoneticians and speech scientists have considerable visual expertise in that they can 'read' spectrograms. In interviewing the expert phoneticians working on the FOPHO project another area of visual expertise was explored - that of 'reading' waveforms [3]. One of the expert phoneticians (P.R.) was particularly good at this and although he generally could not produce a full phonetic transcription of what was said from the waveform alone, he could provide a remarkable amount of information about the phonetic content of the waveform in question. Presented with a section of waveform from an unknown utterance P.R. would first make a series of categorical statements about portions of the waveform which he could immediately identify as having some particular acoustic-phonetic feature. The features identified were sometimes phonemes, a very easy to identify such phoneme being /s/, but very often they were broader phonetic features as as 'voiced', 'nasal' or 'plosive burst'. After making categorical statements P.R. would go on to make a series of more tentative statements about the waveform indicating phonetic features that he believed were probably but not certainly present. However, it is P.R's categorical labellings that are of most interest in this paper. The waveform reading expertise was encapsulated in a set of production rules which were based on a very efficient signal processing technique (described below) which addressed the same waveform primitves that the phonetician used in providing the rationale for his categorical labelling decisions. Initially these rules were merely added to the FOPHO system's knowledge base. However later it was decided that this set of categorical rules might be useful in the primary control mechanism of a dictation system that we proposed building.

A DESIGN FOR A DICTATION MACHINE

The proposed dictation machine, Dicma, is a machine designed to be used for commercial correspondence where a high proportion of the words are predictable. Thus the dictation machine design can make heavy use of a modified word- and phrase-spotting technique. The top-level concept postulated in the Dicma design is that the 'pure' recognition part of the system should produce some form of phonetic encoding (the form of which is discussed below) of the input speech and that this encoding should then be searched for indications the presence of words and phrases which are likely to occur in the dictated material. Sources of such likely-to-occur words are extrapolations from typed-in keywords and studies of the particular user's past correspondence. (For more details of methods used in the prediction of likely words in a dictated passage see [4]). After possible locations of predicted words have been found the presence of these words can be verified or rejected using a test-and-eliminate strategy. A predictive parser can then be used to predict grammatically suitable fillers for the undecoded speech between verified words. In their turn these fillers can be acepted or rejected using the test-and-eliminate strategy.

The basic idea here is not new. The notion of predicting and then searching for likely words in a section of speech was fundamental to the ARPA speech understanding projects [5]. What is being postulated in this paper however is that these ideas can, by a judicious choice of reasoning technique, be pushed a long way for relatively low computational effort and thus enable the production of efficient, low-cost, special-purpose speech recognition devices.

In order to maximise the efficiency of operation of the dictation machine described above we adopted a new structure for the output from the 'pure' speech recognition system component. Generally FOPHO has run according to what is quite a common top-level approach to continuous speech recognition, that is a hierarchial refinement scheme derived from formal phonetic classification theory. An example of this approach is to first classify an unknown sound as either sonorant or non-sonorant, then if it is non-sonorant to see if it is continuant or interrupted and so on. A detailed exposition of this approach has been given by De Mori, Laface and Piccolo [6]. This approach has been generally accompanied by some probabilistic or fuzzy weighting scheme for estimating a degree of belief in any particular classification at any particular level in the hierarchy. However the hierarchical-classification -cum-fuzzy-weighting scheme does not allow us to take advantage of strong categorical inferencing techniques. The issue here is that recognising a particular feature in a stream of speech with near 100 per cent certainty is not nearly as strong a statement as saying that a particular feature has been recognised categorically as being that particular feature. On the basis of this observation we have decided to use two types of reasoning mechanisms in the dictation system - one categorical (or pattern-matching) and the other fuzzy. Categorical reasoning is to be used in likely-word location and a mixture of categorical and fuzzy reasoning is to be used in the test-and-eliminate strategy. This approach is essentially similar to the control mechanism used in many medical expert systems [7].

AN ADEQUATE PHONOLOGICAL ENCODING

To use categorical reasoning for likely-word location the continuous speech input to the system and the set of likely words that are to be searched for must both be encoded according to some robust and adequate phonological encoding scheme. It must be robust in the sense that false encodings must not occur and it must be adequate in the sense that too many ambiguous word locations must not occur. But what constitutes a suitably robust and adequate encoding? First we discuss what might constitute an adequate phonological encoding. Various recent studies on the distributional characteristics of word cohorts that result from encoding complete dictionaries of words according to various phonological encodings (see [8] for an overview) are relevant to this problem. Lai and Attikiouzel [9] carried out a cohort study of Australian English using all the complete (51,018) words of the Macquarie Dictionary [10] as their source. They found that for the various phonological encodings they studied (two of which are close to an encoding we consider below) the expected cohort size is quite small (less than 5) for words of phonetic length seven or greater, certainly small enough to be easily distinguishable with the addition of simple verification techniques. However for words of phonetic length two to six the expected cohort size is rather too high to be easily settled with verification techniques, particularly when it is remembered that the problem we are considering is that of finding words in a continuous stream of speech where false positives can occur across word boundaries.

Accordingly we postulated that a small, but likely vocabulary (such as we would have if we knew the most commonly-occurring words in a user's correspondence) would give rise to a manageable set of words which, when searched for in the input string, should lead to a correct decoding of a fair percentage of the input. To investigate this we carried out a word-frequency study of 33 consecutive letters written by the first author. It was found that 83 words could be classed as high-frequency words [11]. Under quite a weak phonological encoding such as the following:

(voiced), (unvoiced), (vowel),
(nasal), (/s/), (/p/), (/t/), (/k/),

the 83 high-frequency words gave rise to 56 words cohorts, 45 of which were unique. The largest cohort was a five word cohort of phonetic length two. The 56 cohorts were searched for (using rast Bibliographic search techniques [12]) in ten test letters also extracted from the writer's correspondence and also encoded using the eight-class phonological encoding. The high-frequency words accounted for 57% of the words in these letters. A 'best interpretation' of the search on a particular letter was defined to be a reading of the letter that was obtained by assuming that that locations of words of high phonetic length were more likely to be correct than locations of words of low phonetic length. Word locations were not allowed to overlap. With this notion of interpretation, 52% (or 73% if the simplifying assumption of exact phonetic length is made) of the high-frequency words that occurred in the test letters were in the best interpretations of the letters and every high-frequency word was at least in the best or a second-best interpretation. This result together with the smallness of the cohort sizes involved means that even with this weak phonological encoding the verification task to check for correct words and eliminate false positives is quite straightforward. Also as there were no false positives for words of phoentic length six or greater it is probably unnecessary to apply verification to words of this length. Thus for

the environment in which the proposed dictating machine would operate only quite a weak phonological encoding would seem to be adequate to ensure computational efficiency in the operation of the first phase of the machine. But the question that still needs to be addressed is: can such a phonological encoding be achieved robustly for continuous speech?

In the next sections we address this question; first describing the role that key features could play in providing this robust encoding, then giving a means of finding key features and finally discussing the results of running key feature rules over passages of speech read by several speakers.

ABSOLUTE KEY FEATURES

What we call a key feature is a block of continuous speech which has been recognised with absolute certainty as having some particular acoustic-phonetic attribute. In particular, although some sections of speech having this acoustic-phonetic attribute may be missed after the application of key feature labelling rules, one can be certain that there are no false positive labellings, i.e. that no block has been incorrectly labelled. For the categorical reasoning phase of the operation of the proposed dictation machine however we need to find a sub-set of the key features that has the additional property of no false negatives, i.e. all blocks of those features will be found. We will refer to these as absolute key features.

Generally however it should be noted that the indisputable certainty of a key feature label enables immutable anchor points to be established in speech. Around these anchor points a range of hypotheses suggested by phoneme or higher-level prediction rules can be tested using either categorical or fuzzy inferencing. The key feature labels also provide a context which allows context-dependent recognition inferences to be made.

LOCATING KEY FEATURES

When giving explanations for his categorical labellings of printed speech waveforms the phonetician P.R. generally couched them as as arguments about two waveform primitives - zero-crossings per unit time and waveform amplitude. We developed a set of simple waveform analysis techniques which efficiently produces a measure of these two waveform primitives simultaneously. The fundamental procedure of this technique can be described as follows:

> The sum of the absolute values of the valley and peak waveform amplitudes for each adjacent valley-peak pair is calculated as a function (called W1) of the time mid-way between each valley-peak pair.

By repeating this procedure twice on the successive outputs from the procedure we obtain a function (W3) which is similar to the waveform energy [3].

A second procedure is derived from W1. It is referred to as M1 and is obtained by taking the inverse of the time between adjacent W1 points as a function of the mid-point in time between those points. This function can be displayed in a frequency versus time graph.

The third style or procedure used is a function which averages the M1 points for pre-defined window sizes and sampling rates. This function gives (with appropriate choice of sampling rates) a rough guide to various formant trajectories [13]. Graphs of W3 and averaged M1 (for a sampling rate of 20,000Hz) for a female speaker saying "insects may be" can be seen in figure 1. The three procedures described above are all computationally extremely fast.

Several key feature location rules were written which were based on the output from these procedures but which reflected the categorical labellings of the waveform-reading phonetician. This set included rules for the following phonetic labels 'heavy stress', 'voiced', 'voiceless speech', 'nasal', 'liquid', 'voiceless fricative', 'inter-vocalic voiced plosive or fricative', 'plosive burst', 'syllabic peak', 'not high front vowel' and '/s/' and for various broad vowel categories. These rules were written in what is essentially a production rule form. For example the rule for a nasal is the following:

```
label
   name        :  nasal,
   wave        :  speech,
   requires    :  [M1(20000)],
   association :  M1(20000) is long_low_amplitude
                                        (3,300)
end.
```

These rules are written in a special language front-end to Prolog. The declarative style of the rules makes them easy to construct and de-bug. Figure 1 is an example of a typical screen display

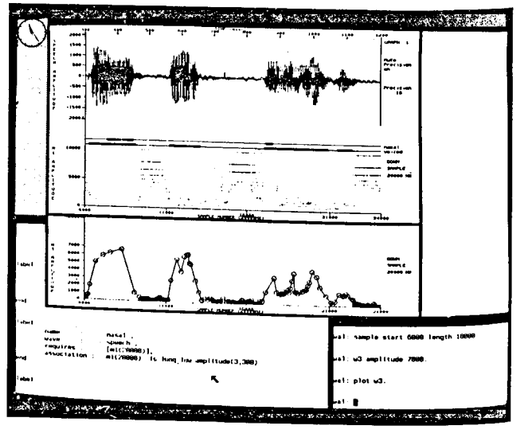


Figure 1: Multi-window screen display for key feature rule system. The top signal processing window displays averaged M1; the bottom signal processing window displays W3. Key features are marked by heavy horizontal lines in the top window.

seen during the running of the rules on a SUN workstation. The central windows are used to graphically display recognition results - the dark lines in the centre show where the key features, 'voiced' and 'nasal' are located - and signal processing results while the window in the right-hand corner is the command window and the window in the left-hand corner is an editor window, the presence of which allows for rapid prototyping of new rules. The various key feature rules written to date were tested on 32 speakers (15 male and 17 female) reading lists of words and on six speakers (three male and three female) reading a short reading passage. Only four absolute key features were found - 'voiced', 'voiceless speech', '/s/' and 'heavy stress'. 92% of the voiceless fricatives and 78% of all the nasals present were located by the appropriate key feature rules. Other key features were located less than 50% of the time using such rules. For key features that are not absolute key features the results varied considerably from speaker to speaker. In particular one of the speakers reading a passage produced plosive bursts less then 10% of the time. However it should be noted that under the condition of stressed speech the key feature 'nasal' is always located 100% of the time.

A ROBUST PHONOLOGICAL ENCODING

From these results it is clear that the current key feature rules would not yield a very strong phonological encoding. The best robust encoding that could be got from the present versions of the key feature rules is:

(voiced), (unvoiced), (stressed vowel), (/s/).

This encoding is even weaker than the eight-class encoding discussed in a previous section. Nevertheless even this encoding does not result in an impossible number of cohorts particularly for words of phonetic length greater than six. Also the fact that at least one key feature becomes an absolute key feature in stressed speech suggests that several encodings should be used when likely words are being searched for. Thus after all key features have been found two phonological encodings of the unknown passage of speech could take place - one the phonological encoding for stressed speech and the other the (weaker) phonological encoding for unstressed speech. The words being searched for could be similarly bi-encoded. First the stressed version could be searched (this would be the faster and more productive search) and then the second, weaker encoding of the unlocated portions could be searched with the weaker encoding of the searched-for words. After that key features which were not explicitly used in the two encodings could be used to eliminate some false word locations. After this still, fuzzy verification strategies could be used.

Furthermore it should be emphasised that key features do not have to be located by the means given in this paper. Any speech segmentaton rule, based on any form of signal processing, that results in no false positives is a key feature rule. Thus a re-examination of speech segmentation studies would doubtless yield a wide range of key feature rules some of which might be absolute key feature rules and this might give rise to stronger phonological encodings for use in the categorical reasoning process.

CONCLUSION

In this paper we have argued that relatively fast and unsophisticated speech processing should yield reasonable recognition rules if categorical reasoning strategies are used.

REFERENCES

[1] M. O'Kane, 'The FOPHO Speech Recognition Project', Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, 1983, pp.630-632.

[2] R.A. Cole, A.I. Rudnicky, V.W. Zue & D.R. Reddy, 'Speech as patterns on paper', in Perception and production of fluent speech, R. Cole (ed.), Lawrence Erlbaum Associates, Hillsdale, 1980, pp.3-50.

[3] M. O'Kane, J. Gillis, P. Rose & M. Wagner, 'Deciphering speech waveforms', Proceedings of IEEE-IECEJ-ASJ International Conference on Acoustics, Speech & Signal Processing, Tokyo, April 1986, pp.2227-2230.

[4] M. O'Kane, D. Mead, J. Newmarch, K. Byrne & R. Stanton, 'The DICMA Project', Proceedings of the First Australian Conference on Speech Science and Technology, Canberra, November 1986, pp.278-283.

[5] D.H. Klatt, 'Review of the ARPA Speech Understanding Project', J. Acoust. Soc. Amer., 62, 1977, pp.1345-1366.

[6] R. De Mori, P. Laface & E. Piccolo, 'Automatic detection and description of syllabic features in continuous speech', IEEE Trans. Acoustics Speech & Signal Processing, ASSP-24, 1976, pp.365-378.

[7] P. Szovolvits & S. Pauker, 'Categorical and probabilistic reasoning in medical diagnosis', Artifical Intelligence, 11, 1978, pp.115-144.

[8] V.W. Zue, 'The use of speech knowledge in automatic speech recognition', Proceedings IEEE, 73, 1985, pp.1602-1615.

[9] E. Lai & Y. Attikiouzel, 'A comparison of several coarse phonetic classification schemes', Proceedings of the First Australian Conference on Speech Science and Technology, Canberra, November, 1986, pp.361-321.

[10] The Macquarie Dictionary, rev. ed., Macquarie Library, Bee Why, 1985.

[11] M. O'Kane & D. Mead, 'On the feasibility of a continuous speech dictation machine', Technical Report No. 14, Canberra College of Advanced Education, 1987.

[12] A. Aho & M. Corasick, 'Efficient string matching: an aid to bibliographic search', Comm. ACM, 18, 1975, pp.333-340.

[13] M. O'Kane & J. Gillis, 'Efficient derivation of formant-like information from speech waveforms', Proceedings of the First Australian Conference on Speech Science and Technology, Canberra, November 1986, pp.322-327.

SPEECH RECOGNITION SYSTEM ON A MICROCOMPUTER

V.G.Lebedev

Lab. of Technical Cybernetics
Novosibirsk State University
Novosibirsk, USSR 630090

S.A.Khamidullin

Dept. of Informatics
Institute of Mathematics
Novosibirsk, USSR 630090

ABSTRACT

Real-time discrete and connected speech recognition system is described. The system includes the following levels:
- pre-processing of speech signal and feature extraction;
- recognition of isolated words;
- recognition and understanding of discrete phrases.

The first level has a hardware implementation and the others have software ones.

A phrase of discrete speech consists of a chain of utterances separated by pauses. Utterances are recognized first, after these and through them phrases are, use being made of semantico-syntactic and pragmatic restrictions.

The word recognition in phrases is carried out by dynamic programming procedure with adaptive corridor. The comparison of the test pattern with all reference patterns is made in parallel. Simultaneously the rejection of misleading candidates over threshold defined in the recognition process is carried out.

Syntactic and semantic districtions are presented in the system as a tree of word connections in phrases. Using this diagram of permissible phrases the system organizes the current word recognition, defines the end of test phrase or goes to the next recognition step and defines in this case the set of permissible phrase continuations.

The system is implemented on microcomputer "Electronica-60M" and also includes pre-processor for the initial speech signal processing and features extracting. As features the output intensities of six bandpass filters sampled every 16 ms are used.

The 120 words occurring in 140 phrases of 3-7 words each are recognized, with speaker adaptation, in the real time scale with a realiability of 98%.

INTRODUCTION

A system is discussed which is designed to understand phrases of discrete speech. Discrete speech phrases Φ_j, $j=1\div N$, are defined as chains of utterances S_i, $i=1\div K$, separated by short pauses. Here N is the total number of possible phrases, and S_i is a single word form or block of connected words (S_i will further be referred to as words). The minimal duration of pauses between utterances is 100 msec. The system operates under two main modes: training and recognition. Under the mode of training a set of templates $E=\{E_i\}$ is formed corresponding to the set $S=\{S_i\}$, $i=1\div K$, and under that of recognition to each phrase F_κ presented for recognition there corresponds a phrase from a set $\Phi=\{\Phi_j\}$. From the whole set Φ of phrases we will consider only admissible phrases, i. e. phrases meeting the semantico - syntactic and pragmatic restrictions adopted in the system. The syntactic restrictions include among admissible only grammatically correct phrases.

The semantic restrictions admit from the syntactically correct phrases only those making sense.

The pragmatic restrictions make it possible to select from semantically correct phrases only those admissible in particular situations specified by a concrete applied field in which this system of phrase recognition is operating. Note that the process of checking a test phrase for admissibility coincides in the system with that of phrase recognition.

From the now available systems of speech recognition the one under discussion differs in that templates and uses a more effective strategy of taking into account semantico-syntactic restrictions [1]. Unlike the systems of recognizing continuous speech the present system has a restriction consisting in requiring obligatory pauses to separate S_i. On the other hand, introducing this restriction allows the labour input to be essentially reduced as it is no longer necessary to divide phrases into words. An important peculiarity of the system is also the possibility of its quick modification through the means of the operation system (RAFOS or RT-11) maintaining the functioning of programs involved in the recognition system. Most of the present-day systems of recognition on the base of microcomputers function without any operation systems or are created as a special processors and hence are deprived of this possibility.

ARCHITECTURE AND ALGORITHM OF SYSTEM OPERATION

The system is realized on a microcomputer "Electronika-60" involving a pre-processor for the initial speech signal processing and a complex of programs written in the languages MACROASSEMBLER and FORTRAN in the "RAFOS" operation system.

Through a microphone a speech signal is fed to the pre-processor which every 16 msec determines the values of intensities at the outputs six filters covering a band of 400 to 5000 Hz.

The value of the total intensity which is to exceed the assigned threshold determines the start of signal input. The final decision on the start of input is taken if several successive input segments meet this condition. The decision to end the input is taken if several successive segments have a total intensity below the threshold. Otherwise the segments of low intensity correspond to speech pauses.

The threshold value and the required number of segments in the first and the second case are assigned by the user.

Performed parallel with the process of input initial intensity vectors worked out by the pre-processor are the operations:
- replacing the assigned number of input intensity vectors by one averaging,
- uniting the averaged close contiguous vectors of intensity into groups (segments) and the subsequent averaging(segmentation) of the latter.

The secondary averaging is performed for a group of vectors in which the distance between the first vector and all the subsequent ones is less than the so-called segmentation threshold also assigned by the user. After the segmentation the secondary features are constructed:

$$R_\ell = 256\,(P_\ell + G_\ell)/(Q+\mathcal{D}),\quad \ell = 1\div 6,$$

where P_ℓ is the value of averaged intensity in the l-th band after segmentation, $Q=\sum_\ell P_\ell$, G_ℓ and $\mathcal{D}$ are reguleting additives.

The software of the system includes the program of input, training and discrete speech prase recognition, and the programs intended for preliminary vocabulary compilation and for forming a tree of word compatibility in phrases.

The program of input, training and recognition performs the following functions
- distribution of on-line storage among the templates E_i and a test realization, which is defined as the next word from a phrase F_κ at the stage of recognition;
- training on the basis of a given vocabulary;
- recognition of discrete speech phrases;
- recording temlates in the file;
- replacing separate templates.

The entire working process is carried on as a dialogue between the user and the computer. To each action required from the user the program supplies prompting requests. This permits a wide range of users to ceadily master the system. In the stage of training the speaker pronounces into the microphone once each word S_i from a given vocabulary. There is a possibility of replacing templates introduced erroneously at the stage of training. The words in phrases are recognized by the method of dynamic programming with an adaptive corridor [2].

Let an utterance S_m consist of ℓ segments. On introduction of the next segment with a number j there occurs a recalculation of all R_i ($i=1,2,\ldots,K$) there R_i are the distances from S_m to templates E_i calculated with the help of an algorithm of dynamic programming. Such an arrangement of calculating R_i allows, in the process of recognition S_m, unpromising templates to be cut off. The cutting-off condition has the form: if $R_i/R_{min} > \beta$ where $R_{min} = \min R_i$ then the template E_i is cut off as unpromising. The cutting-off threshold decreases monotonously with growing j. The cut-off templates take no part in subsequent calculations till the recognition of S_m is over.

Such a behaviour of the cutting-off threshold makes it possible, at the first steps, to exclude from consideration the templates the most different from the test realization by their initial segments. As the number of control realization segments grows account is taken of ever finer distinctions. The recognition comes to an end if at a certain step there remains but one template. Otherwise after the calculation of distances is over a template is chosen with the minimal distance to the test realization. If $R_{min} > R^*$ where R^* is the value of the refusal threshold, the system refuses to recognize the words of S_m.

ALLOWANCE FOR LINGUISTIC RESTRICTIONS

Admissible sequences of words in phrases can be presented in the form of a tree where each branch reflects a continuation of the admissible phrase. Such a presentation is more economical than the conventionally used matrix of word combinability. The tree is built in the form of a two-dimensional array uniting a sequence of units each of which is a set of admissible nodes (word numbers) having a bond with the node of the previous level (the first line of the unit), and reference addresses

to the units of the subsequent level (the second line). The reference addresses equal to zero determine the end of the phrase.

The input data describing the phrases are represented as a sequence of lines each of which describes a particular phrase (a group of phrases) or part of it and has the form: $[j]\,A_1, A_2, \ldots, A_i, \ldots, A_k\,[*]$ where j is the current phrase number, A_i - a number or a set of numbers for words that could stand in the i-th place of a phrase, * is the simbol for the phrase continuation in the next line if its description fails to fit into one line.

Example:

1. (10,12), 11(1,2,3,4,5,6,7,8,9), (13,14)
2. (25) 24(23,21)(13,14)
3. (15)(29,92)

The first line of this file describes 36 phrases in which the first place may be occupied by the 10-th or 12-th word, the second by the 11-th, the third by the 1-st, 2-nd,...,or 9-th; the fourth by the 13-th or 14-th word of the fixed vocabulary. The second line describes 4 phrases in which the first place may be occupied by the 25-th word, the second by the 24-th, the third by the 23-d or 21-st, the fourth by the 13-th or 14-th. The program of forming a tree of word compatibility in phrases operates in the dialogue mode and makes it possible to introduce initial data determining the sequence of words in a phrase from the terminal keyboard or from an earlier prepared external file. Taking into account the large variety of identical branches the program eliminates repeated branches which allows the required volume of memory to be reduced 5 to 6 fold.

The array of phrases constructed according to the above example has the form

1	2	3	4	5	6	7	8	9	10	11	12	13	14
28	10	12	25	15	-2	11	-2	1	2	3	4	5	6
4	6	6	21	26	1	8	99	18	18	18	18	18	18
15	16	17	18	19	20	21	22	23	24	25	26	27	28
7	8	9	-2	13	14	-1	24	-1	23	21	-1	29	92
18	18	18	2	0	0	1	23	2	18	18	2	0	0

The initially constructed array describing the tree of word compatibility in phrases would contain 306 words instead of 56 as is the case after the optimization.

EXPERIMENTAL RESULTS

The system was tested on phrases of a problem-oriented vocabulary belonging to the language of an air-traffic-dispatcher. The vocabulary contained 120 words. On the material of 140 phrases made up of 3 to 7 words with speaker adaptation the recognition reliability obtained amounted to 98%. The branching factor varied from 1 to 48 and on the average was equal to 13. The system worked in the real time scale. At present the system is in experimental operation.

REFERENCES

[1] G.Ya.Vysotsky, B.N.Rudny, V.N.Trunin-Donskoy, G.I.Tsemel, "Experience of Speech Control by a Computer", Izvestiya Akademii Nauk SSSR. Tekhnicheskaya kibernetika, Moscow, pp.134-143, no.2, 1970.

[2] H.Sacoe, S.Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. Acous. Speech and Signal Processing, vol. ASSP-26, no.1, pp.43-49, 1978.

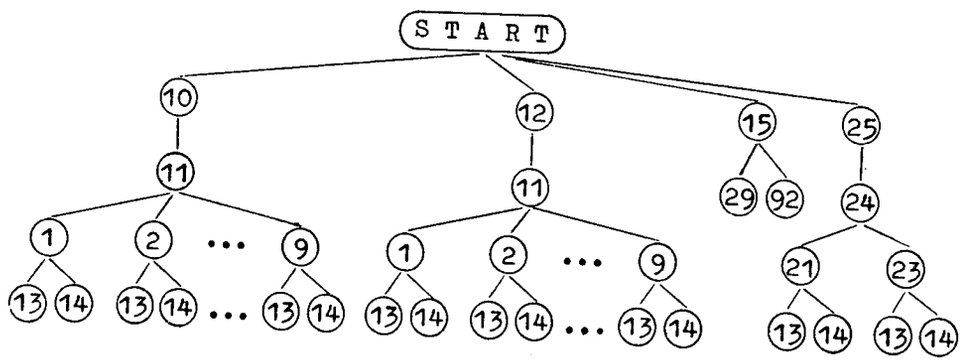


Fig. 1. Example of the tree of permissible phrases.

THE EFFECTS OF WORD BOUNDARY AMBIGUITY IN CONTINUOUS SPEECH RECOGNITION.

JONATHAN HARRINGTON[+] ANNE JOHNSTONE[++]

The Centre for Speech Technology Research, University of Edinburgh, Scotland.
Also Department of Linguistics[+] and Department of Artificial Intelligence[++], University of Edinburgh, Scotland.

ABSTRACT

This study assesses the effect of employing different phonological units on parsing a given string of phonemes into words in a continuous speech recogniser. It is shown that when the input utterance is encoded using a representation intermediate between the broad classes in Huttenlocher & Zue [3] and the 44 phonemes of Received Pronunciation, the number of possible word strings found from the input utterance is usually in excess of 10 million. Even when all 44 phonemes are implemented, an input utterance of 4 - 10 words in length can be parsed into in excess of 10,000 word strings if word boundaries are not identifed prior to lexical access. In the final part of the study, it is shown that the number of such parses can be reduced if stress is represented in the input utterance and lexicon.

INTRODUCTION

In some of the speech understanding systems of the *ARPA* project [1] as well as the feature-based, continuous speech recogniser being developed at Edinburgh University, one of the main tasks of the syntactic and semantic components is to filter out grammatically unacceptable and meaningless word strings which are the output of a lexical access component: they must, then, be able to identify the target word string *please let us know*[1] from a list of other strings such as *please letter snow* and *please let a snow*, both of which are possible word parsings of the input phonemic string /p l ii z l e t @ s n ou/[2]. The total number of word strings which can be derived from a given phonemic string depends on several factors, such as the number of entries in the lexicon, the parsing strategy and the units which are used to represent words phonemically. This paper forms part of a larger study of which the main goal is to devise a set of units which is optimal both from the point of view of acoustic-phonetic processing (i.e. it must ultimately be possible to identify such units with a high degree of accuracy from the acoustic waveform) and from the point of view of syntactic/semantic filtering (i.e. the number of word strings passed to the syntactic/semantic components should be minimal).

An initial aim has been to implement a *mid-class* representation in the continuous speech recogniser [2] being developed at Edinburgh University. The prime motivation for analysing the acoustic waveform into mid-classes is that they should be easier to identify than the 44 phonemes of Received Pronunciation (R.P.): for example, an analysis of the acoustic waveform into mid-classes such as /B/, voiced stop, is (arguably) likely to result in better identification scores than its analysis into the members of /B/, that is /b/, /d/, /g/. At the same time, it has been shown that when all the words of a 20,000 word lexicon are represented in classes that were much 'broader' than our mid-classes (i.e. there are fewer broad-classes than mid-classes and therefore a greater number of phonemes, on average, in a broad-class than a mid-class), around 1/3 of the words are still uniquely identifiable [3]; when the lexicon is represented in mid-classes, the percentage of uniquely identifiable words will presumably be considerably greater. This may, therefore, be a strong argument for analysing the acoustic waveform as far as the mid-class level and allowing the syntactic and semantic components to filter out the impermissible word strings which have resulted from using mid-classes rather than phonemes. However, the statistics on discriminability in the lexicon do not take account of the fact that in continuous speech, word boundaries are more difficult to identify from a given mid-class string compared with a phonemic string. Thus, while at a phonemic level, the sequence /m g l/ can only be parsed into /m # g l/ [4] (e.g. *same glass*), at a mid-class level (i.e. /N B L/), the unambiguous identification of the word-boundary is no longer possible: since the mid-class category /N/ includes /n/ and since /B/ includes /d/, /N B L/ could also be parsed as /N B # L/ (e.g. *sand layer*), or indeed /N B L #/ (e.g. *sandle*). Since phonotactic constraints often no longer successfully apply at the mid-class level, the total number of ways of parsing a given mid-class string into words is likely to increase considerably despite the fact that the lexicon remains highly discriminable when represented in mid-classes. The first experiment was designed to determine the magnitude of this increase and to assess whether this would place an unmanageable burden on syntactic and semantic filtering.

METHOD

The lexicon of the continuous speech recogniser includes the 4000 most frequent words from the American Heritage Dictionary [5]. Each entry consists of an orthographic form, a phonemic citation form (R.P.) and a key for accessing syntactic tag information. Using a phonological rule interpreter written in INTERLISP-D to run on the Xerox 1100 series [6], a set of phonological reduction rules was applied to this lexicon to derive fast speech forms (known as reduced forms) which were stored together with the citation form under the corresponding orthographic entry. Details of the reduction rules are given in [7].

The lexicon containing citation and reduced forms was then compiled into a discrimination tree in which, working from left-to-right, phonemic entries with identical phoneme sequences share the same branch(es). Thus, *tee, tea, teach, teacher* and *tedious* share the same branches as far as the second phoneme /ii/ at which point there is a division to /d/ (the continuation through the tree for *tedious*) and to /ch/ (the continuation for *teach* and *teacher*). A terminal branch is attached to a phoneme node whenever a

sequence of phonemes forms a word. A fragment of the tree is shown in Figure 1.

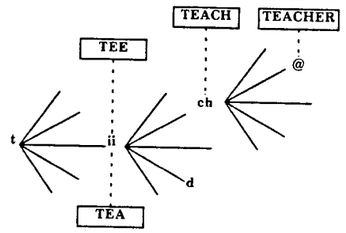


FIGURE 1: The tree-structured phonemic lexicon implemented in the continuous speech recogniser.

The acoustic front end of the continuous speech recogniser analyses the acoustic waveform into a string of phonemes (henceforth *input utterance*) which are matched against the discrimination tree to locate possible word boundaries. The matching process takes place from left to right through the input utterance and when a word is matched, it is stored on a word lattice. Thus, if the first phonemes of a string were /t ii ch i ng w i l/ (*teaching will...*), *tea* and *tee* would be stored on the word lattice. Subsequently, two searches, or paths, are continued: the first is the continuation from /ii/ to /ch/ and /i/; the second is from the initial /ch/ node that begins words such as *chide, choke* etc. to the /i/ node, in this case, of words such as *chin* and *chimpanzee*. The second path in this example would be terminated for two reasons: there are no citation, or reduced forms, beginning with /ch i ng/; and also because the fragment /ch i/ is not a citation or reduced form of any word. Only those paths which enable a complete parsing of the input utterance are passed to the syntactic and semantic components *and only such complete paths are considered in the statistics on total number of paths in this paper*. An example of the parsing process is shown in Figure 2.

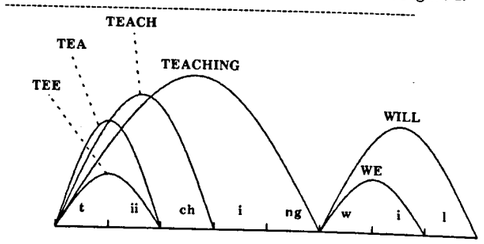


FIGURE 2: the paths show that there is only one possible parsing of the sequence /t i ch i ng w i l/, into *teaching + will*.

When an input utterance is represented in mid-classes, it is first expanded into all possible phonemic representations which are then each matched against the tree as described above.

Phonemic transcriptions were made by a trained phonetician of 50 sentences produced by one R.P. speaker: these sentences included a mixture of 'phonemically dense' sentences (e.g. *I know no minimum* whose consonants are entirely nasal); sentences taken from a 'phonemically balanced' passage; sentences from Section H of the Lancaster-Oslo-Bergen corpus [8]; sentences from a corpus of business dictation collected at C.S.T.R; and sentences that consisted of words which were uniquely identifiable (in isolation) when represented in mid-classes. These transcriptions were then converted automatically to their corresponding mid-class representations[3].

The hand transcriptions, rather than the phonemic strings derived automatically by the acoustic front end of the continuous speech recogniser, were input to the discrimination tree. As such, the hand transcriptions can be considered as a perfect analysis by the continuous speech recogniser of an acoustic waveform into a string of mid-classes excluding any representation for word boundaries, syllable boundaries or stress.

RESULTS I

Parses into words	$< 10^3$	$10^3 - 10^4$	$10^4 - 10^5$	$10^5 - 10^6$	$10^6 - 10^7$	$>10^7$
Number of utterances	1	4	2	4	7	32

TABLE 1: distribution of mid-class input utterances in terms of number of word strings found. The first column on the left denotes, for example, that 1 (out of 50) utterances was parsed into less than 1000 word strings.

From Table 1 it can be seen that 32 out of the 50 input utterances of 4 - 10 words in length were parsed into 10 million, or more, word strings. The smallest number of parses was 82, the largest 3.25×10^{18}. The average number of parses was 8.47×10^{16}.

DISCUSSION I

When *isolated* words are represented in mid-classes, the number of words that are uniquely identifiable decreases from around 98% (for words represented phonemically) to 70%. Such a statistic would suggest that analysing the acoustic waveform to the mid-class level is a viable alternative to performing a phonemic analysis. However, if word boundaries cannot be identified prior to matching the input string against the lexicon, the syntactic and semantic components could have to find the correct word string from over 10 million competing strings. This may be too much of a burden to place on higher level processing; furthermore, the hand-transcriptions represent the best possible mid-class analysis of the acoustic waveform by a connected speech recogniser. The number might increase if the input utterance were derived from an acoustic front end analysis that contained a large number of errors. Second, the lexicon only contained around 4000 lexical items; if larger lexicons are implemented (e.g. 20,000 words), the number of word strings found will undoubtedly increase. Finally, the speaker produced utterances of 4 - 10 words in length. It is certainly possible to produce longer utterances without pausing; in this case, the total number of ways in which the corresponding mid-class representation could be parsed into words would again increase.

The next section reports statistics on the number of possible word parsings of *phonemic* input utterances. In this case, the original hand-transcriptions were matched against the tree-structured lexicon.

RESULTS II.

Parses into words	0 - 10	11 - 100	$100 - 10^3$	$10^3 - 10^4$	$> 10^4$
Number of utterances	15	16	8	8	3

TABLE 2: distribution of phonemic input utterances in terms of number of word strings found. The first column on the left denotes, for example, that 15 (out of 50) input utterances were parsed into 10, or less, word strings.

Table 2 shows that when the input utterances are phonemically represented, over half of them parse into 100, or less, competing word strings. The average number of word strings was 2491; there were two utterances that could only be parsed into one word string; the maximum number of word strings for any utterance was 66,528.

It was not necessarily the case that longer input uttterances (where length is defined as number of phonemes or number of words intended by the speaker) necessarily gave the greatest number of parsings into word strings: there was no correlation between number of phonemes in the utterance and number of parses into words (r = -0.07, not significant); neither was there a significant correlation between number of words produced by the speaker and number of possible parses of its phonemic representation into word strings (r=0.11), although there is a trend to show that these two variables are positively correlated.

DISCUSSION II

Even when the input utterance is phonemically encoded, the syntactic and semantic components could have to filter out over 10,000 competing word strings from the target word string. While this figure may not be unmanageable for syntactic and semantic filtering, it is clearly desirable to seek to reduce this figure further.

As a means towards reducing the number of word strings, we considered the possibility of increasing the number of 'sound units' by using *allophones* in both the input utterance and the lexicon. The fact that the number of word strings should decrease using an allophonic representation is easily demonstrable. Phonemically, *plea* is represented as /p l ii/ which also embeds the lexical item *Lee*, phonemically /l ii/. On the other hand, *Lee* would not be embedded within *plea* in an allophonic representation since these would be encoded as [li] and [pl̥i] respectively. However, this advantage would be lost if the allophones that were the product of word-internal context-effects were also caused by context-effects across word boundaries: thus if /l/ in *Lee* were realised as a voiceless [l̥] in a moderately fast production of *..type Lee some results*, *Lee* would once more be embedded within *plea* even at an allophonic level of representation. There is some experimental evidence [9] to suggest that such word-boundary coarticulation of /l/ is possible. If the majority of identifiable allophones can occur as a result of coarticulation both across word boundaries and word-internally, the case for introducing this kind of phonetic representation is considerably weakened. Furthermore, the acoustic front end is faced with the additional difficulty of identifying *specific* allophones; given the fact that our recogniser does not obtain a perfect analysis even at the mid-class level, it may be unrealistic to assume that it would be able (in the short-term, at least) to differentiate, for example, between aspiration before stressed vowels, the various allophones of /h/ and [l̥].

An alternative means of increasing the number of units in the input utterance, and thereby decreasing the number of word strings found, would be to include *stress* in the lexicon and input utterance. In order to test this hypothesis, stressed vowels were differentiated from unstressed vowels in the lexicon by inserting a '*' symbol before the former; no distinction was made between different levels of stress; thus, *conversation*, which is normally marked for secondary stress on /k o n/ and primary stress on /s ei/ is represented as /k *o n v @ s *ei sh @ n/ in citation form. With this type of representation, in which each vowel phoneme (except schwa) can be marked for stress, an additional 19 units are introduced into the phonemic inventory. The same set of hand-labelled transcriptions were then re-transcribed including the '*' symbol to identify word-stress. Most of the function words were not marked for stress either in the lexicon or in the hand-labelled data. These modified transcriptions were matched against the modified tree-structured lexicon as described above.

RESULTS III

Parses into words	0 - 10	11 - 100	$100 - 10^3$	$10^3 - 10^4$	$> 10^4$
Number of utterances (U)	15	16	8	8	3
Number of utterances (S)	20	14	10	5	1

TABLE 3: distribution of phonemic input utterances in terms of number of word strings found. (U)/(S) denote the input utterances unmarked/marked for stress as described above.

The results in Table 3 show that when the input utterance and the lexicon are marked for stress, 34 utterances are parsed into less than 100 word strings and 6 utterances are parsed into 1000, or more, word strings; the corresponding results for a phonemic input utterance unmarked for stress are 31 and 11 respectively. For the stress-marked input utterances, the average number of word strings for a given utterance was 624 compared with 2481 word strings for the unstressed, input utterances.

Table 4 shows similar statistics for stressed and unstressed *mid-class* utterances. The stressed mid-class utterances were derived from the stressed phonemic utterances by automatic conversion into mid-classes, but with stressed vowels marked: thus, /k *o n v @ s *ei sh @ n/ is represented as /P *BV N V CV S *D S CV N/ where /P/, /BV/, /CV/, /S/ and /D/ are the mid-classes *voiceless stop, back vowel, central vowel, strong fricative* and *diphthong* respectively.

Parses into words	$< 10^3$	$10^3 - 10^4$	$10^4 - 10^5$	$10^5 - 10^6$	$10^6 - 10^7$	$>10^7$
Number of utterances (U)	1	4	2	4	7	32
Number of utterances (S)	3	6	5	8	8	20

TABLE 4: distribution of mid-class input utterances in terms of number of word strings found. (U)/(S) denote the input utterances unmarked/marked for stress as described above.

As in the phonemic utterances, the inclusion of stress in the mid-class utterances improves performance: there are 14 stressed,

mid-class utterances that were parsed into 100,000, or less, word strings and 20 such utterances parsed into 10 million, or more, word strings: the corresponding results for the mid-class, *unstressed* utterances are 7 and 32 respectively. The average number of word strings for mid-class, stressed utterances is nevertheless very high at 5.54×10^{13}.

DISCUSSION III

The results show that including stress in the process of matching an input utterance to the lexicon clearly decreases the average number of word strings found. Implicit in this result is the assumption that the acoustic front end would be able to identify stressed vowels in an utterance. There are some reports [10], [11] of a high level of success in the automatic identification of stressed vowels; work in this area on our own continuous speech recogniser is currently in progress.

This study has also not been able to take account of *sentence* stress which could cause some unstressed vowels in our lexicon to be stressed. Thus, since *can* (auxiliary) is marked as unstressed in the lexicon, our lexical access model would fail to find the appropriate word string in the utterance *I can come* (emphasis on *can*); the effects of sentence stress on the automatic identification of lexically stressed vowels is the subject of a future investigation.

CONCLUSIONS

If the acoustic front end of an automatic speech recognition system is unable to locate word boundaries, the syntactic and semantic components must be implemented to identify the target word string from a potentially large number of competing word strings. The number of ways in which a given input utterance can be parsed into word strings depends on at least two factors: the type of parsing strategy and the units which are used for the phonemic representation of words. This study has been concerned with the latter problem and has shown that an input utterance represented entirely in mid-classes, or even phonemically, may place an unmanageable burden on syntactic and semantic filtering. The possibility of introducing stress into the input utterance and lexicon was explored with preliminary, promising results. The introduction of stress in this way may provide a basis for implementing a mixed mid-class and phonemic representation: if stress enables a substantial reduction in the number of word strings found, it may be possible to represent some of the phonemes which are notoriously difficult to identify from the acoustic waveform (such as weak fricatives /th, dh, h, f, v/) by their mid-classes.

REFERENCES

[1] Smith A. & Sambur M. (1980) Hypothesizing and verifying words for speech recognition. In Lea W. (ed.) *Trends in Speech Recognition* 139-165 Englewood Cliffs: New Jersey.

[2] Dalby J, Laver J. & Hiller S.M. (1986) Mid-class phonetic analysis for a continuous speech recognition system. In Lawrence R. (ed.) *Proceedings of the Institute of Acoustics* 8.7, 347-354. Institute of Acoustics: Edinburgh.

[3] Huttenlocher D.P & Zue V.W. (1983) Phonotactic and lexical constraints in speech recognition. *Proceedings of the American Association for Artificial Intelligence Conference* 172-176.

[4] Lamel L. & Zue V.W. (1984) Properties of consonant sequences within words and across word boundaries. *IEEE Institute of Acoustics, Speech and Signal Processing* 42.3.1 - 42.3.4.

[5] Carroll J., Davies P. & Richman B. (1971) *The American Heritage Word Frequency Book*. Houghton-Mifflin: New York.

[6] Cutting D. & Harrington J.M. (1986) Phongram: an interpreter for phonological rules in automatic speech recognition. In Lawrence R. (ed.) *Proceedings of the Institute of Acoustics* 8.7, 461-470. Institute of Acoustics: Edinburgh.

[7] Harrington J.M., Laver J. & Cutting D. (1986) Word-structure reduction rules in automatic, continuous speech recognition. In Lawrence R. (ed.) *Proceedings of the Institute of Acoustics* 8.7, 451-460. Institute of Acoustics: Edinburgh.

[8] Johannson S., Leech G.N. & Goodluck H. (1978) *The Lancaster-Oslo/Bergen Corpus of British English*. Oslo University: Department of English.

[9] Bladon R.A.W. & Al-Bamerni A. (1976) Coarticulation resistance in English /l/. *Journal of Phonetics*, 4, 137-150.

[10] Marshall C. & Nye P. (1983) Stress and vowel duration effects on syllable recognition. *Journal of the Acoustical Society of America* 74, 433-443.

[11] Lea W. (1980) Prosodic aids to speech recognition. In Lea W. (ed.) *Trends in Speech Recognition* 139-165 Englewood Cliffs: New Jersey.

NOTES

1 Our thanks to Jim Hurford for this example.

2 The machine readable alphabet for the 44 phonemes of R.P. is shown below:

/p/	*pea*	/f/	*fan*	/l/	*lee*
/b/	*bead*	/v/	*van*	/r/	*road*
/t/	*tea*	/th/	*think*	/w/	*win*
/d/	*day*	/dh/	*then*	/y/	*you*
/k/	*key*	/s/	*sing*	/m/	*man*
/g/	*guy*	/z/	*zoo*	/n/	*name*
/ch/	*chew*	/sh/	*shoe*	/ng/	*sing*
/jh/	*judge*	/zh/	*measure*		
		/h/	*hat*		
/ii/	*we*	/o/	*hot*	/ei/	*stay*
/i/	*hit*	/oo/	*saw*	/ai/	*sigh*
/e/	*head*	/u/	*could*	/oi/	*toy*
/a/	*had*	/uu/	*who*	/au/	*now*
/aa/	*hard*	/@/	*the*	/ou/	*go*
/i@/	*here*	/u@/	*sure*	/e@/	*there*
/@@/	*first*				

3 We thank Maggie Cooper for her assistance in converting from phonemes to mid-classes.

4 We are grateful to Julian Kupiec for writing software to count the number of word strings.

This research was supported by SERC grant number GR/D29628 and is part of a collaborative project with Plessey Research & Technology, the Husat Research Group at Loughborough, Imperial College of Science and Technology, and C.S.T.R., Edinburgh University. Our thanks to John Laver and Henry Thompson for many helpful comments on an earlier draft of this paper.

ON THE SPEAKING MODULE OF AN AUTOMATIC READING MACHINE

GABOR OLASZY

Institute of Linguistics
Hungarian Ac. of Sciences
Budapest 1250 Pf. 19
Hungary

GÉZA GORDOS

University of Technology
Budapest XI. Stoczek u.2
Hungary

ABSTRACT

The speaking module (Scriptovox) of the automatic Hungarian Reading Machine(RM) was developed in the years 1983--86 by a four-member research team of electrical engineers of the University of Technology, the Institute of Linguistics and the Research Institute of the Hungarian Post and Telecommunication. Scriptovox --using the MEA 8000 type integrated circuit for speech generation -- was developed for the fully automatic conversion of any Hungarian text into good quality speech in real time.

INTRODUCTION

The primary requirement a text to speech (TTS) converter system has to meet is that it should convert every character of a text in a given language (including not only letters but other characters as well) into control codes with the aid of which intelligible speech can be generated by a speech synthesizer. At the same time an important requirement is that it should recognise the different types of sentences (statements, questions, etc). This recognition is the basis of the automatic generation of melody and rhythm. Last but not least, a fundamental requirement is the real time operation of conversion and speech generation.

The conversion of ASCII characters of the text into synthesizer control codes is realised in the Scriptovox system in 3 steps.

1. Conversion of ASCII characters into "phoneme codes".
2. Conversion of phoneme codes into MEA control codes (speech frames) and their concatenation.
3. Realisation of melody patterns by rewriting the pitch control bits of some speech frames of the group of frames concatenated in step 2.

LETTER TO PHONEME CODE CONVERSION

Thirty-three phonemes are used for generating Hungarian speech. Only the short versions of speech sounds are included among these thirty-three phonemes, the long versions are represented by doubling the phoneme code of the short counterpart. When processing the graphemes of the text into phoneme codes we distinguish three types of ASCII characters.

The first -- and simplest-- type comprises those characters with which a phoneme code can be associated directly in one step, e.g. A,O,V,F,H etc.

The second type of characters cannot be converted directly into code numbers: their conversion requires an examination of the neighbouring characters. Examples for such characters are S,Z,C,T, etc. For instance, the letter S occurs in the combinations SZ, ZS, SSZ, ZZS, CS, CCS, denoting different sounds in each case.

The third group of ASCII characters includes numbers,abbreviations, and other symbols.

This algorithm works as follows :

1. Setting of initial values .
2. Accepting ASCII codes into buffer-1 (B1) which is 1 kbyte large .
3. Identification of ASCII characters by scanning the text left to right.
4. If it is a special type of a letter then going to the rules for setting the appropriate phoneme code. Writing the phoneme code into buffer-2 (B2) which is 1 kbyte large.
5. If it is a number then going to the number routine where phoneme codes of the number will be inserted into B2.
6. If some other symbol then going to the lexicon and set the appropriate codes into B2.
7. Setting all other characters into B2.
8. Identifications of ASCII characters representing punctuation marks and setting the appropriate codes in B2.

Finally in B2 we find the original text in a form as if everything in it had been written using only letters. For example, the sentence MOST 12 ÓRA VAN. 'It's 12 o'clock now' takes the form MÖST TIZENKETTÖÖ OORA VAN.

PHONEME CODE - MEA CONTROL CODE CONVERSION

In the next step the program converts the content of B2 into a series of speech frames which are stored in a 4 kbyte buffer (B3). For the conversion, a collection of speech frames (225 different types) and a 33x33x6 element concatenation matrix (rule system) is used. The initial content of the speech frames of the data base and the rules were defined in 1983 and have continuously been refined thereafter. The rule system includes rules for the concatenation of frames picked from the data base when converting the phoneme codes of B2 into a series of frames. Choosing the appropriate elements of the 225-element data base every Hungarian sound and sound combination (VV,CV,VC,CC), as well as all the assimilations can be realised.

The rule system

In order to get speech from the group of phoneme codes stored in B1 these codes must be converted into very many speech frames to be stored in the buffer B3. This buffer size (4 kbyte) is enough for 40 s of speech in on conversion process. The rule system works as follows. In the rule matrix every row and column represents a phoneme code. The program -- before turning to the rule matrix-- makes a diad-like interpretation of the phoneme codes in B2 scanning it from left to right. So by the interpretation of consecutive phoneme codes a row and a column of the rule matrix are determined. This row and column pair points at a matrix entry, the contents of which are six bytes. These bytes represent the identifiers of speech frames that must be picked from the data base and placed into B3 one after the other. If the desired sound effect of a step during the conversion requires less then six speech frames ∅'s are inserted in the superfluous bytes. If during execution the program finds ∅'s it goes on to the next step. It should be noted that one complete sound combination is realised by the program totally after performing three steps: the step before the sound combination concerned in B2, the step of its own, and the next one. When the step by step conversion of phoneme codes is completed, B3 contains a series of speech frames of the text to be uttered. Sending these frames to the synthesizer a monotonous, robot-like speech will be produced. Thus the realisation of TTS conversion has been acomplished at the segmental level only.

Automatic generation of melody

To make speech more natural melody patterns must be superimpesed on the segmental realisation. In Scriptovox system a fully automatic melody generation of the mostly used types of patterns is working. What are the elements of this melody generation?

1. Building microintonation patterns into appropriate sound combinations.
2. Recognising the articles and some conjunctions in the text and making them unstressed.
3. Recognising comma(s) in the text and changing the intonation (and rhythm) before the comma(s).
4. Superimposing the intonation of declarative sentences characterised by a full stop at the end.
5. Superimposing the appropriate melody patterns on the various types of questions (question mark at the end). The types of questions distinguished for Hungarian are as follows:

a) Questions beginning with Q-word.
b) Questions without Q-word -- further divided into three subcases(see below).

Microintonation

Quick variations in fundamental frequency independent of context and of the speakers' will are called microintonation. The variation ranges between 10--15 Hz inside a sound. It is built in the speech frames.

Articles and conjunctions

The system identifies the articles a,az 'the' and conjunctions és 'and', hogy'that'. To make these words less stressed the pitch is decreased by 8 Hz in them. In the article being at the very beginning of a sentence the pitch is decreased 16 Hz.

The interpretation of comma(s)

A comma in a written text corresponds to a change in the melody and rhythm of live speech. To implement these changes the proper place of comma(s) in B3 has to be marked. For this purpose a special frame with short duration and zero amplitude is inserted wherever there is a comma in the text. Then by scanning right to left the earliest 32ms long frame of the vowel immediately preceding the comma is searched for and its pitch is increased by 8 Hz. The search goes on to the left up to the next vowel and in its first 32ms long frame the pitch is increased by 4 Hz. In the frame following the comma the pitch is then restored to the former value.

The intonation of statements

The automatic generation of this type of pattern begins with the calculation of the length of the sentence. Three types of modifications are used in the algorithm, i.e. reducing the pich by 4 Hz, or by 8 Hz or reducing it in the last word of the sentence by an additional 4 Hz(see Table 1.) .

Table 1.

Sentence length categories	Minimum number of frames	Number of parts	Reduction in every part	Reduction in the last p add.	Reduction in the last word add.
very short	10	3	8Hz	-	-
short	20	4	4Hz	4Hz	-
normal	30	5	4Hz	4Hz	-
long	45	6	4Hz	4Hz	4Hz

The intonation of questions

The punctuation of questions is characterised by the use of question mark. In Hungarian several types of questions are used. To make an automatic realisation of questions, a multi-level algorithm has to be designed. Regularities lending themselves for algorithmic procedure, as well as unambiguous orthographic forms have been found for the following types of Hungarian questions.

Question beginning with Q word. The system recognises twenty-one Q words and implements the intonation pattern of Fig. 1a.

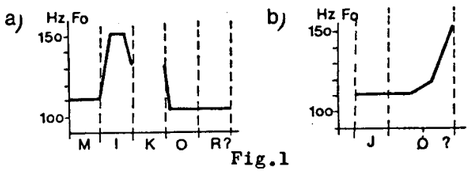


Fig.1

In the next step the algorithm of declarative sentences is applied in the rest of these questions.

Questions having no Q word are as follows: One syllable question, where the intonation peak must be at the very end of the vowel (Fig.1b).

Two syllable questions have the peak also in the last vowel but the pattern differs from the earlier one. The peak must be placed at the beginning of the vowel and the peak must be reduced towards the end of the same vowel. The duration of the peak must not exceed 30 ms. By this type of questions three types of word endings are distinguished (Fig.2). The intonation patterns differ radically in the way the pitch decreases at the end of the wovel.

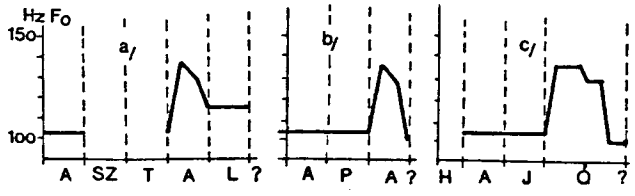


Fig. 2

Three or more syllable questions have the intonation peak in the last but one vowel followed by a decrease of pitch in the last one. The quality of the resulting pattern is further improved by distinguishing two subcases (Fig.3a,b). If the last but one vowel is long then the decrease of pitch has to commence already in the long vowel.

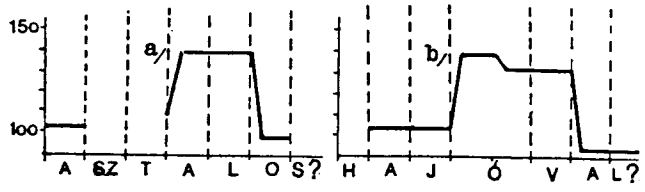


Fig. 3

THE PERCEPTUAL EXAMINATION OF SPEECH QUALITY

The complete process of designing and constructing a TTS system has to end in a scientifically based perceptual examination of the speech quality. The phonetically balanced speech material for the test consisted of four groups of sound sequences: 30 syllables, 30 meaningless bisyllabic sequences, 30 one or polysyllabic words and 10 sentences. This material was recorded by a male announcer and by the Scriptovox. The natural speech material was given for two groups of 18 students (18 year old) and one week later they listened to the synthesized material. They had to put down what they thought they heard. The score for words was 84% for both the natural and synthesized items, as to the sentences, 99% for the natural and 98% for the synthesized was obtained.

CONCLUSIONS

The Scriptovox TTS system displays several differences when compared with conventional unlimited vocabulary speech synthesis systems. Its data base comprises only 225 speech frames (1 kbyte). The rule system converting letters and other symbols of the text into a concatenation of speech frames uses, at one point, a novel diad-like representation. Extensive experimentation was involved in formulating the fully automatic generation of rules of intonation for the various classes and subclasses of sentences. The Scriptovox system seems to accomplish a good compromise among low cost, high speech quality, fully automatic TTS (no need for diacritics or auxiliary symbols in the text), small memory requirement (12 kbyte) and very low bitrate (approx. 100 bytes/second).

REFERENCES

[1] MEA 8000 voice synthesizer. Philips Technical Publ. 101. 1983, Netherland

[2] Gordos G.--Takács Gy.:Digitális beszédfeldolgozás. Müszaki K. 1983, Budapest

[3] Olaszy G.:A magyar beszéd leggyakoribb hangsorépitő elemeinek szerkezete és szintézise. NyÉrt. 121. 1985, Budapest

[4] Olaszy G.:A phonetically based data and rule system for the real time text to speech synthesis of Hungarian. X-th Int.Cong.of Phon.Sciences, Utrecht 1983, Abstracts 398.

USE OF SPEECH SYNTHESIS IN AN INFORMATION SYSTEM FOR HANDICAPPED TRAVELLERS

B.C. DICKSON, S.J. EADY, J.A.W. CLAYARDS, S.C. URBANCZYK AND A.G. WYNRIB

Centre for Speech Technology Research, University of Victoria,
P.O. Box 1700, Victoria, B.C., V8W 2Y2

ABSTRACT

Communicaid is an interactive microcomputer-based information retrieval system that employs synthesized speech and visual displays to assist handicapped travellers at an international airport. Audio output is accomplished using an LPC-based synthesis system in which the units of synthesis are individual words and phrases of English or French. Concatenated synthesis units are modified by a set of phonetic liaison rules and by a pitch smoothing algorithm.

INTRODUCTION

This paper describes a microcomputer-based system, called Communicaid, that has been designed to provide information to handicapped travellers at the Vancouver International Airport. Communicaid uses a combination of visual displays and synthetic speech output to provide information in English and French on topics such as transportation, accommodation and airport facilities. The video presentation is designed for travellers with hearing impairments, whereas the audio presentation is intended for those with sight impairments. Both modes of presentation are automated on a microcomputer. Since video mode of presentation does not involve audio speech output, it will not be discussed further here. The remainder of this paper deals exclusively with the audio portion of the Communicaid system.

DESIGN CONSIDERATIONS

Information Format

Communicaid presents audio information to the traveller in the form of information menus [1], similar to the one shown in Figure 1. Each menu provides a list of topics from which to choose. The user listens to the menu and selects an item by pressing a button after hearing a topic of interest. This process is repeated several times with increasingly specific menus, until the user obtains the desired information. At that point, the user can return to the main information menu and choose another topic of interest.

The information menus have been designed to take into account the temporary nature of the speech signal (compared to written text) and the way this interacts with the limitations of human information processing [1]. Consequently, each menu contains a maximum of seven information items, so as not to exceed the limits of human short-term memory [2]. In addition, the user has the option of listening to each menu several times. During the repetition, menu items are presented in a more cryptic form to introduce some variability into the audio presentation. Finally, for each menu item, the audio presentation is augmented by the visual display of a key word in large letters on a video monitor. All of these features are designed to facilitate easy information access in the audio mode.

Voice Output

In the initial stages of this project, we considered several different types of automated voice output for the presentation of audio information. Due to the interactive nature of the audio presentation, we needed a method that would provide fast access to the audio messages in an order that would be determined by the user. Thus, analog recordings of the information menus on audio tape would not be appropriate. On the other hand, commercially available text-to-speech conversion systems were judged unsuitable for this project, because of relatively low speech quality [3].

We then considered the use of digitally recorded speech materials, a method that has been demonstrated in applications such as telephone directory assistance [4], weather forecasts [5] and time-of-day announcements [6]. The advantage of using digitally recorded speech is that the quality is quite high, and the audio material is also easy to access in an interactive application of this kind. The major drawback with digitized speech, however, is that it requires enormous amounts of storage space. It is possible to reduce storage requirements by recording the speech materials in the form of isolated words and phrases, each of which can then be used in a number of different contexts. However, the problem with this strategy is that the prosodic aspects of each digitized word (i.e., pitch, duration and intensity) are fixed and cannot be easily modified to produce an intonation pattern that is appropriate for different sentence contexts. Speech output systems that make use of digitally recorded words typically require several versions of each word with a different intonation contour for each version [4].

MAIN INFORMATION MENU

YOU ARE NOW / AT THE START. (600) / TO CHOOSE INFORMATION ABOUT / ONE /
OF THE FOLLOWING / SEVEN / TOPICS (100)/ PRESS THE SELECT BUTTON /
AFTER YOU HEAR / THE TOPIC / YOU WANT. (600)/ TO END THIS SESSION /
AT ANY TIME (100)/ JUST RETURN / THE HEADSET / TO THE HOOK. (600)/

ONE (100)/ FOR FRENCH (100)/ POUR CONTINUER / EN FRANCAIS (100) / APPUYEZ /
SUR LE BOUTON DE SELECTION. (600) /
TWO (100) / LOCATIONS OF / AIRPORT FACILITIES. (600)/
THREE (100) / GROUND TRANSPORTATION / TO AND FROM / THE AIRPORT. (600)/
FOUR (100) / LOCATIONS OF / AIRLINE / TICKET COUNTERS. (600) /
FIVE (100) / HOTELS / IN THE VANCOUVER AREA. (600) /
SIX (100) / ASSOCIATIONS OF / THE BLIND AND PARAPLEGIC. (600) /
SEVEN (100) / FOR INFORMATION ON / FLIGHT DEPARTURES AND ARRIVALS (100)/
PLEASE CONTACT / THE TRANSPORT CANADA / INFORMATION
BOOTH (100) / DIRECTLY BEHIND YOU. (600) /

THIS IS THE END OF / THE LIST. (600) /
TO REPEAT / THE LIST / YOU HAVE JUST HEARD (100) / PLEASE WAIT. (600)/

FIGURE 1: Contents of the main audio information menu for the Communicaid Centre. The diagonal slashes demarcate the preprocessed vocabulary items that are concatenated to produce the speech output for this menu. Numbers in parentheses indicate the location and duration (in msec) of pauses in the audio output.

Due to space limitations imposed by the present application, this method of voice output was not suitable. Communicaid's information menus required a vocabulary of some 1,100 different words and phrases, each of which would be used in an average of seven different contexts. The digitization of several different versions of each vocabulary item would require an inordinate amount of storage space.

As an alternative to the use of unmodifiable digitally recorded speech items, we chose to produce the voice output for this system by encoding prerecorded words and phrases using linear predictive coding (LPC). This method maintains a relatively high quality of speech output. At the same time, it allows modifications to the prosodic aspects of encoded words and phrases, so that they can be joined together to form complete sentences. In addition, this technique greatly reduces the storage requirements for the encoded speech materials. Thus, our strategy for voice output was to use LPC-encoded words and phrases to generate sentences in a variation of the word-concatenation method proposed by Olive [7,8].

As pointed out by Olive, the production of good-quality synthetic speech using the word-concatenation technique requires that certain modifications be made to the word units when they are concatenated. The parameters that require modification include the pitch contour of each word, the spectral shape and amplitude at word boundaries, and the duration of each syllable.

Due to time constraints for this project, we were not able to develop complex prosody rules for this application. Instead, we chose to address this issue by making the basic units of synthesis (called "vocabulary items") as large as possible.

Thus, each vocabulary item could contain as many as four or five words (see Figure 1 for examples). By using large synthesis units, we were able to reduce the need for modifying prosodic features, because each vocabulary item would already contain pitch and duration patterns that would be appropriate for one or more sentence contexts. The rules that we did develop for modifying prosody acted mainly to alter pitch and energy parameters at the boundaries between vocabulary items.

SPEECH SYNTHESIS METHOD

Synthesized speech for this system is generated on a microcomputer using a Texas Instruments TMS-5220C speech synthesis chip. A control program is used to provide the synthesis chip with a series of quantized values for pitch, energy and ten LPC reflection coefficients. These parameter values are stored as preprocessed vocabulary items corresponding to individual words or short phrases. English and French sentences are synthesized by concatenating these preprocessed vocabulary items in a specified order and then applying rules to modify pitch patterns and to eliminate spectral discontinuities at word boundaries.

Vocabulary Production

As indicated above, the strategy for speech synthesis was to employ preprocessed vocabulary items that contained as many words as possible. Considerable care was taken to parse the audio menu scripts so as to maximize the length of vocabulary items, while also ensuring that each item would be used in several different contexts, within the various menus. The chosen vocabulary items were then embedded in carrier sentences, which replicated as closely as possible the sentential environments in which they would be found in the information menus. This strategy ensured an appropriate intonation pattern for each item.

The carrier sentences were then read by a male speaker whose voice was recorded on a reel-to-reel tape recorder. Each item was digitized (at a 10-kHz sampling rate with 10-bit resolution) and excised from its sentence environment. The digitized vocabulary items were then analyzed using the autocorrelation method of LPC [9] to derive values of energy, pitch and 10 LPC reflection coefficients at 20-msec intervals. These parameters were quantized for output on the synthesis chip. Each encoded vocabulary item was then edited to eliminate any spectral discontinuities, and to provide a uniform energy maximum. This method of digital encoding produces a compression ratio of approximately 80 to 1, compared to the original sampled speech data.

Concatenation of Vocabulary Items

At the time of synthesis, encoded vocabulary items are concatenated to form complete sentences of English or French. The input to the system is a set of command files corresponding to the audio information menus. Each command file contains the vocabulary items that are to be concatenated, along with diacritics to indicate the ends of sentences, and the location and duration of pauses. The system verifies the existence of each word in the list of encoded vocabulary items, and the requested items are joined together in the order specified. At this point, the encoded vocabulary items are modified by a number of phonetic liaison rules and by a pitch smoothing algorithm.

Phonetic Liaison

The phonetic liaison rules act to change energy and segment durations at the boundaries between vocabulary items. The application of each rule is determined by the particular phonetic segments that are present at each boundary. The phonetic segments at the beginning and the end of each item are identified by a rule and classified into one of several categories. There are five phonetic categories for item-initial segments and four such categories for item-final segments. These are listed in Table 1.

A total of 13 phonetic liaison rules handle all possible combinations of initial and final phonetic segments. Eight of these rules involve a smoothing of the energy contour at the boundary between vocabulary items (e.g., for a vowel-vowel or fricative-fricative combination). The other five rules involve the replacement of aspiration with silence (e.g., for a combination of two voiceless plosives), the insertion of a short silent interval (e.g., for a fricative-voiced plosive combination), or the repetition of a voiced frame immediately preceding the boundary (e.g., for a vowel-voiced plosive combination).

TABLE 1

Classification of Segments
for Phonetic Liaison Rules

Item-Initial Segments	Item-Final Segments
1. Vowels and [j].	1. Vowels, liquids, nasals and glides.
2. Nasals and liquids.	2. Fricatives and affricates.
3. Fricatives.	3. Voiced plosives.
4. Voiced plosives and [w].	4. Voiceless plosives.
5. Voiceless plosives, affricates and [ð].	

Pitch Smoothing

Following application of the phonetic liaison rules, a pitch smoothing algorithm modifies the pitch contours at the boundaries between vocabulary items. The aim of this algorithm is to eliminate abrupt pitch changes that may occur at such boundaries. Modifications of this type are required when a vocabulary item recorded for use in

sentence-final position occurs in a sentence-medial location. Sentence-final items are characterized by a falling terminal pitch contour. This falling contour must be flattened somewhat if the vocabulary item is to be used in the middle of a sentence.

In general, this relatively simple pitch-smoothing strategy was found to be adequate for concatenating the large synthesis units used in the present application. This was due to the fact that the information delivery system did not require the use of an interrogative intonation pattern. The intonational requirements for the synthesized speech included a declarative (falling) pitch contour for major syntactic boundaries in non-final position [10]. The former was handled by the falling pitch contour that accompanied most vocabulary items; the latter was generated as a result of the pitch smoothing algorithm and by the insertion of a pause at the major syntactic boundary.

The success of the simple pitch-smoothing rule used in this application is due to the restricted intonational requirements and the use of relatively large synthesis units. A more elaborate pitch assignment algorithm has since been developed for English to handle the concatenation of smaller synthesis units (i.e., individual words), as well as interrogative pitch patterns and the synthesis of focused words in English sentences [11,12]. It is anticipated that this more complex algorithm will be incorporated into the Communicaid system at a later date.

SUMMARY

The Communicaid system uses speech synthesis to provide audio information to handicapped travellers at an international airport. The traveller specifies the desired information by choosing from a list of items presented in an audio menu. Voice output is generated by concatenating LPC-encoded words and phrases of English or French. The concatenated vocabulary items are modified by a set of phonetic liaison rules and by a pitch smoothing algorithm. This application illustrates the use of speech technology for automated information delivery.

ACKNOWLEDGEMENT

This project was done under contract to Rutenberg Design Inc. of Montreal, with funding from the Transportation Development Centre of Transport Canada. We thank Mr. Uwe Rutenberg and Dr. Ruth Heron for technical advice and assistance.

REFERENCES

[1] Waterworth, J.A. (1982). "Man-machine speech dialogue acts." Applied Ergonomics, vol. 13, pp. 203-207.

[2] Miller, G.A. (1956). "The magical number seven, plus or minus two: Some limits to our capacity for processing information," Psychological Review, vol. 63, pp. 81-97.

[3] Carlson, R. and Granstrom. B. (1984). "Text-to-speech conversion in telecommunications," Behaviour and Information Technology, vol. 3, pp. 73-78.

[4] Waterworth, J.A. (1983). "Effect of intonation form and pause durations of automatic telephone number announcements on subjective preference and memory performance," Applied Ergonomics, vol. 14, pp. 39-42.

[5] Andersen, D.P. (1984). "A talking computer gives weather forecasts by telephone," First International Conference on Speech Technology (J.N. Holmes, ed., Brighton, U.K.).

[6] Waterworth, J.A. (1984). "Interaction with machines by voice: A telecommunications perspective," Behaviour and Information Technology, vol. 3, pp. 163-177.

[7] Olive, J.P. (1974). "Speech synthesis by rule," Proceedings of the Speech Communication Seminar, Stockholm, vol. 2, pp. 255-260.

[8] Olive, J.P. and Nakatani, L.H. (1974). "Rule-synthesis of speech by word concatenation: A first step," J. Acoust. Soc. America, vol. 55, pp. 660-666.

[9] Markel, J.D. and Gray, A.H. (1976). Linear Prediction of Speech (New York).

[10] Eady, S.J. (1986). "The influence of syntactic structure on fundamental frequency patterns of Canadian French sentences," Proceedings of the 12th International Congress on Acoustics, paper A6-7.

[11] Eady, S.J., Dickson, B.C., Urbanczyk, S.C., Clayards, J.A.W., and Wynrib, A.G. (1987). "Pitch assignment rules for speech synthesis by word concatenation," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, in press.

[12] Eady, S.J. and Dickson, B.C. (1987) "Synthesis of sentence focus in English declaratives," Proceedings of the 11th International Congress of Phonetic Sciences, in press.

MODELLIERUNG VON INTONATIONSKONTUREN DES DEUTSCHEN - ANWENDUNGEN FÜR SPRACHKOMMUNIKATIONSGERÄTE

DIETER MEHNERT

Sektion Rehabilitationspädagogik und
Kommunikationswissenschaft, Bereich Technik,
Humboldt-Universität zu Berlin, DDR

ZUSAMMENFASSUNG

Aus einer umfangreichen Untersuchung repräsentativen Sprachmaterials konnten für die Wahrnehmung relevante segmentale und suprasegmentale Grundfrequenzkonturen ermittelt werden, die die Grundlage für schematisierte Grundfrequenzmodelle bilden. Es wird an Beispielen gezeigt, welchen Einfluß die Grundfrequenzkonturen auf die Wahrnehmung künstlich erzeugter Sprachsignale ausüben und inwieweit Grundfrequenzmodelle geeignet sind, Sprachkommunikationsgeräte zu qualifizieren.

EINLEITUNG

Stellt man die Kommunikationsfähigkeiten des Menschen denen moderner Rechner gegenüber, so wird das Problem der Kommunikationsgeschwindigkeit im Dialog mit dem Automaten deutlich. Die Kommunikationsgeschwindigkeit ist im Vernältnis zu Verarbeitungszeiten der Information sehr langsam. Sie kann nur erhöht werden durch die Anpassung des Rechners an die Kommunikationsfähigkeiten des Menschen. Möglich ist das über Technologien, die eine Kommunikation mit dem Rechner über die natürliche Sprache realisieren. Die mündliche Sprache ist dabei von besonderer Bedeutung, weil sie das direkteste Kommunikationsmittel darstellt und ohne Verwendung von Zwischendatenträgern verwendet werden kann.
Die Erfahrungen haben gezeigt, daß durch die Sprachkommunikation mit dem Automaten herkömmliche Ein- und Ausgabegeräte keineswegs ersetzt werden können, es hat sich aber herausgestellt, daß sich neue Anwendungsmöglichkeiten anbieten, die aus den Besonderheiten der Sprache resultieren.
Der enorme technologische Fortschritt auf den Gebieten der Mikroelektronik hat die erforderlichen Voraussetzungen für eine Sprachkommunikation mit Automaten geschaffen. Dies gilt gleichermaßen für die Spracherkennung und Sprachsynthese. Da für die Verwirklichung des Dialogs beide Komponenten benötigt werden, steigt mit den Anwendungsmöglichkeiten der Spracherkennung auch der Bedarf an leistungsfähigen Synthesesystemen.
Effektiv wird die Mensch-Maschine-Kommunikation nur dann sein, wenn die Spracherkennung sicher arbeitet und die Sprachqualität bei der Sprachausgabe einem Standard entspricht, der weitgehend die Natürlichkeit des vom humanen Sprachgenerator erzeugten Sprachsignals erreicht und damit eine hohe Verständlichkeit hat, wenn sich beide Kommunikationspartner also problemlos verstehen.
Intonationsstrukturen spielen bei der Realisierung dieser Forderung, wie überhaupt im formalen linguistischen Kommunikationscode, eine wesentliche Rolle.

WIRKSAMKEIT PROSODISCHER PARAMETER

Intonation wird als ein Komplex verstanden, in dem die 3 prosodischen Parameter Tonhöhe, Lautstärke und Dauer kompliziert zusammenwirken. Es erhebt sich die Frage, welche Wirksamkeit die prosodischen Parameter einzeln oder kombiniert auf das synthetische Sprachsignal ausüben, wie groß der Einfluß auf die Verständlichkeit und Natürlichkeit ist und ob es eine Hierarchie unter den Parametern gibt. Daraus sollte abgeleitet werden, welchem oder welcher Gruppe von Parametern man sich bei der Sprachsynthese besonders widmen muß. Dazu wurden in einer größeren Untersuchung /1/ Sätze synthetisch, zunächst monoton, aufgebaut und mit den prosodischen Merkmalen Grundfrequenz, Intensität und Zeit entsprechend Abb. 1 ergänzt.
Hörergruppen hatten die Aufgabe, das synthetische Sprachsignal hinsichtlich Verständlichkeit, Akzentwahrnehmung und Natürlichkeit zu beurteilen. Die Zusammenfassung aller auditiv ermittelten Ergebnisse der Perzeptionsversuche ist in Abb. 2 dargestellt.
Es ist zu erkennen, daß, sobald der Grundfrequenzparameter allein oder mitvertreten ist, die Werte für Verständlichkeit, Natürlichkeit und Akzentwahrnehmung erheblich zunehmen. Der Parameter Grundfrequenz ist in der Lage, beim Hörer mehr akustisches Referenzwissen zu aktivieren als es die Parameter Intensität oder Zeit oder beide zusammen vermögen. Der Hörer wird bei Anwesenheit des Parameters Grundfrequenz deutlich zu größerer Perzeptionsleistung veranlaßt, das bedeutet für

die Anwendung in der Sprachsynthese, daß es sicher genügt, sich vorerst dem Parameter Grundfrequenz und seinen Variationen zuzuwenden. Der Gewinn bei zusätzlicher Berücksichtigung der anderen beiden Parameter wird, im Verhältnis zum Aufwand gesehen, gering sein. Letztlich ist, zumindest in den europäischen Sprachen, die Funktion der Intonation hinsichtlich ihrer kommunikativen Aufgabe im wesentlichen an Grundfrequenzbewegungen gebunden. Der Parameter Grundfrequenz und dessen zeitlicher Verlauf ist dominierend für die Akzentwahrnehmung, die Natürlichkeit und somit auch für die Verständlichkeit.

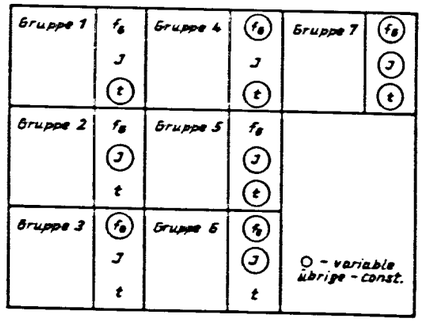


Abb. 1 Perzeptionstest - Strategie

INTONATIONSMODELLE IM DEUTSCHEN

Betrachtet man bisherige Untersuchungen zur Intonation, so kann man feststellen, daß neben einigen Ausnahmen fast alle sprachwissenschaftlich orientierten Methoden das Ziel verfolgt haben, die Tonhöhenphänomene in der Sprache hinsichtlich ihrer distinktiven Funktion zu beschreiben. Danach war zu erwarten, daß die Analyse der deutschen Intonation von einer anderen Methode profitieren könnte, die z.B. alle Verbindungen zur sprachwissenschaftlichen Funktion vermeidet und Fragen zu beantworten versucht, wie:
- Ist es überhaupt notwendig, die Intonation als eine kontinuierliche Grundfrequenzbewegung zu beschreiben oder könnte eine Approximation durch diskrete Grundfrequenzänderungen als Funktion der Zeit erfolgen?
- Welches ist die kleinste Einheit für eine Wahrnehmungsbeschreibung und wieviele Einheiten gibt es?
- Welches sind ihre akustisch-phonetischen Eigenschaften und wie groß ist der akustische Toleranzbereich?

Alle diese Fragen befassen sich mit den melodischen Aspekten der Äußerungen und münden in eine Darstellung der wahrnehmungsrelevanten Tonhöheneigenschaften der Sprache, die in enger Beziehung zu den physiologischen Tätigkeiten des Sprechers stehen. Das heißt, daß eine Beschreibung der Tonhöhenbewegungen zu den sprecherischen Aktivitäten führt und zu Modellen, die in der Kommunikation zwischen Sprecher und Hörer repräsentiert werden.

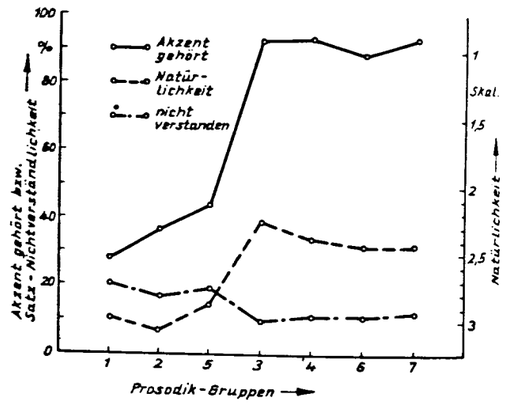


Abb. 2 Wirksamkeit prosodischer Parameter

Verfolgt man auditiv den Intonationsverlauf einer Äußerung, so bemerkt man, daß sich Tonhöhenanstiege und -abfälle und Abschnitte einer fast gleichbleibenden Tonhöhe abwechseln. Betrachtet man Mehrkanalregistrierungen derartiger Äußerungen, findet man den Gehörseindruck bestätigt. Demnach ist also für eine mögliche Approximation der Grundfrequenz von Interesse, wieviele unterschiedliche Anstiege, Abfälle und quasistationäre Abschnitte sich unterscheiden lassen und ob diese segmentalen Tonhöhenbewegungen eine gemeinsame Bezugslinie aufweisen.
Das analytische Zuhören ebenso wie die apparative Analyse eines repräsentativen Sprachmaterials ergaben, daß in den meisten Äußerungen, wo keine bemerkenswerten Anstiege oder Abfälle vorkommen, die Tonhöhe nicht monoton bleibt, sondern allmählich nach unten abweicht. Dieses Phänomen wird als Deklination bezeichnet /2/. Die Auswertung des gesamten Analysematerials ergab für die Deklination eine Abhängigkeit nach Abb. 3.
Danach wurden für die Sprachsynthese 3 Deklinationstypen, abhängig von der Länge der Äußerung, vorgeschlagen (Abb.4). Diese gefundenen Werte für deutsche Satzrealisierungen sind mit denen in der Literatur veröffentlichten Daten anderer europäischer Sprachen vergleichbar/1/.

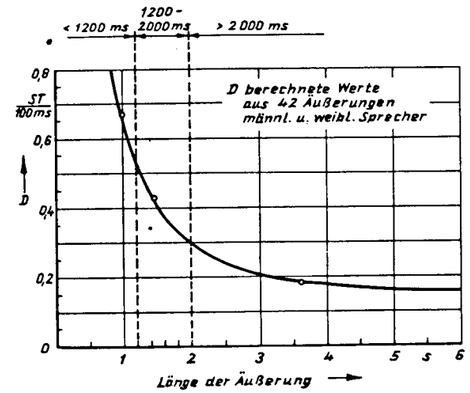


Abb. 3 Deklination als Funktion der Länge der Äußerung

Die nach unten abweichende Linie kann nun als Grundlinie (O) angesehen werden, der die übrigen Tonhöhenbewegungen überlagert sind.

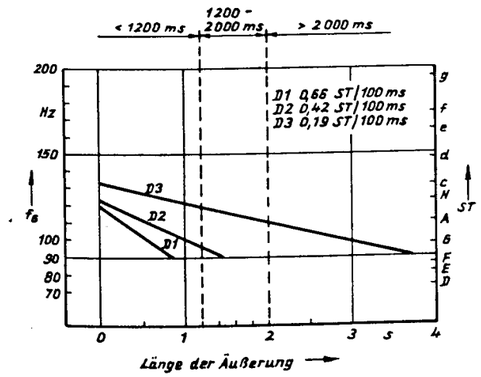


Abb. 4 Deklinationstypen in Relation zum f_G-Bereich

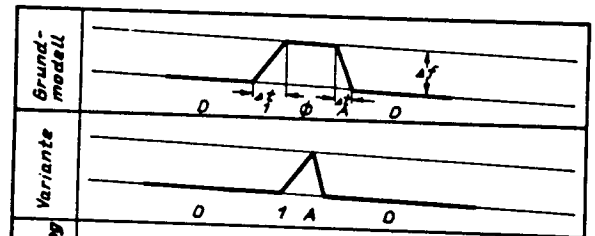
Die Deklinationslinie verbindet ebenfalls Anstiege und nachfolgende Abfälle, d.h. sie kann auch auf einem höheren Niveau (Ø) weitergeführt werden. Für den Abstand beider Linien, den Frequenzhub Δf, sind entsprechende Werte gefunden und in Abb. 5 gemittelt dargestellt.
Die Deklination hat für den synthetischen Aufbau der Grundfrequenzkontur eine große Bedeutung. Wird die Deklination durch eine monotone Grundlinie ersetzt, wird der Eindruck der Natürlichkeit sogleich erheblich herabgesetzt.
Zu hören und im Registriermaterial zu beobachten sind desweiteren Tonhöhenanstiege und -abfälle. Sie werden als Übergänge von einer niedrigen zu einer höheren bzw. umgekehrt, von einer hohen zu einer niedrigen Tonhöhe wahrgenommen. Sie repräsentieren sich in verschiedenen Formen als steile und flache Übergänge.

Tonhöhenbewegung		zur Akzentstelle beitragend	Position in der Silbe	$\Delta f_{gem.}$ [ms]	$\Delta f_{gem.}$ [ST]
Anstiege	1	ja	Anfang	100 - 150	4 - 5
	2	ja	Ende	50	2 - 3
	3	nein	Ende	100	4 - 5
	4	ja oder nein	geht über variierende Anzahl von Silben	offen	4 - 5
Abfälle	A	ja	Ende	75 - 100	4 - 5
	B	ja	Ende	50	2 - 3
	C	nein	Anfang bis Mitte	75 - 100	4 - 5
	D	ja oder nein	geht über variierende Anzahl von Silben	offen	4 - 5

Abb. 5 Daten für wahrnehmungsrelevante segmentale Tonhöhenbewegungen

In dem gesamten Material konnten nun verschiedene Anstiegsformen und Formen des Abfalls gefunden und hinsichtlich Position und Funktion in der Silbe klassifiziert werden.
Schließlich wurden 4 Anstiegsformen (1, 2, 3, 4) und 4 Formen des Abfalls der Grundfrequenz (A, B, C, D) unterschieden. Aus den experimentell gesammelten Daten sind so für die Wahrnehmung relevante segmentale Tonhöhenkonturen (Minimaleinheiten) herausgefunden worden, die die Grundlage für schematisierte größere Intonationseinheiten, für suprasegmentale Grundfrequenzkonturen bilden.
Die Feststellung eines Grundmodells und einiger Varianten führt nun zu der Frage, ob ein derartiges Modell zur Beschreibung der Intonation überhaupt ausreicht. Vom Standpunkt der Kommunikation gäbe es keinen Grund, wonach es weitere Modelle geben sollte, denn das Grundmodell ist in der Lage, eine Äußerung durch die Intonationskontur zu komplettieren, es beinhaltet Tonhöhenbewegungen, die Silben zu betonten Silben machen können und es kann syntaktische Strukturen innerhalb einer Äußerung verbinden usw..
Es gibt jedoch Hinweise, daß weitere Modelle existieren müssen, die sich vom Grundmodell unterscheiden, letztlich auch aus der Feststellung, daß die All-

tagssprache nicht so lebendig klingen würde, wenn sie nur aus der stereotypen Verkettung von Grundmodellen bestände.
Alle diese Modelle sind synthetisch mittels eines speziellen f_G-Konturengenerators aufgebaut und danach perzeptiv überprüft worden.



Abh. 6 Modelle suprasegmentaler f_G-Konturen

Die Aussagen der Hörer waren erwartungsgemäß eindeutig, die Intonationskontur 'Deklination + Akzent' wurde höher bewertet als nur 'Deklination' und beide zusammengenommen wesentlich höher als 'monoton' /1, S. 134/.
Die Auswertungen der Registrierungen und die Hörergebnisse lassen die Schlußfolgerung zu, daß sich die Intonation des Deutschen durch eine Reihe von Grundfrequenzkonfigurationen beschreiben läßt.

ANWENDUNGEN FÜR SPRACHKOMMUNIKATIONSGERÄTE

Wie nachgewiesen werden konnte, leisten die beschriebenen Grundfrequenzmodelle einen Beitrag zur Qualitätsverbesserung synthetischer Sprache.
Realisiert wurde eine zunächst halbautomatische Intonationssteuerung des Sprachsynthesegerätes ROSY 4201 (TU Dresden), d.h., daß durch gesetzte Zusatzzeichen in der Phonemkette an den zu akzentuierenden Stellen der Rechner die gewünschte suprasegmentale f_G-Kontur selbständig aufbaut. (Später werden diese Zusatzzeichen bereits bei der Graphem-Phonem-Umsetzung mitbehandelt.)
Der Rechner ermittelt aus den weiteren Satzzeichen die Länge des Satzes, bestimmt die entsprechende Deklinationslinie und berechnet, unter Berücksichtigung der Position der Akzentstellen, eine der Hochlautung angenäherte suprasegmentale Grundfrequenzkontur, die dann über den Synthetisator realisiert wird.
Das Programm erlaubt die terminale Intonationskontur mit einer oder mehreren Akzentstellen und auch die progrediente Intonationskontur, wenn die Satzteile durch 'Komma' oder 'und' voneinander getrennt (gekennzeichnet) sind. Der zweite Satzteil wird dann hinsichtlich der Akzentstellen wie ein autonomer Satz behandelt.

Die gleichen Grundfrequenzmodelle lassen sich auch zur Tonhöhensteuerung eines Elektrolarynx verwenden, was zu einer erheblichen Verbesserung der Verständlichkeit der Elektrolarynx-Sprache im Vergleich zur monotonen beiträgt. Bisher war mit derartigen Anregungsgeneratoren für Laryngektomierte nur monotone Sprache möglich. (Frühere Versuche mit steuerbaren Geräten brachten nicht den gewünschten Erfolg /1, S. 158/).
Aus den Erfahrungen, die mit der Grundfrequenzsteuerung bei der Sprachsynthese gemacht worden sind, wurde ein Steuerteil für den Elektrolarynx entwickelt, der eine Reihe von 'normierten' f_G-Bewegungen allein auf Abruf erzeugen kann /3/.
Mit diesen Elementen ist die Erzeugung des Grundmodells und seiner Varianten möglich. Das bedeutet zwar eine Einschränkung, sie scheint aber aus dem Grund erlaubt, da sich etwa 60 - 70 % aller Äußerungen im Deutschen mit dem Grundmodell und seinen Varianten realisieren lassen. Ein Grundmodell kann auch ohne wesentliche Natürlichkeitseinbuße eine kompliziertere f_G-Kontur ersetzen.
Zusammenfassend kann festgestellt werden, daß bei einer Gegenüberstellung der intonationsgesteuerten zur monotonen Elektrolaryfx-Sprache die Natürlichkeit der erstgenannten höher eingeschätzt wird. Davon profitiert indirekt auch die Verständlichkeit, das Hören von richtig intonierter Elektrolarynx-Sprache ist für den Kommunikationspartner angenehmer, er kann sich besser auf das erzeugte Sprachsignal konzentrieren.

/1/ D. Mehnert, Analyse und Synthese suprasegmentaler Intonationsstrukturen des Deutschen, ein Beitrag zur Optimierung technischer Sprachkommunikationssysteme
Diss. B, Technische Universität Dresden (1985)
/2/ A. Cohen, H't Hart, On the anatomy of intonation, Lingua 19 (1967), 177-192
/3/ D. Mehnert, Anwendung suprasegmentaler Intonationskonturen zur Verbesserung von Elektrolarynx-Sprache, Studientexte zur Sprachkommunikation 2 TU Dresden (1986), 110 - 118

Doz. Dr.sc.techn. D. Mehnert, Humboldt-Universität zu Berlin, DDR 1086 Berlin, Unter den Linden 9 - 11

A FULL HUNGARIAN TEXT-TO-SPEECH MICROCOMPUTER FOR THE BLIND

G. Kiss(1), A. Arató(2), J. Lukács(3), J. Sulyán(2), T. Vaspöri(2)

1. Hungarian Academy of Sciences Institute of Linguistics
2. Hungarian Academy of Sciences Central Reas.Inst.for Physics
3. Eötvös Loránd Univ. Faculty of Natural Sciences

ABSTRACT

The authors introduce Braille-Lab, a Hungarian-speaking microcomputer developed for the blind. This Z80 microprocessor-based personal computer is fitted with a Philips MEA 8000 formant synthesizer, providing for Hungarian text-to-speech conversion. The original version of the machine contains a speaking BASIC interpreter. The new version, Braille-Lab+, is also furnished with a speaking word processor and a speaking database management system running under a speaking CP/M compatible operating system. Braille-Lab has been approved and adopted by the Hungarian National Federation of the Blind, 95 sets have been installed so far.

INTRODUCTION

In the past few decades, intensive research into speech synthesis has been going on in a number of countries including Hungary. This research work has three main types of motivation.

1. Fifth-generation computers are to create a new, humanized type of man-machine-man relationship. Hence one of the main objectives of research is viva voce 'conversation' between man and machine. The various links of the man-machine-man communication chain (each constituting a research area in its own right) and the way artificial speech production fits into that chain are represented in Fig. 1.

2. Another impulse for attempts at speech synthesis was the desire to achieve a better understanding of the acoustics of speech. Indeed the principle of analysis by synthesis is more effective than any measuring apparatus, however sophisticated the latter may be: it shows what the essential components of speech really are [7]. That principle can be best implemented by formant synthesis. These considerations led to the establishment, under Kálmán Bolla's leadership,of a complex acoustic

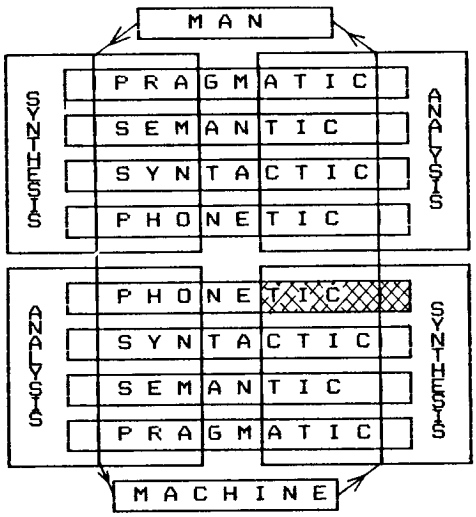


Fig. 1. The Man–Machine–Man communication chain

speech synthesizing system in the Linguistics Institute of the Hungarian Academy of Sciences, in the late 1970s. The hardware configuration includes an OVE III (Swedish-made) formant synthesizer [1] and a PDP 11/34 computer. The effective operation of the system is guaranteed by a specially designed interactive program called FOPRO [11]. The utility of the system for phonetic research is demonstrated by a number of scholarly papers [5, 10]. The program was also used for designing an inventory of speech frames for a Hungarian text-to-speech (TTS) system based on the principle of formant synthesis in the early 1980s [13]. The inventory, in turn, was used in HUNGAROVOX, a Hungarian real-time TTS system for speech synthesis [9, 12]. Later, a developing system was also made for a Philips MEA 8000 formant synthesizer [3].

3. The third type of motivation for research on speech synthesis is a desire to develop various appliances to help handicapped people (afflicted with speech disorders, blindness, etc.). The area was

given a vast impetus by the appearance, in the early 1980s, of speech synthesizer contained in a single IC, e.g. UAA 1003, IMS 5200, SC-01, MEA 8000 [2, 4, 6, 8], since these could be built into various appliances. These considerations led to the development of Braille-Lab (B-L), a speaking computer to be used by blind people, introduced in the present paper. This Hungarian-speaking microcomputer fitted with a text-to-speech conversion system effectively helps the education of blind people in computational technology (thus creating high-qualification employment possibilities for them). Also, it accelerates their full integration into society.

THE HARDWARE OF BRAILLE-LAB

B-L is a Hungarian-made, Z80 microprocessor-based personal computer. Its memory is organized on a page basis,and consists of 64 kbyte RAM and 20 kbyte ROM. The card containing the speaking module has been built into the computer with MEA 8000. The TTS software is located on page 2 of ROM. The keyboard of B-L contains every letter of the Hungarian alphabet, arranged in a way almost identical with the keyboard of standard Hungarian typewriters. The built-in small loudspeaker makes it possible for the speech produced by the system to be heard without an external loudspeaker. The built-in BASIC interpreter leaves 48 kbyte free memory capacity available for the user.

The basic version of B-L has been further developed. B-L+, the new version, runs under a CP/M compatible operating sysem. Along with a 64 kbyte operative memory, it is also furnished with a 192 kbyte RAM disk and a 1 Mbyte floppy disk drive. The new version further contains a speaking word processor and a speaking database management system. With these two programs,its possibilities of application by the blind have been multiplied.

THE TEXT-TO-SPEECH SOFTWARE SYSTEM OF BRAILLE-LAB

The basis for Hungarian TTS conversion by B-L is a text in Hungarian orthography, with no special symbols added. The program translates that text into a series of frame code numbers for the MEA 8000 synthesizer. The frame code numbers designate the elements of a 218-member frame inventory, devised earlier. The TTS conversion is implemented in the following four steps:

1. First of all, the text to be converted to speech is transformed by the program into a series of (code numbers of) speech sounds. Hungarian orthography is a fairly accurate indicator of the series of sounds to be uttered. However, not only single letters but also combinations of two, and even three, letters may stand for single sounds. In the letter-to-sound transformation, the program basically relies on Fig. 2.:

1	2	3	4	5	1	2	3	4	5
1.	a	1	ɔ	-	34.	nn	22	n:	+
2.	á	2	a:	-	35.	ny	23	ɲ	-
3.	b	10	b	-	36.	nny	23	ɲ:	+
4.	bb	10	b:	+	37.	o	6	o	-
5.	c	11	ʦ	-	38.	ó	6	o:	+
6.	cc	11	ʦ:	+	39.	ö	7	ø	-
7.	cs	12	ʧ	-	40.	ő	7	ø:	+
8.	ccs	12	ʧ:	+	41.	p	24	p	-
9.	d	13	d	-	42.	pp	24	p:	+
10.	dd	13	d:	+	43.	r	25	r	-
11.	e	3	ɛ	-	44.	rr	25	r:	+
12.	é	4	e:	-	45.	s	26	ʃ	-
13.	f	14	f	-	46.	ss	26	ʃ:	+
14.	ff	14	f:	+	47.	sz	27	s	-
15.	g	15	g	-	48.	ssz	27	s:	+
16.	gg	15	g	+	49.	t	28	t	-
17.	gy	16	ɟ	-	50.	tt	28	t:	+
18.	ggy	16	ɟ:	+	51.	ty	29	c	-
19.	h	17	h	-	52.	tty	29	c:	+
20.	hh	17	h:	+	53.	u	8	u	-
21.	i	5	i	-	54.	ú	8	u:	+
22.	í	5	i:	+	55.	ü	9	y	-
23.	j	18	j	-	56.	ű	9	y:	+
24.	jj	18	j:	+	57.	v	30	v	-
25.	k	19	k	-	58.	vv	30	v:	+
26.	kk	19	k:	+	59.	z	31	z	-
27.	l	20	l	-	60.	zz	31	z:	+
28.	ll	20	l:	+	61.	zs	32	ʒ	-
29.	ly	18	j	-	62.	zzs	32	ʒ:	+
30.	lly	18	j:	+	63.	sp	33	-	-
31.	m	21	m	-					
32.	mm	21	m:	+					
33.	n	22	n	-					

1= number, 2= letter /s/, 3= code number
4= IPA symbol , 5= length of sound

Fig. 2. Table of letter-to-sound correspondences

2. The second step of TTS conversion is the designation of the series of frames that will realize the speech sounds of the text to be uttered. This designation is basically of a diadic nature.The 218 frames utilized are arranged in the inventory in a very special order. Each combination of sounds is realized by adjacent frames. Thus we can dispense with storing what is called a combination matrix and consequently save a significant amount of memory capacity. In order to further optimize the utilization of the frame inventory, various sound sequences can be realized by overlapping series of frames,as illustrated in Fig. 3. Long sounds are also produced at this stage by multiplying some component of the frame of the corresponding short sound (2 to 5 times, as the case may be) in the series of frame code numbers. Each element of the series of frame code numbers will be an integer between 1 and 218. That series then serves as input to the melody generating part of the program.

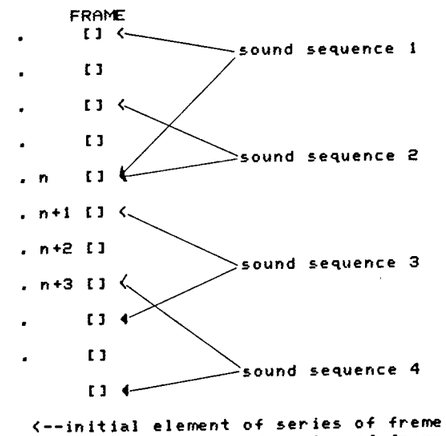


Fig. 3. The structure of the frame inventory and the way frames realizing sound sequences are specified

3. The melody is generated by the program by selecting the appropriate value of the PI parameter of the MEA 8000 synthesizer frame-by-frame. The first step in producing the melody is the segmentation of the text into intonation units. The intonation units are marked off by .(full stop) ,(comma) ?(question mark) !(exclamation mark)or RETURN. Triggered by those punctuation marks, the program will supply the segmental structure produced so far with one of the melody patterns.

4.When the coding of segmental and suprasegmental structure is completed, B-L forwards the resulting series of code numbers to the MEA 8000 speech synthesizer, the speech is simultaneously heard.

THE USE OF BRAILLE-LAB

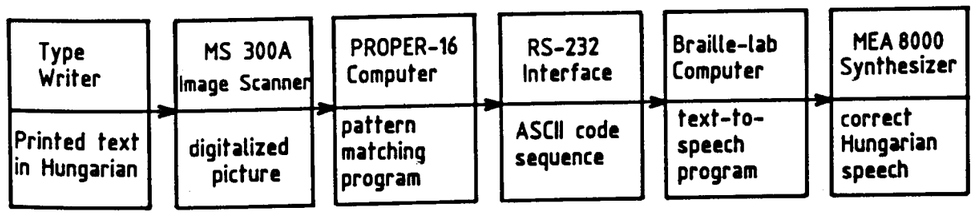
The computer is able to speak as soon as it is switched on. The following introductory words appear on the screen and are simultaneously heard [in Hungarian]: "B-L computer, version 3.1. 48843 bytes of free memory capacity."

After that, each time a key is pressed, the system utters the corresponding speech sound, in order to make it easier for a blind person to avoid typing errors. Names of non-letter keys, including numerals, are uttered as words. E.g. on pressing % the machine says "százalék" (percent), etc. Using the cursor keys, the user can aurally check the contents of any character position of the screen.

Basically there are two situations in which B-L actually speaks: 1. during entering and editing BASIC programs; 2.at run-time when any information appearing on the sreen is simultaneously said aloud.

1. During program editing, the echoing function mentioned above is in operation; in addition, at the end of each line when RETURN is pressed the computer reads out the whole line as connected text. Numerals at this point are not read character-by-character but as wholes (e.g. twenty-five rather than two, five). The English terms of BASIC are read out according to the Hungarian value of

The hardware of reading machine



The software of reading machine

Fig. 4 The system of the reading machine when the Braille-lab is speaking terminal

letters, rather than in proper English pronunciation. At program listing, the list can be heard as it appears on the screen. In short, any information appearing on the screen including e.g. error messages, is also uttered without any special command.

2. The information appearing on the screen during the running of BASIC program will also be heard automatically. For instance, as a result of the running of the following short program, all Hungarian numerals between 1110 and 1125 will be heard first with a question intonation and then with a statement intonation (i.e. "Is the next number 1110? Yes, 1110." etc):

```
10 FOR I=1110 TO 1125
20 PRINT "A következő szám" I "?"
30 PRINT "Igen" I "."
40 NEXT
```

BRAILLE-LAB AS THE SPEAKING PERIPHERY OF A READING MACHINE

At an exhibition called "Hungarians in the World" held in Budapest in August 1986, the authors, assisted by researchers of SZKI (Institute for Computer Research) connected B-L whith an IBM compatible PROPER-16 computer. On the other side, an MS 300A Image Scanner was also connected with PROPER-16. The image recognition program developed by the SZKI people recognized printed Hungarian text. PROPER-16 forwarded the resulting ASCII code, via a standard RS 232 interface, to B-L which uttered the text real time, with a proper Hungarian intonation, intelligibly.

HOW TO MAKE BRAILLE-LAB SING

One of the special features of B-L is that it can also sing. To make the computer sing, the user has to specify the correct rhythm and the correct sequence of pitches. Rhythm can be represented by lengthening the vowels appearing in the words of the song, by entering vowel letters more than once. The length of the syllable containing the vowel will increase in proportion with the number of identical vowel letters entered. The melody has to be given in relative sol-fa letters, according to Zoltán Kodály's method. The pitch defined by sol-fa letter assigned to a syllable will be superimposed by the program on the appropriate syllable which has been rhythmically defined as above. By that procedure, any Hungarian-text song can be produced. This special feature of the system opens up a novel area of application in the on-line representation and correction of Braille music notation [15].

BRAILLE-LAB AS AN AUTHORIZED APPLIANCE

B-L is an appliance authorized for use by the Hungarian National Federation of the Blind. By March 1987, a total of 95 sets have been installed in the schools of the Federation and by individual users. Based on the speaking BASIC of B-L, the Federation organized two beginners' courses on computation in spring 1986 and 1987. The speaking computer effectively helped the blind participants to acquire knowledge and skill in computation and to put them to creative use. The Users' Manual for B-L has been published on cassette tape and in Braille print as well.

REFERENCES

[1] AB FONEMA: OVE III c Speech Synthesizer Manual. Type 21001.

[2] ANDERS, B.: Digitale Sprachsynthese für Low-Cost Anwendungen. Bauelemente der Elektronik 7. 1981, 246-250.

[3] ARATÓ A.--KISS G.--TAJTHY T.: A MEA 8000 beszédszintetizátor Commodore 64 számítógépen működő fejlesztő rendszere. /The developing system of MEA 8000 speech synthesizer operating on Commodore 64 computer/ Magyar Fonetikai Füzetek (MFF) 15. 1986, 143-154.

[4] ASTHEIMER, A. Sprachsyntese in LPC-Technik, Elektronik 12. 1981, 73-81.

[5] BOLLA K.: Folyamatos beszédszintetizáló rendszer magyar nyelven (VOXON). MFF 10, 118-129.

[6] BRÜCK v. H. E.--TEULING, D. J. A.: Integrated voice synthesizer. Philips, Technical publication 48. Electronic Components and Applications Vol. 4 No. 2, February 1982.

[7] FONÁGY I.: Utószó. In.: Laziczius Gy.: Fonetika. Budapest, 1963, 189-206.

[8] FONS, K.--GARGAGLIANO, T.: Articulate Automata: An Overview of Voice Synthesis, BYTE Publications Inc. 1981.

[9] KISS G.: Parolsintezo kun nelimigata vortaro en la spegulo de la hungara lingvo. In.: Perkomputila Tekstoprilaboro, redaktis Koutny I., Budapest, 1985,33-47.

[10] KISS G. A magyar magánhangzók első két formánsának meghatározása szintetizált hangmintákat felhasználó percepciós kísérlet segítségével. /Determining the First Two Formants of Hungarian Vowels by a Perception Test Involving Synthesized Sound Samples/ Nyelvtudományi Közlemények 87/1. 1985, 159-170.

[11] KISS G.--OLASZY G. Interaktív beszédszintetizáló rendszer számítógéppel és OVE III szintetizátorral. /An Interactive Speech Synthesizing System with Computer and OVE III. Synthesizer/ MFF 10. 1982, 21-46.

[12] KISS G.--OLASZY G.: A HUNGAROVOX magyar nyelvű, valós idejű, párbeszédes beszédszintetizáló rendszer. Információ Elektronika 2. 1984, 98-111.

[13] OLASZY G.: A magyar beszéd leggyakoribb hangsorépítő elemeinek szerkezete és szintézise. Nyelvtudományi értekezések 121. Budapest 1985.

[14] OLASZY G.--PODOLECZ GY.: A SCRIPTOVOX MEA 8000 beszédelőállító rendszer felépítése és hangelemtára. Kép és Hangtechnika 6. 1986, 49-61.

[15] Revised International Manual of Braille Music Notation 1956. World Council for the Welfare of the Blind, Paris.

E CADUC: FACTEURS DISTRIBUTIONNELS ET PROSODIQUES DANS DEUX TYPES DE DISCOURS

Pierre R. Léon

Phonetics Laboratory, Department of French, University of Toronto
300 Huron Street, Toronto, Ontario, M5S 2X6 CANADA

ABSTRACT/RÉSUMÉ

L'étude du E caduc, dans deux types de discours, la conférence et le débat montre que, à côté de facteurs distributionnels phonématiques et stylistiques, un certain nombre de paramètres prosodiques doivent être pris en compte pour expliquer le jeu de la variation: Accentuation barytonique, jonctures, un nouvel E caduc final fonctionne comme signal identificateur, émotif et sociolectal.

INTRODUCTION

Les principales études que l'on possède sur le jeu du E caduc en français, en particulier celles d'André Martinet (6), Henriette Walter (7 et 8), Vincent Lucci (5), ainsi qu'une synthèse récente (1), examinent le problème phonologique et phonétique, essentiellement sous l'angle distributionnel, en fonction de critères externes sociologiques ou stylistiques. La présente étude, reprend un certain nombre des points examinés par Lucci (5) pour l'incidence du style didactique sur la réalisation du E caduc. On tente de montrer ici que plusieurs facteurs prosodiques jouent un rôle important, quelles que soient les variables en cause.

CORPUS

Le corpus est constitué par 10 échantillons de parole de 5 à 6 minutes, chacun, extrait de conférences, prononcées à Toronto entre 1974 et 1986 par 5 universitaires hommes (BN, FC, PL, PM, MV) et 5 universitaires femmes (AMH, JK, DLB, ML, NM). On a ajouté 10 extraits de même durée, pour les mêmes sujets engagés dans une discussion avec leurs collègues, après leur conférence. L'âge des sujets varie entre quarante et soixante ans. Cette variable, ainsi que l'origine géographique, n'a pas été prise en compte, car tous les locuteurs retenus ont un français similaire, très standardisé (du type généralement pris comme modèle par les orthoépistes). Il faut noter que leurs idiosyncrasies relèvent d'un même type — celui des intellectuels français. Leur discours est caractérisé par une élocution assurée, marquée par une articulation ferme, nette, et un grand nombre d'accents didactiques, même lorsque l'énoncé est coupé de ruptures, calculées, ou non.

TRANSCRIPTION ET DÉPOUILLEMENT

Le corpus a été transcrit en notant, à l'audition, tous les E caducs réalisés ou supprimés, dans toutes les distributions possibles. On a noté les frontières de groupes accentués ainsi que les pauses. On a compté à part les E caducs d'hésitation. Il a parfois été difficile de décider, à la finale, si une détente consonantique longue pouvait être comptée comme un E caduc. On a essayé de ne retenir dans ce cas que les E perçus avec une valeur syllabique. (Une étude sur ce problème est en cours).

RÉSULTATS GÉNÉRAUX

On peut avoir une idée générale des deux types de discours auxquels on a affaire ici en comparant les chiffres de la table 1. Le nombre d'E caducs prononcés, pour l'ensemble de chaque texte, est indiqué en pourcentage par rapport à toutes les possibilités de réalisations des phones. On a donné, à titre indicatif, le nombre d'hésitations réalisées sous forme de E caducs plus ou moins longs. Ce chiffre est en valeur absolue. Le débit a été calculé sur 3 échantillons d'une minute, effectués au début, au milieu et à la fin de chaque texte (60 échantillons au total).

Type =	Conférence			Débat		
paramètres / Sujets	Débit	ə %	Nbr [ə] d'H	Débit	% ə	[ə] d'H.
BN	4.0	49.3	7	3.90	46.1	23
FC	3.94	50	3	3.91	42.8	14
PL	3.87	39.9	12	4.09	58.6	25
PM	3.88	48.2	14	4.69	20.7	32
MV	3.81	63.6	1	3.48	57.7	8
Moyenne	3.9	50.2	7.4	4.01	45.18	20.4
écart type	.07	8.5	6	0.40	15.35	9.4
AMH	4.29	59.7	6	4.49	45.7	5
JK	4.33	29	5	3.61	35.2	0
DLB	4.58	61.7	4	4.77	40.25	16
ML	4.18	51.4	0	4.10	46.28	21
NM	3.84	30	0	3.81	43.8	22
Moyenne	4.24	46.3	3	4.15	42.2	12.8
écart type	.24	15.8	2.8	.47	4.5	8.7

Table 1. Total des E caducs prononcés en toutes positions, débit en seconde pour chaque sujet et nombre d'hésitations à l'intérieur des groupes.

On remarque tout d'abord une grande homogénéité du point de vue du débit moyen, pour les 4 groupes, et le faible écart qui existe entre le débit de la conférence et celui du débat. On pourrait en déduire qu'il y a en fait un débit du discours intellectuel qui varie peu, quelle que soit la fonction discursive exercée en public. Par contre, le nombre d'hésitations (réalisées comme des E caducs) semble bien une caractéristique individuelle qui varie énormément. (Les deux sujets féminins (ML et MN) qui n'ont réalisé aucune hésitation ont lu leur texte. Le sujet féminin (JK) était dans une colère froide très contrôlée lorsqu'elle est intervenue dans le débat.) Quant à la moyenne des E caducs prononcés, on pourrait imaginer qu'elle est inversement proportionnelle à la vitesse d'élocution. Si cela semble vrai grosso modo pour la comparaison de la conférence au débat chez les hommes, cela ne se vérifie plus chez les femmes et les variations individuelles empêchent d'envisager une conclusion sur ce type de rapport.

E CADUC INITIAL DE GROUPE

Il existe, dans notre corpus, un E caduc en initiale absolue, du type "[ə] mon propos". Cet E fonctionne comme voyelle d'appui, brève et inaccentuée, qui ne donne pas l'impression d'un E d'hésitation, à cause de sa netteté et de son enchaînement avec le début du groupe. Son emploi varie selon les individus et ne semble avoir ni rôle phonématique ni stylistique.

Discours / Sujets	Conférence (%)	Discussion (%)
BN	94.1	95
FC	100	96.4
PL	95.6	100
PM	97.2	81.81
MV	95.2	94.1
Moyenne	96.4	93.4
écart type	2.2	6.8
AMH	56	58
JK	75	95.4
DLB	89.2	46.4
ML	83	81.8
NM	100	66
Moyenne	80.6	69.5
écart type	16.5	19.3

Table 2. E caduc prononcé à l'initiale de groupe

Par contre, on a effectué un examen de tout E caduc, premier phone vocalique d'un groupe linguistique, tel que "Je prétends"... qui peut être réalisé comme "J'prétends". Il s'agit, selon les phonéticiens, d'un phone qui apparaît dans cette distribution en variation libre. Son occurence dépendrait du style. L'E caduc réalisé serait la marque sémiotique d'une recherche. Théoriquement, on devrait en prononcer beaucoup plus dans une conférence que dans un débat. On a relevé, pour notre corpus, les chiffres de la table 2.

La distribution phonématique montre qu'une très haute proportion d'E caducs sont maintenus à l'initiale.

Le type de discours semble avoir peu d'incidence pour le groupe des hommes. Il en a un peu plus pour les femmes qui présentent en outre de bien plus grandes variations individuelles dans les deux types de discours.

Un facteur prosodique entre probablement en ligne de compte dans le maintien du E caduc à l'initiale de groupe. C'est la tendance du français moderne à accentuer le début de groupe. Ivan Fònagy (2) à rappelé que cette tendance à la barytonie a été signalée il y a fort longtemps. Dans quelques échantillons du corpus analysé, 92% des E caducs maintenus comportent un accent de durée et d'intensité. Il s'agit, répétons-le, de textes où le sujet énonciateur est fortement impliqué. Ex. "Je n'admets pas", "Je considère", "cela ne peut être nié", "ne nous leurrons pas", "le dialogue est avant tout"...

On trouve peu de séquence d'E caducs successifs dans notre corpus en dehors du groupe Je ne. Dans 100% des cas le premier E est prononcé. Le second n'est supprimé que dans 42% des cas, lorsque le discours n'est pas argumentatif.

E CADUC INTÉRIEUR DE GROUPE

Précédé d'une seule consonne prononcée. Dans ce cas, la règle distributionnelle veut que l'E caduc tombe généralement, comme dans: "là dessus", "comme pour", "lentement". Les chiffres relevés dans notre corpus sont loin de confirmer cette prédiction, comme on le voit dans la table 3.

La distribution phonématique n'entraîne la suppression attendue du E caduc que dans un peu plus de 50% des cas chez les hommes et 60% chez les femmes.

Le facteur stylistique, entre les 2 types de discours, n'a pas eu d'incidence lui non plus. L'un des sujets masculins (FC) a même inversé la tendance attendue. Les écarts types indiquent, ici encore, des variations individuelles. Cependant, dans l'ensemble, le maintien du E caduc dans des exemples comme "là-dessus", "dans ce climat", "on le voit", est sans aucun doute un marqueur du discours intellectuel.

Un facteur prosodique a cependant joué dans cette distribution pour le maintien du E caduc, lorsque celui-ci est suivi d'un mot commençant par un accent didactique. On relève ainsi: "étant donné le peu de temps", "ou de la situation", "il y a aussi le contenu". Dans notre corpus, ces cas représentent un assez faible pourcentage des E caducs maintenus après une seule consonne (de 2 à 9% avec de grandes variations individuelles). Mais le maintien de cet E est systématique chaque fois qu'il est suivi d'une syllabe marquée d'un accent didactique. L'E caduc semble alors être une sorte d'appui phonique, suivi généralement d'une jonction expressive, le plus souvent avec coup de glotte devant la consonne suivante.

Discours / Sujets	% Conférence	% Débat
BN	42	37
FC	37	68
PL	52	43
PM	61	52
MV	77	55
Moyenne	53.8	51
écart type	14.2	10.6
AMH	41	65
JK	66	53
DLB	79	80
ML	70	63
NM	44	45
Moyenne	60	61
écart type	14.9	11.8

Table 3. E caduc supprimé après une seule consonne prononcée à l'intérieur d'un groupe.

E précédé de plus d'une consonne prononcée.
Dans ce cas, la règle distributionnelle veut que le E caduc soit généralement prononcé comme dans "au cours de", "il se trouve". Les chiffres relevés pour notre corpus sont indiqué dans la table 4.

Discours / Sujets	Conférence	Débat
BN	83	100
FC	87	91
PL	70	88
PM	58	55.5
MV	100	100
Moyenne	79.6	86.8
écart type	14.4	16.6
AMH	77	83
JK	100	75
DLB	90	100
ML	100	100
NM	100	96
Moyenne	93.4	90.8
écart type	9.0	10.0

Table 4. E caduc intérieur maintenu après plus d'une consonne prononcée.

La distribution phonématique attendue est meilleure ici que partout ailleurs. Elle est cependant loin d'être appliquée intégralement. Le groupe des femmes présente une plus grande homogénéité que celui des hommes. Mais dans l'ensemble, les écarts individuels sont assez importants.
Dans plusieurs cas, souvent signalés par les phonéticiens, l'E caduc est supprimé dans cette distribution, en raison de la nature de la séquence consonantique. On trouve ainsi de nombreux exemples de parc(e) que. Mais d'autres facteurs peuvent intervenir.
Le facteur de style semble bien ne jouer aucun rôle ici. Le groupe des hommes présente même un plus haut pourcentage d'E caducs dans le débat que dans la conférence. Pour les femmes, l'écart n'est certainement pas significatif.
Les exceptions. Les principales anomalies à la règle de distribution examinée ici sont les suivantes dans notre corpus:

-E caduc apparaît là où il n'a pas d'existence phonématique virtuelle, comme dans: "un contact[ə] réel".
-E caduc tombe en dépit de la distribution phonématique C C + E, comme dans "j'en apport(e) des preuves".

Règles prosodiques. Les 2 exemples ci-dessus, malgré leur aspect contradictoire peuvent s'expliquer par une même règle prosodique. Elle répond à la question d'André Martinet (6), pourquoi "ours [ə] blanc" et "arc-boutant"? question à laquelle j'avais proposé une réponse par la règle rythmique (3) opposant des séries comme:
([ə] + 2 syll) ≠ ([ə]+ 1 syll.),
`porte(e) `crayon ≠ `porte `plume,
`garde(e) `barrière ≠ `garde `côte,
et la règle qui veut que E caduc tombe plus volontiers à la jonction externe qu'à l'intérieur d'un mot:
"au terme de cette étude" ≠ "une fermeté si amicale".
Fònagy (2) m'a fait remarquer que cette règle rythmique peut être envisagée comme règle accentuelle, le français répugnant à accepter deux accents successifs. Le phénomène se vérifie dans le corpus étudié ici où l'on relève:
le `texte `littéraire ≠ le `texte(e) littéraire
`donc [ə] `sa maman ≠ `donc cette maman
un `conte développé ≠ un `conte(e) facétieux
on peut maintenant intégrer à la série le syntagme: un `contact[ə] `réel
et la plupart des exemples de la liste des E caducs avant accent d'insistance, relevés dans notre corpus.
Introduire un E caduc dans cette position revient à formuler une règle analogue à la règle rythmique de la théorie métrique de Marc Liberman et Alan Prince (4): Un E caduc est inséré pour repousser vers la droite un accent secondaire, la langue n'admettant pas deux accents rythmiques consécutifs.

E CADUC FINAL DE GROUPE

Selon la règle distributionnelle, l'E caduc final tombe. En réalité, nous relevons dans notre corpus plusieurs types de phénomènes, dont les deux principaux sont un E caduc à valeur syllabique,

inaccentué, et un E caduc d'hésitation à durée variable. Le nombre des E d'hésitation est donné en valeur absolue; celui des E caducs en pourcentage par rapport à toutes les réalisations possibles dans cette position, dans la table 5.

Discours	Conférence		Débat	
Sujets	E prononcé (%)	E hésit.	E pron.(%)	E hésit
BN	10	4	20	7
FC	6	2	10	11
PL	7	8	13	19
PM	25	3	15	2
MV	0	0	9	4
Moyenne	9.6	3.4	13.4	8.6
Écart type	9.3	2.9	4.3	6.7
AMH	20	1	32	1
JK	2	2	12	0
DLB	10	2	14	8
ML	0	0	38	20
NM	0	0	27	18
Moyenne	6.4	0.89	24.6	9.4
Écart type	8.6	1	10.1	9.31

Table 5. E caduc prononcé à la finale.

Hésitations finales. Le nombre des hésitations finales, comme celles déjà relevées au milieu du groupe, paraît nettement un facteur individuel, indépendant des phénomènes linguistiques et stylistiques. On peut dire cependant qu'il y en a davantage dans le débat que dans la conférence et que les femmes de ce groupe en produisent en général moins que les hommes.

E final à la joncture de continuité. Sa réalistion dépend souvent de la vitesse d'élocution, dans des cas comme: "le fait est acceptable # dans..." Chez certains sujet le e final d'acceptable sera prononcé si le débit est rapide. Si au contraire le degré de cohésion des deux groupes successifs est rompu par un débit lent ou une pause courte l'E caduc tombera plus facilement.

E caduc final accentué, noyau d'un morphème

Parmi les E prononcés en finale, on relève dans ce groupe d'intellectuels de nombreux exemples - en particulier dans les énoncés argumentatifs - d'E caducs accentués du type: "Il n'en est pas moins vrai que, et"...Les principaux exemples relevés sont: lorsque, alors que, bien que, quoi que, devant une phrase enchassée, à valeur phonétique d'incise.

E caduc, signal sémiotique expressif ou identificateur. On remarquera enfin, dans la table 5, le nombre important d'E caducs prononcés par le groupe des femmes, à la finale. Il s'agit généralement d'E caduc inaccentué et en fin d'énoncé, devant pause totale. Le plus fréquemment, cet E se réalise après occlusive sourde et après un accent d'insistance dont il semble le contre-coup. Cet E, dans notre corpus, pourrait connoter une parlure chic, moderne, jeune. Il se répand actuellement en France, surtout chez les jeunes filles. On l'entend souvent dans les exclamations ou les injonctions, du type "Arrête!" Le timbre alors se délabialise et devient presque [a]. Il peut être très long comme en témoigne le spectrogramme (fig.1) ci-dessous, pour "arrête" (a=112 ms, ʁɛ=300 ms, tə=280ms).

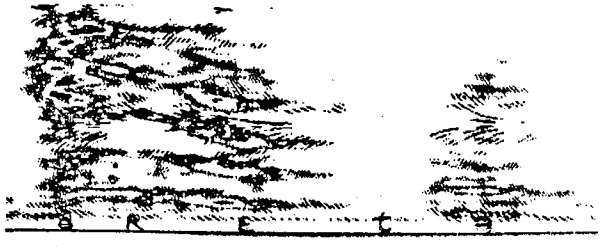


Fig. 1 Arrête! (au spectrographe électronique Ph. Martin).

CONCLUSION

Le jeu du E caduc a montré, dans notre corpus, que les deux types de discours envisagés n'en faisaient souvent qu'un, celui de la parlure intellectuelle. Sa manifestation est redondante puisqu'elle comporte un taux élevé de réalisations d'E caducs non conformes aux règles distributionnelles. Parmi ces réalisations, les unes dépendent de règles prosodiques, accentuelles et rythmiques, (à l'initiale et à l'intérieur), les autres réalisations (en finale) fonctionnent de manière plus aléatoire, comme indices sémiotiques identificateurs, sociolectal ou émotif, ou comme signaux impressifs.

RÉFÉRENCES

(1)S. Bazylko, Le statut de [ə]dans le système phonématique du français contemporain et quelques questions connexes, La linguistique, vol. 17, fasc.1, 1981.

(2)I. Fonagy, L'accent français, accent probabilitaire, in Fonagy et Léon, L'accent en français contemporain, Studia Phonetica 15, Montréal, Paris, Bruxelles, 1979, 122-232.

(3)P. Léon, Apparition, maintien et chute du E caduc, La linguistique, fasc. 2, 1966, 117-120.

(4)M. Liberman and A. Prince, On Stress and Linguistic Rhythm, Linguistic Inquiry, 8, 1977, 249-236.

(5)V. Lucci, Etude phonétique du français contemporain à travers la variation situationnelle, Grenoble, Publications de l'Université des langues et des lettres, 1983.

(6)A. Martinet, La prononciation du français contemporain, Droz, Paris, 2ème éd., 1971.

(7)H. Walter, La dynamique des phonèmes dans le lexique français contemporain, Paris, France Expansion, 1976.

(8)H. Walter, Enquête phonologique et variétés régionales du français, Paris, PUF, 1982.

DURATION AND FORMANT FREQUENCIES OF ITALIAN BIVOCALIC SEQUENCES

PIER LUIGI SALZA
Dept. of Speech Synthesis
CSELT S.p.a.
10148 Torino, Italy

GIOVANNA MAROTTA
Laboratory of Linguistics
Scuola Normale Superiore
56100 Pisa, Italy

DAVIDE RICCA
Institute of Glottology
Universita' di Torino
10100 Torino, Italy

ABSTRACT

A professional speaker read four replications of 34 meanigful sentences containing ia, ai and ua, au bivocalic sequences, under different stress conditions, both in lexical words and at word boundaries. The following measurements were made on each test sequence: onset and offset frequency locations of first and second formants, durations of onset target, glide (transition), and offset target.

Results show that within word clear duration differences exist between diphthongs and vowel sequences in hiatus only in the onglide (rising) case. At word boundary the discriminant issue is the presence or absence of a phrase boundary, according to which word final vowel is preserved or reduced, respectively. In the latter case, the stress carried by final vowel is also lost.

1. INTRODUCTION

Bivocalic combinations are very common in Italian, both within words and at word boundaries. Traditional grammars usually classify the realizations permissible within word as diphthongs or as bisyllabic vowel sequences (hiatus). For word boundaries, they basically account for two phenomena: sinaloephe, i. e. the fusion of the two elements, and elision, where the final vowel of the word is dropped.

Moreover, as for diphthongs within words, traditional Italian phonology makes a distinction between onglide (rising) and offglide (falling) diphthongs, see e. g. Battaglia and Pernicone [1], Romeo [2], Muljačić [3], Tagliavini and Mioni [4], Canepari [5]. The presence of one segment marked [+high] and the presence of stress on the [-high] vowel are invoked as the main conditions for diphthong realization. Nonetheless, also in unstressed contiguous vowels contrasting pronunciations are meant to occur depending on morphological and phonological rules, see e. g. Romeo [2] and Muljačić [3]. In conclusion, the studies quoted above point to the phonetic transcription scheme shown in the left column of Table I (where [a] is selected as [-high] full vowel).

Experimental work undertaken in the past year aimed at measuring the segmental durations which a speaker performs to produce contrasting diphthongs and vowel sequences in a wide corpus of nonsense words [6]. The results indicated that stressed structures could barely be differentiated in diphthongs and hiatuses, as long as reduction phenomena were not present; in unstressed cases only onglide diphthongs were clearly characterized by very short durations of the first element, namely [j] and [w], while no strong evidence was shown for postvocalic reduction.

In the present work a further assessment of the pertinence of the classification adopted by traditional phonology in lexical words, and of its possible extension at word boundary, is carried out on the basis of durational and spectral measurements in a carefully selected corpus of meaningful sentences.

2. SPEECH MATERIAL

The corpus consists of two sections: a) words from the lexicon; b) pairs of lexical words including vowel combinations at word boundaries. In both sections, [a] was chosen as [-high] full vowel. In the following, the stress mark will be reported for the lexical stress and will precede the stressed syllable.

Section a). All the items to be tested are placed in the same syntactical and intonational context, corresponding to a simple meaningful sentence having the following structure: V SN2 Adv, with the subject omitted (as frequently happens in Italian). The target word always holds the postverbal SN2 position and the syntactical role of direct object. The other two main stresses in the sentence are kept in the same position throughout, as we employed the third singular form of the simple past stressed on the last syllable for V, and modal adverbs ending in -mente for Adv. The test words represent typical cases of hiatuses and diphthongs, as witnessed by phonetical transcriptions or morpho-phonemic rules found in previous standard works on Italian phonology, quoted above. The list of test words is shown in Table I.

Section b). As for word boundary, no clear-cut rules about diphthong or hiatus-like pronunciation can be inferred from the literature. Thus, we attempted to assess the possibility of different pronunciation strategies depending on the existence versus absence of a phrase-boundary. This was done exhaustively only for ia and ai combinations, not only for obvious reasons of economy, but also because the lack in the Italian lexicon of words ending with unstressed u would have made thorough comparisons impossible in any case. As typical

examples of no-phrase-boundary contexts, we chose Aux-V or nominal predicate structures, like dovresti andare ("you should go") and si senti' adatta ("she felt beloved"). On the contrary, we took as a typical phrase-boundary the separation between subject and predicate, SN | SV, in sentences like Anna inizio' il compito ("Anne began the work"). The list of word boundary contexts is shown in Table III: in the left column vowel sequences are given in graphemic form.

3. EXPERIMENTAL METHOD

The list was read four times by a professional speaker and the speech was recorded on high quality video cassettes. The speaker was required to use great care in order to maintain a constant mean speaking rate of about six syllables per second. The recordings were digitally converted using a 12 bit A/D converter with 12 kHz sampling rate. Broad-band digital spectrograms were generated by the FFT computation, with a 2 msec frame rate, and the signal was analysed by means of interactive software and videographic facilities, which helped the operator to determine segment boundary location.

Temporal values were measured from both spectrograms and waveforms by applying widely accepted rules whenever possible [7], [8]. The following acoustic parameters were measured:

1. Onset target, glide (transition) and offset target durations of the second formant, in all the items;
2. Onset and offset frequency locations of first and second formants, in ia and ai contexts of section a).

At times, the starting point of F2 transition was partially masked by the raising towards the locus characterizing the neighbouring consonant, especially for stop consonants. In these cases, we took the behaviour of F1 into account, which showed a clear maximum separating the transition region from the region of coarticulation with the consonant, thus enabling to determine, with good consistency throughout the corpus, a boundary point for subsegmental duration measurements (see Figure I).

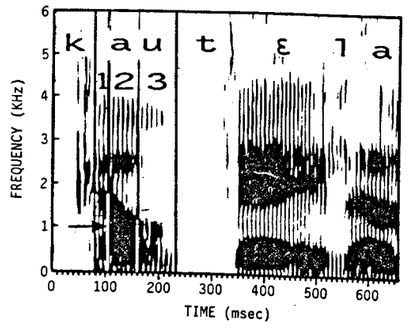


Figure I - Example of target identification strategy in the contiguous vowels [au] of the word cautela, by using the F1 maximum criteria described in the text. ⟶: F1 maximum; 1: onset target; 2: transition; 3: offset target.

4. RESULTS

The analysis of the results will be divided into two main sections (4.1. and 4.2.) corresponding to the two parts of the corpus.

4.1. Lexical items

Table I shows the whole duration of the bivocalic sequences and the percentage duration of the onset target, transition and offset target in each context. Data clearly suggest that target duration is quite short for [j,w], since it reaches 17% of the whole duration at most. Moreover, in the case of unstressed [ja, wa], the whole duration is also very short, i. e. lower than 100 msec, while in every other context it is greater than 145 msec. This result provides evidence for the traditional transcription of rising diphthongs and consequently for the phonological representation of [j,w] as semiconsonantal phonemes. On the other hand, the traditional transcription of falling diphthongs is not confirmed by the data: their whole duration is substantially similar to that of sequences classified as hiatuses; the target duration of i and u in falling stressed sequences (daino, fauci) is not different from that of any other unstressed vowel, ranging between 20% and 30%; finally, in all falling unstressed sequences (arcaicita', aizzatore, cautela, baulino) the i and u targets are even considerably longer than the [a] target.

PHONETIC TRANSCRIPTION	ITEM	% ONSET TARGET MV	% ONSET TARGET SD	% TRANSITION MV	% TRANSITION SD	% OFFSET TARGET MV	% OFFSET TARGET SD	WHOLE DURATION MV	WHOLE DURATION SD
ja	pianale	16	3	53	5	31	5	96	6
'ja	diada	-15	2	25	1	60	2	243	4
ia	sciatore	32	3	34	2	34	1	147	5
aj } ai	arcaicita'	27	3	34	3	39	5	171	9
ai	aizzatore	28	3	31	3	41	1	179	13
i'a	sciata	25	2	25	2	50	1	269	6
'ia	diade	47	1	29	3	24	2	222	6
'aj 'ai	daino	45	7	30	4	25	3	218	,17
a'i	Caino	21	2	24	3	55	1	266	11
wa	qualita'	17	2	46	8	37	8	84	6
'wa	inattuato*	15	2	36	5	49	5	216	15
ua	dualismo	43	7	34	10	23	4	165	13
aw } au	cautela	18	2	37	4	45	5	145	15
au	baulino	27	1	38	3	35	3	171	15
u'a	zuavo	20	2	25	3	55	2	250	8
'ua	tua	54	3	29	2	17	2	253	2
'aw 'au	fauci	48	2	24	3	28	2	242	24
a'u	baule	20	2	33	4	47	4	257	12

Table I - Diphthong versus Hiatus within word. Subsegmental durations expressed as percentage of whole duration, and whole duration in msec. MV: Mean Value. SD: Standard Deviation. *: in most dictionaries it is transcribed as [ua], but in currently spoken Italian and also by our speaker it is pronunced [wa] (see also Muljačić [3, p.86] about luttuoso).

Consequently, we think that purely acoustic data of our speaker show that it is not necessary to maintain the traditional transcription with "semi-vowels" [i̯,u̯] (see the second column of Table I).

In order to check the statistical validity of our conclusions, we performed a T-test analysis (see e. g. Welkowitz et al. [9, p.159]) on the mean durations of the [+high] vowels. The results, as detailed in Table II, are quite satisfactory, since the calculated values are lower than the critical one for [i⟷i], [u⟷u] and higher for [j⟷i], [w⟷u], showing that there is statistically significant difference between the mean durations only for the last two couples.

ITEM	t value	ITEM	t value
ja ⟷ ja	10.98	wa ⟷ ua	11.50
'ja ⟷ i'a	8.19	'wa ⟷ u'a	4.61
ai ⟷ ai	1.29	au ⟷ au	2.81

Table II - T significance test computed for the mean durations of the [+high] vowel of each couple. At the confidence level P<0.01, the critical value is 3.71.

It is interesting to note that bivocalic sequences stressed on the second segment have longer durations than those stressed on the first (see again Table I). Together with the shortness of semiconsonants indicated above, this datum would suggest a trend towards weakening of the first element of the sequence. This tendency becomes even stronger at the word boundary, as will be shown later.

The transition length does not show any clear trend, apart from a longer percentage duration in the case of unstressed sequences. The hypothesis concerning correlation between transition duration and diphthong/hiatus contrast (see Bond [10] for this hypothesis as regards English) is not borne out by evidence.

The rate of change of F2 transition (Hz/msec) is a further acoustic parameter taken into account in the literature (see, among others, Lehiste and Peterson [8], Gay [11], Borzone de Manrique [12]). We measured this parameter in ia and ai contexts of lexical words. Consistent differences between assessed diphthongs and the remaining sequences are not apparent, and the data are not worth publishing. Similar results were also found in Spanish [12]. As it is known, F2 frequency locations of vowels show great variability depending on consonantal context [13]; but even when bordering consonants were kept in the same articulatory class, many of our measurements showed steady targets well above the mean values of our speaker. For example, in un aizzatore and in diade the F2 of [a] was 1722 ± 20 and 1786 ± 9, respectively, compared to the mean value of 1656 ± 20 typical of our speaker in apicodental nonsense contexts. Even more dramatic cases are present in word-boundary data. For instance, the expected [a] of Renata indica heard in isolation was invariably perceived as [e], which is not surprising due to its formant pattern: F1 567 ± 34, F2 1851 ± 51.

As Lindblom and Studdert Kennedy [14] have demonstrated, the identity of a vowel sound is likely to be determined not only by the formant patterns at the point of closest approach to target values but also by the direction and rate of adjacent formant transitions which could compensate for possible undershoot effect. Thus, both the formant pattern and the F2 rate of change per se are likely to have little or no significance. A good proof of these phenomena is provided by our data about i in lexical items. Only stressed [i]'s reach the F2 target value typical of our speaker, about 2500 Hz, while for shorter i's the F2 steady states are located at much lower frequencies. A clear positive correlation between F2 steady-state duration and its frequency location is indeed detectable, as shown in Figure II: the linear correlation coefficient r is .915, which is higher than the critical value 0.798 (at the confidence level P<0.001), so that r is statistically significant [9, p.182].

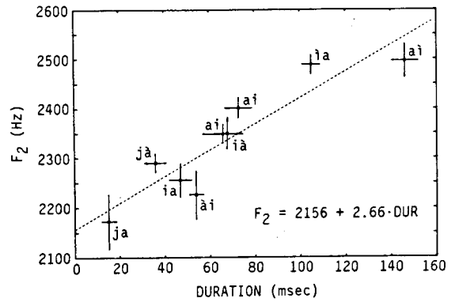


Figure II - F2 frequency location of i in lexical words as a function of its F2 steady state duration. ——: Standard Deviation.

4.2. Word boundary items

Turning now to the bivocalic sequences at word boundary, it must be noted first of all that our data concern ia and ai sequences only, as explained in the Introduction. Here a new factor is important, namely the presence versus absence of syntactic boundary. Moreover, a new type of vowel sequence is possible: one where both contiguous segments bear graphical stress. The mean absolute values of subsegmental durations of sequences at word boundary are shown in Table III. In this case, in fact, absolute values appear more suitable than percentage values for pointing out the relevant feature of our data, that is to say the weakening effect on the first target (word final vowel).

GRAPHEMIC FORM	ITEM	ONSET TARGET MV	ONSET TARGET SD	TRANSITION MV	TRANSITION SD	OFFSET TARGET MV	OFFSET TARGET SD	WHOLE DURATION MV	WHOLE DURATION SD
ia	dovresti andare	16	3	49	9	30	9	87	11
i|a	Giovanni ando'	34	4	52	4	41	5	127	6
ai	sembrava iniziato	10	1	52	7	31	7	94	10
a|i	Anna inizio'	59	21	56	14	65	2	182	7
'ia	si senti' adatta	40	10	53	2	55	8	151	17
'i|a	il salmi' ando'	72	9	59	9	56	15	188	16
'ai	sara' idiota	34	13	74	8	58	13	167	20
'a|i	l' attivita' inizia	96	18	63	9	49	5	205	15
i'a	saresti acida	59	5	52	4	135	2	246	1
i|'a	Giovanni agita	87	9	66	4	106	11	260	23
a'i	restava intimo	53	5	58	3	79	12	191	14
a|'i	Renata indica	43	6	66	15	91	7	198	17
'i'a	si senti' acida	66	9	59	5	133	6	261	6
'i|'a	Mimi' agita	123	6	99	7	123	2	346	6
'a'i	sara' ibrida	30	10	85	6	90	13	206	9
'a|'i	la verita' indica	118	11	77	5	103	5	299	12

Table III - Phonetic events at word boundary. Subsegmental duration and whole duration, in msec. MV: Mean Value. SD: Standard Deviation. | : phrase boundary

In the absence of phrase boundary, the acoustical realization is often very far from the graphemic transcription. When both segments are unstressed, the first undergoes a strong length reduction. Complete elision occurred in many cases not included in the table. For example, in the noun phrase data iniziale the ai context lasted about 70 msec in all replications and its measured and per-

ceived formant pattern was that of a single vowel lying between [e] and [i]. The presence of a syntactic frontier acts as a strong protection against the quoted weakening phenomenon, keeping both targets within comparable durations.

It is interesting to observe that when the boundary is not present, even the graphical stress does not protect the first element from the reduction process: in si senti' adatta and sara' idiota the first target is shortened to 40 msec on the average, while in the corresponding phrase boundary contexts the same target lasts about twice as long. These effects are also quite clear in the case of two-stress combinations: in the absence of syntactic frontier the word final vowel loses its stress completely, whereas in the opposite case it remains well preserved.

5. CONCLUSIONS

From the experimental analysis carried out on bivocalic sequences within lexical word in connected speech, we can claim that the traditional distinction between diphthongs and hiatuses in Italian is not confirmed by purely acoustic data: while for onglide (rising) diphthongs it is possible to establish specific durations different from those typical of hiatuses, offglide (falling) diphthongs do not differ from vocalic sequences classified as hiatuses. Thus, a phonetic transcription based on acoustic data should recognize semiconsonants [j,w], but not semivowels [i̯,u̯], which appear to be the same as the corresponding unstressed vowels.

The hypothesis concerning a possible extension of the distinction between diphthong and hiatus to the context of word boundary has not been proved, since in sentences new factors occur, in particular presence or absence of syntactic phrase boundary between the two test words. Depending on whether this boundary exists or not, word final vowel is quite preserved or reduced, respectively. In the latter case, the stress carried by the final vowel is also lost.

On the other hand, data suggest that the Italian speaker, tending to avoid production of two full contiguous vowels, chooses to reduce the first segment of the vocalic sequence, even when it is marked [-high].

ACKNOWLEDGMENTS

This work was partially supported by an EEC (European Economic Community) contract in ESPRIT "SPIN" Project N. 64.

REFERENCES

[1] S. Battaglia and V. Pernicone, GRAMMATICA ITALIANA, Loescher, Torino, 1951.

[2] L. Romeo, "A phonemic inventory of the Italian bivocalic sequences", FORUM ITALICUM, 1968, 2, 117-143.

[3] Ž. Muljačić, FONOLOGIA DELLA LINGUA ITALIANA, Il Mulino, Bologna, 1972.

[4] C. Tagliavini and A. M. Mioni, CENNI DI TRASCRIZIONE FONETICA DELL' ITALIANO, Patron, Bologna, 1972.

[5] L. Canepari, INTRODUZIONE ALLA FONETICA, Einaudi, Torino, 1979.

[6] P. L. Salza, "La durata dei suoni nelle sequenze vocaliche dell' Italiano", ELETTRONICA E TELECOMUNICAZIONI, 1986, XXXV, 27-34.

[7] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in English", J. ACOUST. SOC. AM., 1960, 32, 693-703.

[8] I. Lehiste and G. E. Peterson, "Transitions, glides and diphthongs", J. ACOUST. SOC AM., 1961, 33, 268-277.

[9] J. Welkowitz , R. B. Ewen and J. Cohen, INTRODUCTORY STATISTICS FOR THE BEHAVIORAL SCIENCES, Academic Press, London-New York, 1982.

[10] Z. S. Bond, "The effect of varying glide durations on diphthong identification", LANGUAGE AND SPEECH, 1978, 21, part 3, 253-263.

[11] T. Gay, "A perceptual study of American English diphthongs", LANGUAGE AND SPEECH, 1970, 13, 65-88.

[12] A. M. Borzone De Manrique, "Acoustic analysis of the Spanish diphthongs", PHONETICA, 1979, 36, 194-206.

[13] F. E. Ferrero, "Caratteristiche acustiche dei fonemi vocalici italiani", PAROLE E METODI, 1972, 3, 9-34.

[14] B. E. F. Lindblom and M. Studdert Kennedy, "On the role of formant transitions in vowel recognition", J. ACOUST. SOC. AM., 1967, 42, 830-843.

SPECTRAL PROPERTIES OF RUSSIAN STRESSED VOWELS IN THE CONTEXT OF PALATALIZED AND NONPALATALIZED CONSONANTS

VLADIMIR B. KUZNETSOV

Moscow State Institute of
Foreign Languages
Moscow, USSR 119034

ARVO OTT

Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 200108

ABSTRACT

Results of an experimental study of the formant frequencies at the steady-state part of russian stressed vowels in the context of palatalized/nonpalatalized consonants are presented. The analysis of the obtained data reveals that there is a considerable difference in the quality of vowels due to the consonantal environment. Traditional static description of vowel quality (F-pattern measured at the stationary portion of vowel) is basically insufficient for representation of the allophonic variation. One must take into account the dynamic properties of transitions as well.

INTRODUCTION

Specific character of the allophonic variation of russian stressed vowels is in great part attributed to the palatalization/nonpalatalization (P/NP) of the surrounding consonants. There is practically no coarticulation between palatalized consonant and the following vowel: the onset frequency of the second formant transition is primarily determined by the palatalization of the consonant [1]. Under these circumstances one should expect the quality of the vowel to undergo some changes. In fact, experiments on the identification of stressed vowels, segmented from monosyllables or words uttered in isolation, have shown that russian listeners are able to recognize some 18 vowels depending on consonantal environment [2]. On the basis of spectrographic analysis of these vowels it was concluded that the listeners' ability to distinguish so many allophones was due to the presence in the structure of vocalic nucleus of the so called "i-like" second formant transition - acoustically and perceptually reliable cue of consonantal palatalization.

Judging from the data on the second formant frequency of the vowels measured at their stationary part [3], the phonetic quality of the vowels as such does not change under the influence of the preceding P/NP consonants of different place of articulation. This observation is supported by the results obtained in a study of formant frequency patterns in russian VCV utterances [4].

Entirely different results are reported in [5,6]. It is shown that palatalization of surrounding consonants produces systematic effect on the steady-state part of the vowels uttered in isolated syllables or words.

In recent years experimental studies of vowel perception have provided some new evidence that leads us to question the correctness of the view that russian listener recognize vowel allophones on the basis of characteristic "i-like" formant transition.

As it has been shown in [7], Russian children, having yet no knowledge of foreign languages, can classify a set of 20 stationary vowels into 13 categories, some of which (for example the front vowels) are not listed in the phonemic inventory of Russian. Experimental study of perception of the vowel-like stimuli with a changing frequency of the second formant [8] has revealed that the onset frequency and the direction of the formant transition are useful perceptual cues for identification of the vowel quality. A recent experiment on the identification of the steady-state part of the vowels segmented from continuous speech has shown that these segments convey information concerning P/NP of the consonantal environment [11].

From all the facts presented above it is clear that both the spectral data and the proposed interpretations of its perceptual significance are contradictory and incomplete. To make some progress in understanding the nature of the allophonic variation in question, one must begin by collecting basic quantitative data on the spectral properties of the vowels. The present study was designed to fulfil in part this task.

SPEECH MATERIAL AND METHOD

Test vowels were uttered in a nonsense monosyllable of CVC type embedded in the carrier phrase "Say ... again". Palatalized or nonpalatalized fricative [s] was used to form a symmetrical environment. Each of the three male speakers recorded a list of 330 sentences (10 vowels * 33 repetitions). To achieve constant speech rate throughout the recording session the speaker was asked to synchronize the onset of the sentence with a periodic light pulse. The phonetic identity of the test vowels was checked up by 8 listeners in an identification experiment. Spectral analysis was computer-implemented. Prefiltered speech signal was sampled at 10 kHz by an 8-bit A/D converter and subjected to Fast Fourier transformation to compute power spectrum of the signal. The analysis window shift was 128 points. An automatic algorithm was used to compute the frequencies of the first three formants [12]. The test vowels were characterized by an F-pattern measured at the point where F2 reached its extreme value or in the middle of the vowel if there was no extreme. To enhance the reliability of formant peak location wideband sonograms were regularly made. By tracking formant trajectories on the sonograms in the vicinity of the vowel segment we were able to identify and discard spurious peaks present in the vowel spectra.

RESULTS AND DISCUSSION

The data obtained are presented in Table 1 and Fig.1. The table lists the means and standard deviations of the formant frequencies of the three speakers. Fig.1 displays the data of speaker M1 on the F1-F2 space. We use cyrillic characters to symbolize vowel allophones. The dots above transcription sign indicate the allophone surrounded by palatalized consonants. The results of the vowel identification test are as follows: speaker M1 - confusion error rate 0.30 %, speaker M2 - 4.00 %, speaker M3 - 1.40 %. The range of the vowel durations of all speakers is from 75 to 120 msec. For speaker M1 and M2 the duration means of the test vowels are reported in [10]. The vowel configuration depicted in Fig.1 is typical, exept for some minor details, to all the speakers. The most obvious conclusion to be drawn from the figure is that the F-pattern of the vowels in the context of palatalized consonants differ greatly from that one in the context of nonpalatalized consonants. The observed differences are much greater than those reported in [5,6] but in both cases they clearly go along the same lines. In the F1-F2 space the vowels [ÿ] and [ö] occupy the areas that are usually labelled as [y] and [øe] respectively in the languages that have front rounded vowel phonemes. Our data does not support the traditional phonetic notion that in Russian [ə] is more close and front than [ä]. All the speakers have their second formant of the vowel [ä] higher than that one of [ə], which indicates a more forward position of the tongue.

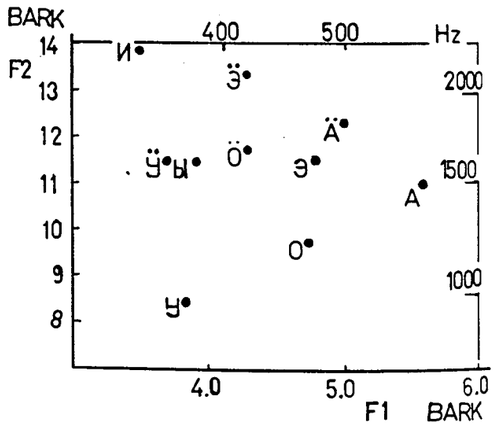


Fig.1. Vowel F-pattern of speaker M1

Examination of Fig.1 and the tabled data shows that the areas occupied by the vowel pairs [ÿ-ы], [ы-ö], [ö-э] and [э-ä] in the F1-F2 space may considerably overlap. Using only F-pattern at the steady-state part of the vowel there is no way to find out which of the two vowels we are dealing with.

The identification data concerning the speech materials of speakers M1,M2,M3 and two more female subjects indicate that the listeners most often confused the following vowel pairs: [ä-и] - 111 errors, [о-у] - 64 errors, [э-ы] - 29 errors, [ö-ÿ] - 28 errors and [а-э] - 26 errors.

Evidently, the vowels are confused when they are uttered in the same consonantal environment. The vowels that are the nearest neighbours in the F1-F2 space (for example, [ÿ-ы], [ы-ö], [ö-э],[э-ä]) but never occuring in the same consonantal context are discriminated quite easily.

The role of the P/NP consonants in the phonetic interpretation of the quality of vowel sounds was demonstrated in the investigation [9] where natural vowel in CVC syllable was replaced by synthetic stationary stimuli. The stimuli with the following coordinates in the F1-F2 space: 470/1560 Hz and 520/1560 Hz - were perceived as [ö] in the context of palatalized consonants and as [э] when the consonants were nonpalatalized; the stimuli with the coordinates 320/1420 Hz and 350/1420 Hz were recognized as [ÿ] in the first case and as [ы] - in the second.

Table 1. Formant frequency means and standard deviations for speakers M1,M2,M3

		F1		F2		F3	
а	M1	578	24	1501	27	2558	30
	M2	520	21	1489	50	2493	106
	M3	560	35	1521	36	2492	64
ä	M1	506	22	1810	36	2696	30
	M2	504	18	1943	76	2459	158
	M3	496	25	1882	98	2514	79
э	M1	487	22	1633	39	2518	29
	M2	479	26	1674	58	2569	100
	M3	452	22	1686	56	2453	32
ы	M1	429	26	2170	50	2801	53
	M2	403	25	2192	90	2498	324
	M3	405	19	2221	69	2621	155
о	M1	470	23	1210	44	2481	48
	M2	434	28	1192	124	2667	167
	M3	441	23	1117	56	2401	76
ö	M1	424	16	1652	51	2510	34
	M2	411	28	1720	112	2210	169
	M3	407	20	1611	179	2209	64
у	M1	372	19	976	61	2410	40
	M2	371	20	1064	94	2773	230
	M3	392	14	951	69	2420	69
ÿ	M1	346	16	1624	130	2391	51
	M2	372	20	1705	115	2325	130
	M3	385	13	1926	194	2297	57
и	M1	375	21	1627	83	2364	50
	M2	385	16	1739	89	2430	176
	M3	405	19	1740	131	2398	85
и	M1	337	21	2322	53	2813	47
	M2	361	20	2329	47	2960	98
	M3	374	20	2246	70	2603	81

An identification experiment [8] with the vowel-like stimuli having F1=300 Hz throughout stimulus duration and F2 consisting of symmetrical initial and final transitions and steady-state part varing from one stimulus to another in the range of 1200-1600 Hz, has shown that the stimuli were perceived as [ы] when the slope of the initial F2 transition was positive and as [ÿ] when it was negative.

In Russian, as it is known [1], the second formant transition from palatalized consonant into the stationary part of the vowel [ÿ] is falling and it is rising from nonpalatalized consonant into [ы]. Thus, it might be concluded, that in Russian the second formant transition from consonant into the following vowel not only conveys information concerning the place of consonant articulation or its palatalization, but is a useful cue for the phonetic classification of vowels as well.

A complex analysis of the data on inherent vowel duration [10], vowel spectra and confusion errors leads us to the assumption that vowel identity is coded not only in its spectral parameters but in its duration as well, though in Russian the duration of vowel is not phonologically significant. There seems to be an inverse relationship between the spectral and the perceptual similarity of the vowels, on the one hand, and their difference in inherent durations, on the other. For example, speaker M1 provides for the poorly discriminated vowel pairs the following duration contrasts, expressed in percent to the range of inherent vowel durations, that in his case is 33 msec: [и-ÿ] - 88 %, [ы-э] - 64 %, [у-о] -45 %, [ÿ-ö] - 35 % and [э-а] -20 %. An interesting aspect of these data is that large differences in duration are established for those vowels that are spoken, judging from the spectral data, with the same tongue height - the most significant and linguistically universal factor determining inherent vowel duration. It is tempting to make a suggestion that in Russian the duration of some vowels does not result automatically from the universal physiological limitations of the articulatory mechanism, but is generated at a higher control level.

It is necessary to touch upon the subject of the vowel production in the context of P/NP consonants. In the conventional view it is assumed that in both cases vowels are produced by the same articulatory gesture. As to the observed differences of the F-pattern at the steady-state portion of the vowels they are ascribed to the effect of the "undershoot". It is believed that the target articulation could be achieved if the duration of the vowel is made large enough. We think that the amount of spectral differences found in this study for the vowels in question could not be explained within the framework of the "undershoot" theory. We support the view developed by Lindblom that the undershoot observed for vowels during faster speech is programmed into the gesture and is not a result of a too fast succession of motor commands. The essence of the speech production process is not an inefficient response to invariant control signals but elegantly controlled variability of response to the demant for a relatively constant end [13].

CONCLUSION

In the present work it has been shown that there is a considerable difference in the spectral properties of russian stressed vowels at the steady-state part depending on the P/NP of surrounding consonants. The

allophonic variation can not be adequately represented if one uses only static characteristics (F-pattern at the stationary part). The dynamic properties of the second formant transition should be taken into consideration as well. Data from the present study together with [10] indicates that there is a tendency to compensate for the lack of spectral contrast between vowels by increasing their difference in duration.

ACKNOWLEDGEMENT

The authors would like to thank T.Timmerman for his helpful assistance in stimuli measurements.

REFERENCES

[1] O.F.Krivnova, V.B.Kuznetsova "Phonetic nature of consonantal feature palatalization/nonpalatalization in Russian" In: "Theoretical and historical problems in Russian" (in russian), Moscow, p.37-67, 1980

[2] L.V.Bondarko, L.A.Verbitskaya et al. "Recognizable sound units in Russian speech" In: "Production and perception mechanism of complex sounds" (in russian), Moscow, p.165-179, 1966

[3] L.V.Bondarko "Phonetic representation of language and phonological representation of speech" (in russian), Leningrad, 1981

[4] E.T.Purcell "Formant frequency patterns in Russian VCV utterances" J.Acoust.Soc.Am., vol.66, p.1691-1702, 1979

[5] B.M.Lobanov "Classification of russian vowels spoken by different speakers" J.Acoust.Soc.Am., vol.49,p.606-608, 1971

[6] M.Romportl "On the vowel system of Russian" In: "Studies in phonetics", Prague, p.37-58, 1973

[7] V.B.Kuznetsov "Classification of steady-state vowels by russian listeners" (in russian) In: "Sensory systems", Leningrad, p.105-116, 1979

[8] V.V.Lublinskaya, N.A.Slepocurova "Perception of vowel-like sounds with a changing spectrum" (in russian), Human psychology, Moscow, 3(1), p.77-84, 1977

[9] N.V.Puhir, L.P.Sherbacova "Perception and phoneme classification of phonetically ambiguous vowels" (in russian) In: "Proc. of seminar on automatic recognition of auditory patterns", Lvov, v.2, p.77-80, 1974

[10] V.B.Kuznetsov, A.Ott, A.V.Ventsov "Inherent vowel duration in Russian: production and perception data" (contribution to the present congress), Tallinn, 1987

[11] V.I.Garbaruk "Feature perception of speech sounds" (unpublished dissertation, in russian), Leningrad, 1984

[12] J.D.Markel "Digital inverse filtering, a new tool for formant trajectory estimation", IEEE Trans. Audio Electroacoust., vol. AU-20, p.129-137, 1972

[13] B.Lindblom "Chairman's comment on session 2" In: "Speech Communication Proc. of SCS", Stockholm, vol.2, IX-XXV, 1974

COMPUTER AIDED ANALYSIS OF STRESSED AND UNSTRESSED BULGARIAN VOWELS FROM 30 MALE AND 30 FEMALE SPEAKERS

PHILIP CHRISTOV

Voice Man-Machine Communication Lab.
Mechanics & Biomechanics Institute
Sofia, Bulgaria, 1090, P. O. Box 373

ABSTRACT

In a previous paper (CHRISTOV, Proc. 11th ICA, vol. 4, pp. 161-164) an algorithm has been reported about the analysis of speech in the spectral domain by the use of phonetic knowledge. Here the results of the application of this algorithm will be discussed in the computer aided analysis of the Bulgarian vowels. The vowels are imbeded in words and uttered by 30 male and 30 female speakers in /b-b/ context. The Computer input is verified by 20 listeners.

INTRODUCTION

The research reported in this paper is aimed at the build up of an acoustic-phonetic fulcrum of the Bulgarian vowel system for the purposes of the mashine recognition of continuous speech.
Following that goal a research strategy is adopted which was implemented successfully in vowel analysis by Potter & Steinberg /1/, Peterson & Barney /2/ and others. The root of it is in using phones of comparable allophones imbeded in words uttered in equal phrases with equal intonation by 25 to 30 male and female speakers.
In addition here central allophones are used together with their unstressed contrasts and the persons who uttered them are carrefully selected among the best professional speakers in the country.

SPEECH DATA

The input speech data /3/ form a body of 614 vowel phones belonging to the central Bulgarian allophones /bV̀b/ and to their unstressed contrasts /bVb/, where V stands for a vowel from the Bulgarian vowel system. They are uttered in Standard Bulgarian by 30 male bariton and 30 female mezzo-soprano professional speakers in a highly damped room and recorded with studio equipment. The vowel /b-b/ context speech fractions are imbeded in words (See APPENDIX) pronounced by the speakers with falling intonation at the end of a standard carrier sentence. A semiautomatic procedure /4/ was used for the verification of the experimental material by 20 listeners. The verification procedure rejected both false listeners and incorrect utterings so that to the computer input have been admited only the utterings for which full agreement was reached between the speaker and each of the reliable listeners.
The analog speech signal was digitalized with a frequency of 20 kHz and stored on 1/2 inch IBM-compatible magnetic tapes. This way four machine compatible sets N, where N=I,II,III,IV, of vowel utterings have been prepared to be used as machine input in this research:

I	- Set of	UNSTRESSED	vowels uttered by	FEMALE speakers
II	- Set of	STRESSED	vowels uttered by	FEMALE speakers
III	- Set of	UNSTRESSED	vowels uttered by	MALE speakers
IV	- Set of	STRESSED	vowels uttered by	MALE speakers.

ANALYSIS

The speech analysis algorithm /5/ uses phonetic knowledge to perform a pitch synchronous Fast Fourier Transform and to determine the formant frequencies in the quasi-stationary region of vowels.
The software realization of the algorithm is modular with input and output modules so that each principal module can be used either as an independent program or as a subroutine in a main program. It has the shape of an application program package including program modules for:

Pitch Extraction
Waveform Analysis
Quasi-Stationary Region &
Representative Section Determining
Fourier Transform
Formant Tracking
Phonetic Classification & Correction

The pitch extraction module is programmed after an algorithm of B. Gold /6/.
The Fourier Transform has been limited to the 30-th harmonic of the fundamental frequency because above that harmonic no traces of higher harmonics have been detected.
The Fourier Transform module gives its output except on cards but also in the form of a computer diagram of the amplitude spectra of the representative section.
After the processing of all vowel utterings with an IBM 360/40 computer it has been established: 1) That the algorithm performes better by male voices and unstressed vowels; 2) That in 18.06% of all vowel utterings it failed to determine the position of the second formant. In all vowel utterings, not resolved by the computer, the position of the second formant has been determined manually using the machine printings of their amplitude spectra.

RESULTS

In result of the analysis the /b-b/ context vowel utterings turned into labeled vectors

$$(F_0, F_1, F_2, F_3)_{LABEL}$$

or points in the measurement space of the fundamental and the firs three formant frequencies. Each vector or point in the measurement space, together with its label, is punched on a single card. The manually corrected vectors are also punched on cards and added to the computer output to form together a set of statistical distributions of equally labeled points

$$W_{NV} \subset W_N,\ N=I,II,III,IV;\ V=i,e,a,ə,o,u$$

in the measurement space. These statistical distributions will be called here, for clarity, also "individual vowel clusters".
A program for statistical processing is used to print cross-section diagrams of the statistical distributions and to compute their statistical estimates: mean, standard deviation, maxima, minima, skewness and kurtosis. The results of the statistical processing of the individual vowel clusters W_{NV} will be presented in a more technologically oriented report.
The commonly accepted two-dimensional graphic representation of three-formant analysis of vowels /7/ is adopted also in this study. The graphs are produced automatically by a computer program which first calculates the value of the approximate measure

$$F'2 = F2 + \frac{1}{2}\,\frac{(F2 - F1)}{(F3 - F1)}\,(F3 - F2)$$

of the effective pitch of the higher formant group and then prints the usual F1xF'2 vowel diagrams in the computer output listing.
Each of the supplemented figures (Fig.Fig. 1 to 4) shows six closed loops plotted manually around the main body of an individual vowel cluster $W_{NV} \subset W_N$ in the machine F1xF'2 graph.
The amount of points inside each closed loop and of these which remain outside or enter alien closed loops is presented in attendant confusion matrixes (Tables 1 to 4).

DISCUSSION

The first impression after looking at the four diagrams in the last page of the report is that by the stressed vowel male utterings (Fig. 4) the closed loops are neatly outlined and distinctly delimited. There is no overlapping at all but many single points, which are far away from their clusters nuclei, remain outside of the closed loops (Table 4).

In contrast, by the stressed female utterings the distributions are more uniform and widely spread so that there is much overlapping (Fig. 2) but very few points remain outside the closed loops (Table 2).
By the unstressed utterings (Fig.Fig. 1 and 3) the clusters shrink around their nuclei but, nevertheless, there is some overlapping between the clusters of /o/ and /u/ for the male voices (Table 3) and between /a/ and /ə/ for the female voices (Table 1).
Despite of the overlapping the general set up of the vowel triangel is well preserved in the F1xF'2-diagrams of all four vowel cluster sets W_N shown in the figures.
A closer examination of the vowel triangles presented in Fig.Fig. 1 to 4 shows that by stressed vs. unstressed contrasting (Fig.2 vs. Fig. 1 and Fig. 4 vs. Fig. 3) there exsists a certain stretching of the vowel triangle along the F1-axis (Left to right in the F1xF'2 area).
A similar stretching can be detected during the female vs. male utterings contrasting (Fig. 1 vs. Fig. 3 and Fig. 2 vs. Fig. 4) but this time along the F'2-axis (Bottom to top in the F1xF'2 area).

CONCLUSION

It appears that, due to the carrefully selected and handled speech material in this study and facilitated by the simple Bulgarian vowel system, it has been possible to come within reach to some natural phenomena concerning the differences between stressed and unstressed vowel clusters produced by men and women.

REFERENCES

/1/ K. R. POTTER, J. C. STEINBERG, "Towards the Specification of Speech", JASA, 22, 807-820, 1950; /2/ G. E. PETERSON, H. L. BARNEY, "Control Methods Used in a Study of the Vowels", JASA, 24, 175-184, 1952; /3/ Ph. CHRISTOV, "A Large Bulgarian Central Allophones Data Base", 11. ICPS, Tallinn, 1-7 August 1987 (Receipt No 0653); /4/ Ph. CHRISTOV, "A Semiautomatic Speech Sound Aural Identification Procedure with Its Application to Speech Analysis", Acustica, 29, 374-349, 1973;
/5/ Ph. CHRISTOV, "An Algorithm Using Linguistic Information and Its Application to the Analysis of Speech in the Spectral Domain", Proc. 11. ICA, 4, 161-164, 1983; /6/ B. GOLD, "Computer Program for Pitch Extraction", JASA, 34, 916-921, 1962; /7/ G. FANT, "Modern Instruments and Methods for Acoustic Studies of Speech", The Royal Inst. of Technology: Rpt.No 8/June 11, 1957, p. 14

APPENDIX

Word List: (In rough phonemic IPA transcription)

Stressed: /bìblija/, /bèbe/, /bàba/, /bəbrek/, /bòbof/, /bùba/

Unstressed: /biblèjski/, /bebètʃ.ef/, /babalək/, /bəbrekoviden/, /bobòvina/, /bubàr/

NOTICE: In Tables are count F1xF'2-labeled points positions, not actual vowels number

Table 1. Confusion matrix to Fig. 1.

	i	e	a	ə	o	u	none	total
i	20	3						23
e		20						20
a			13	4			1	15
ə			3	19				20
o				1	13			14
u					2	9	1	12

Table 2. Confusion matrix to Fig. 2.

	i	e	a	ə	o	u	none	total
i	23	6						23
e	4	25		2				25
a			22	7				24
ə		2	5	21				21
o			3	4	18			23
u					2	13	2	17

Table 3. Confusion matrix to Fig. 3.

	i	e	a	ə	o	u	none	total
i	24							24
e	1	26					1	28
a			18	1				19
ə			1	13	1		1	16
o					17	3	3	20
u					5	17		17

Table 4. Confusion matrix to Fig. 4.

	i	e	a	ə	o	u	none	total
i	25						2	27
e		24						24
a		1	20	1			1	23
ə				19			1	20
o			1	2	19	1	3	26
u					2	20	2	24

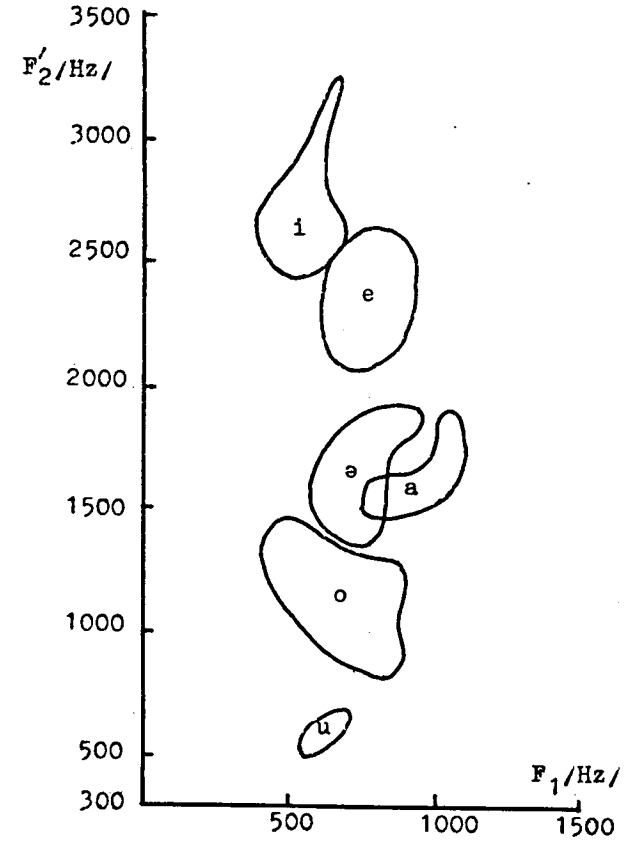


Fig. 1. First vs. second formant diagram of the Bulgarian vowels in /bVb/ context uttered by 30 female speakers

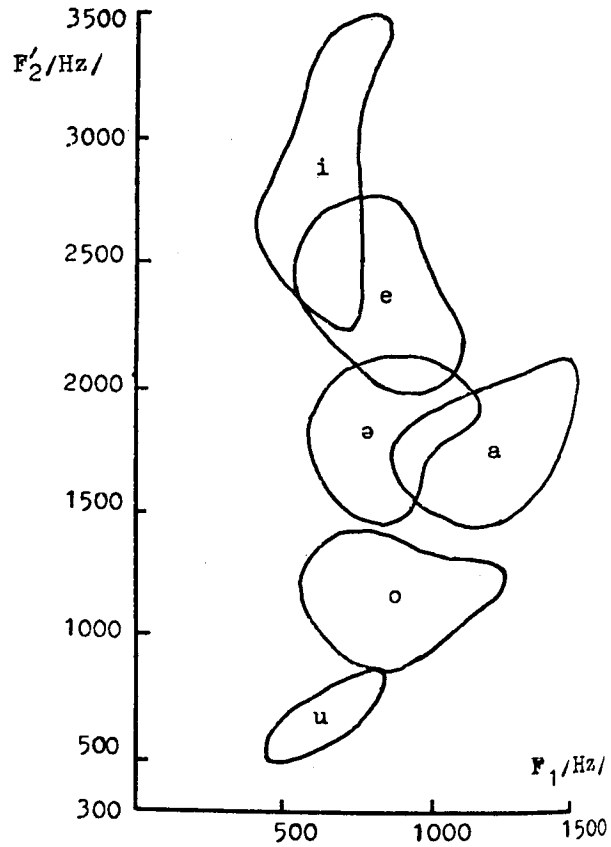


Fig. 2. First vs. second formant diagram of the Bulgarian vowels in /bV̀b/ context uttered by 30 female speakers

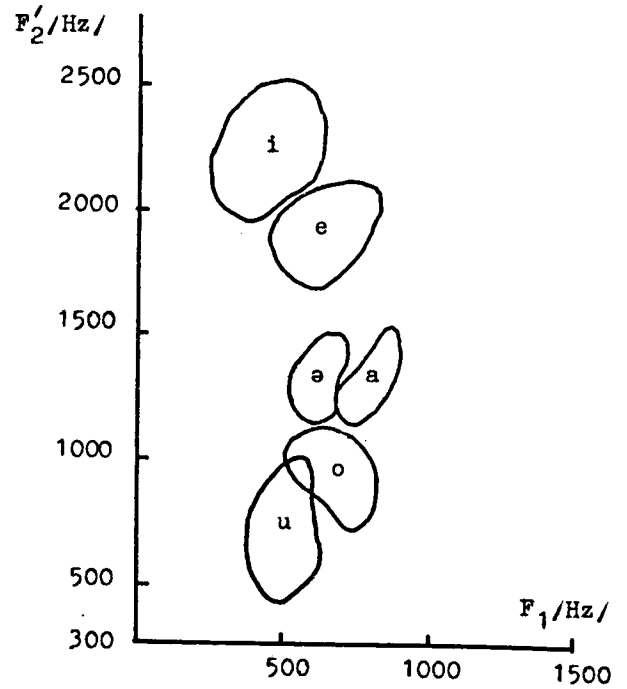


Fig. 3. First vs. second formant diagram of the Bulgarian vowels in /bVb/ context uttered by 30 male speakers

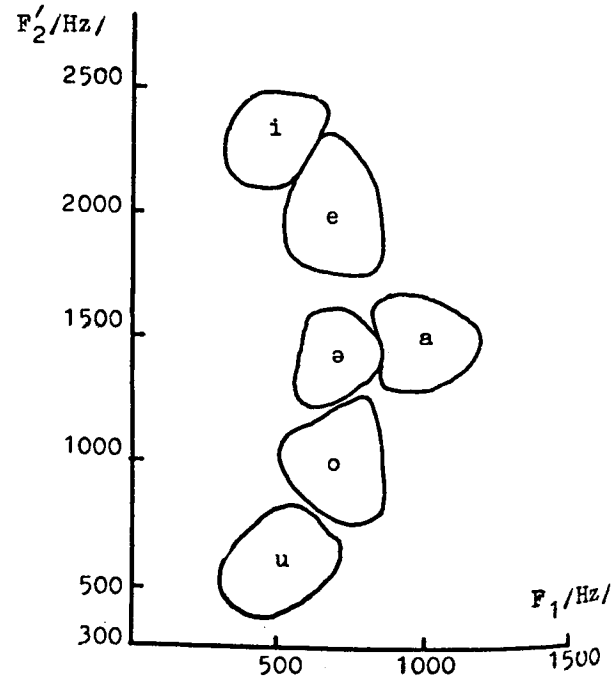


Fig. 4. First vs. second formant diagram of the Bulgarian vowels in /bV̀b/ context uttered by 30 male speakers

INFLUENCE OF SPEAKING RATE IN SPANISH DIPHTHONGS

GUILLERMO A. TOLEDO

NORMA ANTOÑANZAS-BARROSO

Dept. of Linguistics, University of California at Los Angeles
Los Angeles, CA, USA 90024

ABSTRACT

The universal diphthong production model, i.e., invariance in F2 onset frequency and in the rate of change of F2 transition in spite of differences on speaking rate, was reconsidered. Three Argentinian Spanish speakers'data: a corpus of nonsense words, CVVC and CVVCV, consisting of diphthongs /ai ia au ua eu ue oi io/ embedded in a consonantal and syllable environment: initial /t/, final /s/ and /sa/ was analyzed through LPC analysis. A control of the acoustic measurements on three speaking rates (slow, moderate and fast) was taken into account. A specially-written computer program for the measurement of F2 slope (maximum: four steepest points in 30 ms, and normal: from the onset to the offset) was performed. Results indicated significant differences in F2 rate of change. However, F2 onset frequency showed invariance in some data, and significant differences in the rest. Spanish diphthong production would appear to have language- or diphthong-dependent patterns than universal ones.

INTRODUCTION

It has been claimed that the onset frequency and the rate of change of the second formant transition in the production of English diphthongs is constant in spite of differences on speaking rates /1/. In addition, it has been suggested that the rate of change, rather than absolute F2 frequency, is the primary acoustic cue for diphthong recognition /2/.

In an acoustic study, similar results were obtained by Manrique for the fourteen Spanish diphthongs /3/. She worked on corpus of natural words where diphthongs appear in a labial-dental or dental-labial environment. Tokens were emitted under two speaking rates: moderate and fast. F2 rate of change calculation (Hz/ms) through manual spectrographic measurements was made. In so doing, Manrique also observed that the rate of change of the second formant transition shows an invariant pattern under suprasegmental changes of speaking rate. In this previous research, however, three relevant observations should be made: 1) spectrographic analysis is inappropiate for measurements under fast speaking rate: formant crushing interferes with accurate results, 2) selection of a corpus of natural speech (words in a carrier sentence) would be unsuitable because of temporal and spectral distortions caused by different consonantal environments and 3) no statistical differences among results were reported. Besides, in recent literature, this universal model for diphthong production was reanalyzed. Dolan and Mimori /4/ reported two different calculations of F2 rate of change: one, in four points of the transition (30 ms), i.e., the steepest ones, and two, the rate of change of this F2 transition throughout its total duration. They observed that changes in speech rate influenced the F2 slope in English and in Japanese diphthongs even though in Japanese the influence of speech rate was less important than in English.

The purpose of this work, then, was the observation of Spanish diphthongs under changes in speaking rate, related to the universal model previously suggested.

PROCEDURE

Spanish material consisted of nonsense words, CVVC and CVVCV, made up by diphthongs /ai ia au ua eu ue oi io/ embedded in a consonantal and syllable environment: initial /t/, final /s/ and /sa/. These frames were chosen because alveolar or dental consonantal context has less acoustic influence over transitions than the rest of consonants /4/. The corpus, then, was composed of 16 words x 3 male adult Argentinian speakers x 3 repetitions x 3 speaking rates: slow, moderate and fast. Tokens were embedded in a carrier sentence to avoid context effects. Speakers were instructed to control changes in speed through previous listening of reference speech samples emitted by the first author who accurately controlled different speaking rates. The three speakers appeared to be fit for solving the task.

Sentences were recorded on magnetic tape in a sound-proof room. Then, Ss' recorded and selected tokens were sampled on a PDP11/23 Digital computer for acoustical analysis. A 10-KHz sampling rate with a 4.5KHz low-pass filter was used. A 25.6 ms Hamming window with 10 ms intervals was utilized for analysis through linear predictive coefficient (LPC) using UCLA WAVES speech analysis system programs. Tokens were measured by a special computer program. F2 onset and offset, transition duration, maximum rate of change of transition in the steepest four points with a duration of 30 ms (henceforth: MAXSLOPE), and rate of change of complete F2

TABLE I.- MEANS AND STANDARD DEVIATIONS (IN PARENTHESIS) OF F_2 MAXSLOPE AND AVESLOPE. MEANS (Hz) AND STANDARD DEVIATIONS (Hz, IN PARENTHESIS) OF F_2 ONSET, OFFSET AND RANGE.

KEY: SLOW = S MODERATE = M FAST = F

		/ai/	/ia/	/au/	/ua/	/eu/	/ue/	/oi/	/io/
MAXSLOPE	S	10.51 (2.53)	-10.86 (1.99)	-5.20 (1.35)	6.07 (1.22)	-11.05 (8.24)	14.29 (2.87)	15.42 (3.05)	-18.90 (3.62)
	M	9.46 (0.89)	-10.46 (1.39)	-5.00 (1.11)	6.48 (1.78)	-14.44 (2.33)	14.47 (2.74)	17.13 (3.35)	-17.61 (3.27)
	F	10.62 (1.54)	-11.20 (1.39)	-5.28 (1.12)	6.10 (1.89)	-13.99 (2.59)	12.33 (8.64)	20.76 (6.02)	-17.50 (3.90)
AVESLOPE	S	6.10 (1.45)	-7.11 (1.40)	-3.52 (1.17)	4.11 (0.69)	-6.63 (4.78)	8.97 (1.29)	7.85 (1.58)	-8.47 (1.99)
	M	6.19 (0.59)	-6.65 (1.78)	-3.79 (0.77)	4.79 (0.97)	-9.20 (1.01)	9.44 (1.53)	9.07 (1.25)	-9.76 (1.52)
	F	7.08 (1.20)	-7.49 (1.97)	-4.73 (0.71)	5.02 (1.20)	-9.43 (1.88)	8.63 (5.86)	11.44 (1.75)	-10.53 (1.14)
ONSET	S	1390.30 (82.22)	2258.75 (76.37)	1302.04 (136.01)	801.13 (100.43)	1961.55 (135.31)	853.55 (143.35)	974.90 (81.68)	2194.82 (71.57)
	M	1434.03 (57.34)	2112.31 (127.15)	1254.49 (56.56)	818.42 (90.53)	1929.16 (126.98)	910.58 (142.87)	1009.25 60.07	2155.17 (87.29)
	F	1448.17 (56.50)	2085.03 (122.88)	1282.13 (46.89)	871.25 (136.94)	1847.90 (124.03)	959.29 (182.02)	1082.65 (66.55)	2118.69 (62.16)
OFFSET	S	2216.42 (95.03)	1496.33 (82.94)	908.72 (86.06)	1292.52 (63.01)	962.10 (144.05)	1935.23 (93.33)	2066.55 (126.16)	1010.96 (74.13)
	M	2122.71 (93.82)	1458.83 (79.10)	976.87 (94.61)	1215.83 (80.51)	1019.97 (136.51)	1864.23 (47.09)	2119.51 (129.88)	1109.25 (93.28)
	F	2103.65 (122.76)	1473.39 (87.32)	1030.23 (100.90)	1218.88 (59.78)	1100.23 (99.77)	1773.47 (154.39)	2032.18 (168.23)	1165.72 (75.58)
RANGE	S	826.12 (129.30)	762.42 (71.33)	393.32 (123.34)	491.39 (91.96)	999.45 (199.24)	1074.82 (229.56)	1091.65 (115.69)	1183.86 (89.56)
	M	688.68 (102.13)	653.48 (148.69)	277.62 (90.63)	397.41 (149.44)	909.19 (190.10)	953.65 (178.70)	1010.61 (154.50)	1045.92 (143.59)
	F	655.48 (89.20)	611.64 (158.24)	251.90 (69.76)	347.63 (181.61)	747.73 (150.27)	849.97 (202.90)	949.53 (205.12)	952.97 (85.38)

transition (henceforth: AVESLOPE) were calculated. Through this program a detection of transitional segments from steady-state portions was made. The starting point was indicated by differences over 15 Hz in between two consecutive windows, the preceding one with a 10 ms interval duration. Similar treatment for English (with 15 Hz acoustic difference, and for Japanese (with a 20 Hz value) was reported by Dolan and Mimori /4/.

RESULTS

Table I shows means and standard deviations of F2 rate of change on MAXSLOPE and AVESLOPE calculations, and means and standard deviations of F2 onset, offset and range. Results of ANOVA indicated significant differences among speech rates in MAXSLOPE ($F(12, 419) = 97.62$, $p > 0.0001$), and in AVESLOPE ($F(12, 419) = 76.82$, $p > 0.0001$). Similarly, ANOVA calculations showed significant differences of range among the three speaking rates ($F(12, 419) = 182.40$, $p > 0.0001$). ANOVA results also revealed that AVESLOPE and range were strongly influenced by rate: the F ratio for rate related to AVESLOPE was 34.77 (2 and 419 df and $p > 0.0001$), and the F ratio for rate connected with range was 76.22 (2 and 419 df and $p > 0.0001$). However, ANOVA calculations yielded a less significant influence of rate in MAXSLOPE ($F(2, 419) = 2.83$, $p > 0.0599$).

Table II shows statistical differences (Student's t test) between F2 onset frequencies influenced by different speaking rates. In the same manner the table displays statistical differences between F2 offset frequencies. Inspection of the table reveals

TABLE II.- STATISTICAL DIFFERENCES (t-test) ON F_2 ONSET AND F_2 OFFSET INFLUENCED BY CHANGES IN SPEAKING RATE

KEY: ns = NOT SIGNIFICANT x = $p < 0.0001$ xx = $p < 0.002$ xxx = $p < 0.05$
S = SLOW M = MODERATE F = FAST

	F_2 ONSET							F_2 OFFSET						
	/ai/	/ia/	/au/	/ua/	/eu/	/ue/	/oi/	/io/	/ai/	/ia/	/au/	/ua/	/eu/	/ue/	/oi/	/io/
F vs S	ns	xxx	ns	ns	ns	ns	xxx	xxx	xx	ns	xx	ns	ns	xxx	ns	ns
F vs M	ns	ns	ns	ns	x	xxx	xx	xx	ns	ns	ns	ns	xx	ns	ns	ns
S vs M	ns	ns	ns	ns	ns	ns	ns	ns	xx	ns	xx	ns	ns	ns	ns	ns

that only F2 onset in diphthongs /ai au ua/ remained invariant and uninfluenced by speaking rate. The reverse was true in diphthongs /ia eu ue oi io/. Calculations on F2 offset frequencies resulted in an invariant trend in diphthongs /ia ua oi io/ and, on the contrary, resulted in significant differences in diphthongs /ai au eu ue/.

Correlation coefficients (Fig. 1 and Fig. 2), on the other hand, showed that MAXSLOPE was highly

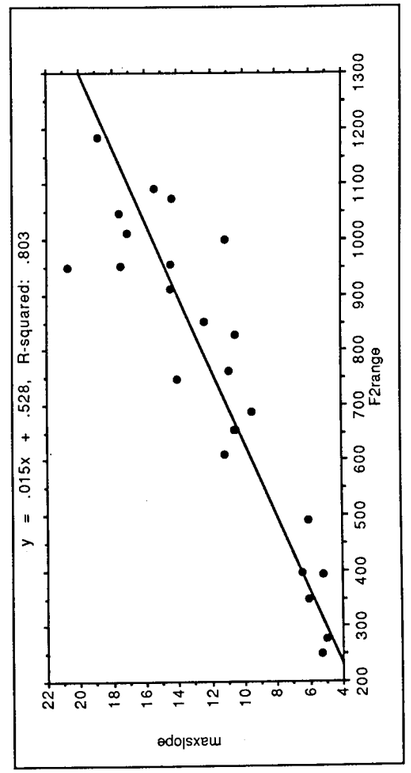


FIGURE 1. CORRELATION COEFFICIENTS BETWEEN MAXSLOPE AND F2 RANGE

correlated with range: R= 0.803, $p > 0.0001$, but poorly correlated with transition duration: R= 0.113, $p > 0.0183$. AVESLOPE showed a similar trend: a high degree of correlation with range: R= 0.654, $p > 0.0001$, but a low one with transition duration: R= 0.119, $p > 0.0001$.

DISCUSSION

ANOVA results obtained from Spanish data due weight to reject Gay's model /1/: changes in speech rates resulted in significant statistical differences in F2 rate of change. Similar conclusions were reported by Dolan and Mimori for English and Japanese complex vowels /4/. In addition, F2 onset frequencies showed significant variations in diphthongs /ia eu ue oi io/ because of changes in speaking rate.

According to Gay's model, no correlation should appear between slope and either range and transition duration if the F2 rate of change remain invariant and uninfluenced by different speech rates. Correlation coefficients in Spanish data reject this notion, at least in one aspect: the high degree of correlation between slope and range.

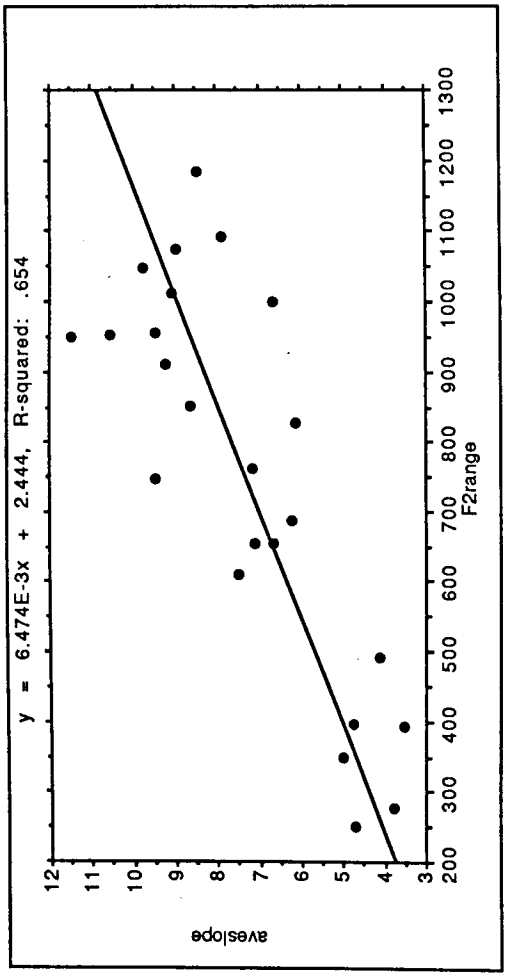


FIGURE 2. CORRELATION COEFFICIENTS BETWEEN AVESLOPE AND F2 RANGE

In brief, the production of diphthongs in Argentinian Spanish shows a language-specific trend rather than an universal-specific one previously suggested by Gay /1/ and by Manrique /4/. This position is supported by recent research on diphthongs in different languages: English and Japanese /4/, Hausa /5/, Chinese /6/, and -in a cross-phonetic study- Hausa, Arabic and Chinese /7/.

REFERENCES

/1/ Gay, T. (1968). Effect of speaking rate on diphthong formant movements, The Journal of the Acoustical Society of America 44, 6, pp. 1570-1573.

/2/ Gay, T. (1970). A perceptual study of American English diphthongs, Language and Speech 13, 2, pp. 65-88.

/3/ Manrique, A.M.B. (1979). Acoustic analysis of the Spanish diphthongs, Phonetica 36, 3, pp. 194-206.

/4/ Dolan, W. and Mimori, Y. (1986). Rate-dependent variability in English and Japanese complex vowel F2 transitions, The Journal of the Acoustical Society of America 80, Suppl.1, S96.

/5/ Lindau-Webb, M. (1985). Hausa vowels and diphthongs, Studies in African Linguistics 16, 2, pp. 161-182.

/6/ Ren, H. (1986). A truncation model for the F2 trajectory in syllables with complex vocalic components, The Journal of the Acoustical Society of America 80, Suppl.1, S98.

/7/ Lindau, M., Norlin, K. and Svantesson, J. (1985). Cross-linguistic differences in diphthongs, UCLA Working Papers in Phonetics 61, pp. 40-44.

ACKNOWLEDGEMENTS

This work has been supported by a grant awarded to the first author, Senior Scholar Fulbright Program: Advanced Research Award in the United States (N°85-08139), by the Council for the International Exchange of Scholars (CIES) affiliated with the American Council on Education. This research has been also supported by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina through a salary appointment to the first author at the Laboratorio de Investigaciones Sensoriales, Buenos Aires, Argentina.

We are particularly indebted to Peter Ladefoged, Mona Lindau, Ian Maddieson and William Dolan for fruitful comments, guide and inspirational help.

Some micro-effects of tempo change on Timing in French

Janet Fletcher

The Centre for Speech Technology Research, University of Edinburgh

Abstract

A number of factors influence the temporal organization of connected speech. One such factor, tempo change, is an important variable whose effects on segment duration may highlight important language-specific and language universal features of speech timing. When French speakers are asked to read a passage at increasing tempo from slow to normal to fast, they modify their pause behaviour substantially. In addition, consonant and vowel segment durations are compressed, but not to the same degree. Whilst the overall levels of temporal compression are lower for vowels than the levels of compression reported for other languages, they are still somewhat higher than previously reported for French. A probable reason for the former is the smaller degree of vowel reduction that occurs in French under conditions of increased tempo. This factor along with other compression effects such as syllable, consonant and vowel elision are considered in the light of traditional timing typologies./fR

Introduction

The durational structure of a stretch of connected speech is determined by a number of factors. Speech tempo is one such factor that has received some attention in the experimental phonetics literature. However like the majority of instrumental studies of speech in general, past research has concentrated on a limited sample of languages, most notably English. In addition, the type of corpora analysed have usually been restricted to short controlled utterances. Few researchers have looked at the effects of tempo change over a large stretch of connected speech. Consequently, this study constituted an attempt to examine in detail the effects of tempo change on one feature of timing organization, namely, segment duration in a long stretch of connected speech. These effects can be described as the micro-effects of tempo change. Furthermore, the language under investigation was French. We wished to examine any possible strategies at the segmental level to modify tempo in spoken French, that differed from those reported for other languages.

Micro-effects of tempo change

Previous studies of tempo change have found that the greatest quantitive effect of increasing or reducing tempo is to do with reducing or increasing the amount of pause time relative to articulation time. Smith (1976) and Duez(1983) found this to be the case in French. However there was also evidence of changes in articualtion rate between tempo that can be explained by temporal changes at a micro-level.

A number of authors have investigated the effects of tempo change on the acoustic duration of segments. However, most of these studies have concentrated on corpora composed of isolated sentences and not of larger tracts of connected speech. Whilst the first kind of study provides us with some idea of the effects tempo change on certain categories of acoustic segments, we need to extend these findings to a broader range of speech situations. We are aware of only one controlled study of micro-effects of tempo variation in a passage of spoken French(Smith 1976) Consequently we will report on the findings of some studies of tempo change that have used single utterance-based corpora, in addition to those studies based on larger stretches of read English, Swedish and Dutch.

Gay(1978) and Port(1984) for American English, Lindblom(1963) for Swedish, and Nooteboom and Shs(1972) for Dutch, found that overall decreases in syllable duration during rapidly delivered speech were reflected primarily in the durations of the constituent vowels, and only secondarily in the durations of the constituent consonants. Port(1984) for example, provided an explanation for this based on universal mechano- inertia effects of increasingly rapid articulation. Consonant articulations are generally more complex than vowel articulations, in so far as they imply the attainment of a specific constriction, or closure at some point, or points in the vocal tract. An increase in tempo from normal to fast. results in consonant and vowel articulations following each other in quicker succession. Consonant gestures needing more execution time are therefore maintained, or even strengthened at the expense of articulatory gestures associated with vowel production. Consequently vowels tend to reduce in duration and quality, particularly if unstressed.

Although this may be the case for the languages cited above, it is not absolutely certain to what extent this could be classified as a universal of articulatory behaviour. Gay also stressed that the coordination of articulatory movements may be adjusted in some way to preserve the information-bearing elements across changes in tempo. For example, stressed vowel gestures may be maintained at the expense of unstressed vowel gestures. It is also possible that vowel gestures are only weakened if the phonology of the language allows it to happen. Certain languages tend to preserve vowel quality in unaccented or usrtressed position to a greater extent than others. As a loss of vowel quality is usually associated with durational shortening (Lindblom 1963), one has to consider to what extent phonology and mechanical inertia of the articulators interact under conditions of rapid articulation rate. Thus it seems likely that we need to put forward

language-specific rules of temporal organization that can account for physiological limitations, whilst maintaining "phonologically invariant temporal relations" (Port1974 p,272).

It is generally accepted that vowels in spoken French tend to preserve their distinctive quality to a much greater extent than vowels in languages such as English, under varying conditions of stress or accent. Delattre's (1968) comparative study of vowel reduction in French, German, English, and Spanish, describes in detail the tendency of English vowels to be modified acoustically and articulatorily, approaching more "centralised" qualities in unstressed position. French vowels on the other hand are modified vary slightly in comparison. This is often cited as a reason why spoken French gives the auditory impression of syllable timed whereas English sounds stressed timed. In spoken English stressed vowels may often be the only ones to sound fully distinct.

It is possible that given conditions of increased tempo, as well as presence or absence of accent, vowels in a French utterance may shorten until a certain limit is reached beyond which further shortening would result in a loss of informative acoustic cues to vowel identity, unless there was specific articulatory modification to preserve quality. This seems unlikely in view of the tendency of the speech production system to prefer hypo-articulation as opposed to hyper-articulation(Lindblom 1973). One might, nevertheless expect a smaller degree of vowel compression in French spoken at rapid tempo, than in other languages, like Dutch, Swedish or English, whose unstressed vowels reduce to short schwa-like segments. Furthermore, there may be a more equal degree of consonant and vowel duration compression in French than in these languages. Smith(1976)found this to be the case in her study. However the limitations of her corpora(data from one speaker) make it necessary to see whether this finding can be replicated in a much larger corpus of French.

Peterson and Lehiste(1960) and Port(1980) also investigated the interaction between tempo, and other duration influencing factors such as presence or absence of stress. Stressed syllables in many languages, including English, Dutch, and Swedish are generally much longer than unstressed syllables. Similarly accented syllables in French are also reportedly longer than unaccented syllables (Delattre 1968, Wenk and Wioland 1982). Peterson and Lehiste(1960) and Port(1960) found that stressed syllables in their American English data were less affected by rapid tempo increase than unstressed syllables. This is not surprising if it is interpreted in a stress-timing framework. If the basic "rhythm" of English is mainly governed by the alternation of stressed and unstressed syllables, it seems likely that unstressed syllables would be compressed to a greater extent than stressed syllables in order to maintain the rhythm in the light of increasing "articulatory demands" particularly if stressed syllables are somehow "more important" than unstressed syllables either from the articulatory, or perceptual point of view.

The interaction of tempo and accent in French has not to our knowledge been examined in any detail, so we do not know whether French behaves in similar ways to English. Proponents of the syllable- timing label of French would claim that there should be no difference at all given that this typology infers that accented syllables in French have no important role at the level of temporal organization. The basic unit of timing is said to be the syllable, whether accented or not. Consequently any re-organization of the articulatory plan due to increased articualtion rate should be evenly distributed across all syllables, and by extension their constituent segments. Thus the effects of tempo change on accented and unaccented vowel duration should be similar.

A handful of studies have examined the micro-effects of slowing down tempo in as much detail as speeding up tempo. There is a certain amount of consensus in the timing literature that slowing down and speeding up tempo are not converse processes. Port(1981) found that when subjects slowed down their neutral tempo, all consonant and vowel segments increased in duration by a constant ratio. Pickett(1980) reported similar findings. Smith (1976) reported a similar results for her French data; remembering also that she found no differences in the way consonant and vowel segment durations change as a function of tempo increase from normal to fast. Once again, the limitations of her experimental corpus make it necessary to test whether this finding can be replicated in a much larger data sample.

Experimental procedure

The following experiment was designed to investigate the above micro effects of tempo change in an extended corpus of spoken French. We predicted that there would be modifications in articulation rate between tempi. Furthermore, We wished in part, to test Smith's(1976) claim that there is no major difference in the degrees of consonant and vowel duration change between tempi (i.e. consonants and vowels compress or expand to a similar extent), due in part to the so-called syllable timing nature of French. On the basis of potential "braking" effects due to the tendency in French to maintain distinctive vowel quality in unaccented position, we hypothesized that the degree of consonant and vowel compression would not be as great between normal and fast tempo as the levels of consonant and vowel compression between normal and slow. We also hypothesized that if a strict syllable-timing description is relevant for French there should be no difference between the degree of shortening of accented and unaccented vowels from slow through to fast tempi.

Subjects and Materials

A transcript of a French radio interview was found that was not too long(112 words), that consisted of colloquial syntax and morphology, conceivably making it easy to read at either slow or fast tempo. Six native speakers of non-meridional French acted as informants in this experiment. All had been educated to university level. They were asked to read the text three times, once at their normal tempo, then at a slower rate, still maintaining natural delivery, and finally at a speeded up version of their normal tempo, likewise maintaining naturalness.

Analysis procedure

The informants' performance of this task was recorded under laboratory conditions on a FM Recorder (Racal Store 4D) at a speed of 30 ips. Oscillograms were made from these recordings using a Siemens-Elema Mingocard 4 at a paper speed of 25mm/s. The recordings had been slowed down by a factor of 4. This technique enables a more accurate segmentation of the audio trace into consonant and vowel segments.

Potentially informative changes in amplitude associated with the presence of nasal and lateral articulations, or intervals of aperiodic noise associated with frication are more readily identifiable. Certain segments such as semi-vowels were not easily identified on the waveform. These were analysed spectrographically, using a Kay Sona-graph. Duration measurements were obtained for intervals corresponding to consonant and vowel segments, following the procedures described by Peterson and Lehiste(1960) and Delattre(1965). Any intervals on the acoustic trace of 200ms and over, showing no signs of periodic or aperiodic noise above the level asociated with background noise were labelled pauses, following the procedure of Duez(1983). Syllables were identified according to phonolgical criteria, and their durations calculated by adding the durations of the constituant segments. In the case of elided segments or syllables, the segment or syllable were given the durational value 0.

The rate of articulation was measured according to the procedures oulined in Grosjean and Deschamps(1975), Butcher(1981),and Dauer(1983, and represented the number of syllables unttered per second of total articulation time. Durational data obtained from consonant and vowel measurements were tabulated according to tempo and speaker, and the means, standard deviations computed of segment duration (consonant and vowels combined), vowel and consonant duration independently, irrespective of accentuation and finally accented vowels and non-accented vowels. Withregards to the latter only vowel segments that occurred in all three readings for each speaker were included in the analysis. Analyses of variance were performed on segment duration in general, then consonant and vowel durations under the conditions of tempo change, and presence or absence of accent.

Results and Discussion

Tables 1-2 illustrate the results of these computations for each speaker. There seemed to be a consistant trend across all but one speaker to increase articulation rate from slow through to fast tempo. It was not evident from our results that articulation rate change was consistantly greater from normal to fast tempo than from normal to slow, as found by Butcher(1981) for German . However there was a significant degree of inter-speaker variability in determining each speakers fast, normal and slow tempi, as one would expect. Speaker 5 did not vary her articulation rate at all between normal and fast tempo, but reduced it by 35% between normal and slow.

We assumed that the above changes in articulation rate due to speeding up or slowing down tempo reflected a certain amount of segment compression and expansion. An examination of the means, standard deviations and results of the analyses of variance showed that segment durations were significantly affected by tempo for all speakers. There was a significant effect of Tempo across segments and subjects, although there was a certain amount of inter-speaker variability ($F = 111.01$, $P<.0001$). There was also a significant tempo/segment interacation ($F = 10.77$, $P<.0001$). This indicated that consonant and vowel segment durations were not interacting with tempo in the same way. An examination of the means and standard deviations of consonant and vowel duration highlight this difference. Vowel segment durations were compressed more than consonant segment dduration due to tempo increase. Thus our results do not replicate those of Smith(1976), and our French data seem to reflect similar patterns in this respect to the English data in Peterson and Lehiste(1960) and the Dutch data in Nooteboom and Slis (1972).

Speakers and	Tempo	TST(ms)	TAT(ms)	AT/TST	SR	AR	MDP	NP
1	*Fast*	38159	31528	83%	4.38	5.28	469(248)	12
	Normal	44611	33834	76%	3.76	5	790(542)	12
	Slow	55803	39020	70%	3.2	4.6	646(422)	21
2	*Fast*	26329	24673	94%	6	6.4	440(226)	4
	Normal	36082	29896	83%	4.65	5.6	478(368)	8
	Slow	47883	38818	77%	3.51	4.3	480(320)	16
3	*Fast*	30670	26900	88%	5.7	6.5	342(129)	8
	Normal	33471	29008	87%	5.2	6	372(195)	11
	Slow	40129	33573	84%	4.3	5.2	312(182)	12
4	*Fast*	30519	25959	85%	5.5	6.5	455(142)	10
	Normal	39684	32040	81%	4.3	5.3	507(241)	15
	Slow	57036	43033	75%	3	4	610(310)	23
5	*Fast*	36785	30270	82%	4.5	5.4	650(340)	10
	Normal	44138	30120	68%	3.7	5.4	825(554)	17
	Slow	87667	42000	62%	2.4	3.9	723(478)	25
6	*Fast*	25057	23387	94%	6.7	7.2	417(161)	4
	Normal	31781	28706	91%	5.3	5.9	514(173)	6
	Slow	38590	32994	84%	4.4	5.1	573(239)	11

Table 1. TST: total speaking time (ms); TAT: total articulation time (ms); AT/TST: the percentage articulation time to total speaking time; SR: speaking rate, number of syllables per second; AR: articulation rate (syllables per second); MDP: mean pause duration in ms (standard deviations in brackets); NP: number of pauses

Speaker	Tempo	Consonant (ms)	durations	Vowel (ms)	durations	Combined (ms)	segment
		mean	**standard deviation**	**mean**	**standard deviation**	**mean**	**standard deviation**
1	*slow*	92.18	(42.85)	102.47	(71.21)	96.92	(55.41)
	normal	81	(35.68)	91.16	(59.85)	85.68	(48.17)
	fast	72.7	(44.36)	75.72	(58.06)	74.09	(46.43)
2	*slow*	82.16	(38.99)	90.29	(49.33)	85.91	(34.84)
	normal	74.03	(36.04)	79.8	(52.94)	76.68	(36.00)
	fast	69.49	(34.56)	70.12	(38.71)	69.78	(33.75)
3	*slow*	84.08	(43.08)	93.05	(68.27)	88.22	(48.08)
	normal	78.43	(35.65)	76.61	(48.46)	77.59	(41.38)
	fast	69.53	(39.29)	62.80	(47.18)	66.43	(35.68)
4	*slow*	82.08	(33.07)	103.76	(84.54)	92.07	(55.08)
	normal	76.91	(49.45)	81.78	(60.09)	79.16	(46.61)
	fast	64.62	(25.94)	72.39	(61.7)	68.21	(35.66)
5	*slow*	90.73	(42.56)	108.18	(67.09)	98.77	(52.78)
	normal	81.9	(44.31)	76.84	(42.43)	79.62	(40.02)
	fast	83.14	(72.61)	74.18	(40.57)	79.01	(38.14)
6	*slow*	87.51	(53.64)	88.22	(62.14)	87.84	(50.29)
	normal	72.47	(36.85)	69.82	(41.14)	71.24	(38.47)
	fast	68.61	(32.67)	59.84	(30.74)	64.57	(33.26)

Table 2. Means and standard deviation values of consonant and vowel segment durations across tempi, and speaker.

The above result could be explained by the comparatively low values of articulation rate computed for these data. On the other hand Port's(1981) explanation of similar results in his data could account for our findings. That is, the similar degrees of compression from slow to normal and normal to fast could be explained by an interaction between physiological, mechanical needs of increasingly rapid articulation being offset by the need to maintain more or less invariant temporal relations determined by the phonology, or indeed rhythm, of the language in question.

Our results do not rule out the possibility that the need to maintain distinctive vowel quality in the fast readings could have been responsible for the paradigmatic relation between consonant and vowel duration compression between slow and normal, and normal and fast tempo. Although the proportion of vowel compression to consonant compression was greater from slow through to fast tempo, the fact that this proportion did not vary significantly between tempi seems at odds with results reported for English, for example. Pickett(1982) found that consonants and vowels in his data were expanded by 33% from normal to slow tempo, but the ratio of vowel compression to consonant compression was much greater from normal to fast tempo. however further tests need to be carried out on our data before we can draw any further conclusions.

Comparing the effects of accent and tempo, there was once again a significant degree of inter-speaker varibility withregards to segment duration ($F=13$; $P<0.0000$). The effects of accent were also strongly significant ($F=243$;$P<0.0000$). The interaction between accent, tempo and subject was extremely pronounced ($F=3.59$;$P<0.0001$). Thus vowel duration in all of the readings was substantially affected by accent and tempo.

A further analysis of variance and covariance was performed on the vowel duration data to test whether there were differences in the way accented and accented vowels were affected by tempo change. This proved to be the case. Tempo affected stressed and unstressed vowels in different ways (F=12.33; P>0006). However unlike the results of Peterson and Lehiste(1963) for American English, and Nooteboom and Slis(1972) for Dutch, tempo change affected accented vowel duration more substantially than unaccented vowel duration. Moreover, there were significant differences in the degree of duration change between slow and normal tempo and normal and fast tempo. The mean differences for accented vowel duration between slow and normal tempo were of the degree of 37%, whereas between fast and normal tempo they measured 16%. However the values for unaccented vowels, 11.5% and 10.95% between slow and normal tempo and fast and normal tempo, were not only a good deal smaller than for accented vowels, but there was no significant difference between tempi.

Speaker	Tempo	Accented Vowel Duration (ms)		Unaccented Vowel Duration (ms)	
		mean	standard deviation	mean	standard deviation
1	*slow*	172.23	73.95	78.21	26.45
	normal	142.18	70.04	66.47	20.80
	fast	129.3	89.43	56.32	17.11
2	*slow*	146.97	45.04	70.61	19.4
	normal	128.33	49.38	62.09	19.67
	fast	120.25	45.3	54.68	18.9
3	*slow*	170.41	82.58	67.87	28.10
	normal	131.66	50.04	55.05	21.67
	fast	103.46	44.02	45.42	14.61
4	*slow*	182.95	84.28	69.31	19.08
	normal	142.18	59.2	58.49	17.87
	fast	120.38	51.36	51.74	15.67
5	*slow*	175.8	69.02	81.1	22.58
	normal	134.64	48.76	66.56	19.54
	fast	120.66	40.33	59.16	19.48
6	*slow*	164.36	76.09	67.36	18.85
	normal	112.18	47.49	56.63	18.04
	fast	97.25	35.18	45.89	15.85

Table 3: Means and standard deviation values of accented and unaccented vowel duration across tempi and speakers.

There are a number of possible reasons why this pattern emerges in our results. The mean durations for unaccented vowels are a great deal shorter than the mean duration of accented vowels, in general. Consequently the smaller tempo effects on unaccented vowel duration could be due to a compressibility limit operating to prevent over-shortening that may lead to loss of distinctive vowel quality. Furthermore, Smith (1976) found that longer segments in her data(whether consonants or vowels) tended to be shortened a great deal more than intrinsically short segments. However, we have not at this stage taken into account the interaction between phonological or intrinsic consonant or vowel length, tempo and accent. A further analysis of the data bearing this factor in mind is therefore necessary.

More significantly, the large difference between accented vowel duration in the slow and normal readings is more than likely concomitant with the reduction in the number of pauses between the three readings(see table 1). Given that accented syllables in French are largely group final, they are often pre-pausal, and subsequently are longer than non-pre-pausal syllables(Smith 1976). Therefore a reduction in the number of pauses between readings necessarily reduces the number of pre-pausal syllables. we are in the aprocess of investigating this possibility.

Smith(1976) claims that the only truly "long" segments in French are those belonging to pre-pausal syllables. But as accent and the possible associated durational properties were not taken into account in her study, it remains unclear if or how non-pre-pausal accented syllables interact differently with tempo than pre-pausal accented syllables. On the basis of our finding that the margin of difference between the amounts of shortening of accented and unaccented vowels was substantially reduced between normal and fast tempo as opposed to slow and normal tempo, we could hypothesise that if pre-pausal vowels were excluded from the analysis the margin of difference would disappear.

Finally, we did not take into account phenomena such as vowel and syllable elision reported to occur as a result of increasing tempo from normal to fast (Smith 1976). Comparative instrumental studies of the micro-effects of tempo increase in French and English are non-existant. Consequently we do not know for sure whether French relies more on syllable and segment elision than segment compression to increase tempo. It is also uncertain whether the number of accented syllables is reduced to counteract overshortening of individual segments. If both of the above possibilities are in fact the case, this could explain why our results differ from those reported for other languages. Although the number of accented syllables in our data did vary, there was some indication that there was some kind of resistence to eliminate too many. A strict syllable-timing hypothesis would predict a reduction in the number of accented syllables. It seems that stressed syllables in English almost never disappear under conditions of rapid articulation rate associated with tempo increase. Once again we need to examine the data in more detail before we can make any further assumptions about the differences between the two languages, and whether we can involve the syllable-timing and stress-timing dichotomy.

Conclusion

It certainly appears to be the case that there are some language- independent ways in which the acoustic durations of a stretch of connected speech reflect tempo change, and in particular tempo increase. The patterns of segment compression in French due to tempo increase are more complex than reported previously. Like English, Dutch or Swedish, vowel duration is more strongly affected by changes in speech rate associated with tempo change, than consonant duration. However it seems that at this stage there may be strong differences between the degree of duration change experienced by different languages.

The situation is much less clearcut when it concerns the interaction between vowel duration, positional and accentual variables, and tempo change. So little experimental data exists to show how tempo interacts with the above duration influencing variables in any language. It seems reasonable at this stage to acknowledge that French accented syllables do not behave in the same way as English or Dutch stressed syllables. Whether this is due to the so-called syllable/ stress timing dichotomy is not absolutely clear. There are definite differences in the way accented and unaccented vowels respond to tempo change in French. A simplistic typology label such as syllable-timing cannot account for these differences.

However, it seems likely that any differences between French and the above-mentionned languages is due to not so much the behaviour of accented vowels, but unaccented vowels. That is to say the degree of difference is not so much due to accented vowels not resisting durational compression unlike unaccented or unstressed vowels in Dutch or English. The reverse seems to operate in our corpora of French. The smaller degrees of temporal compression of unaccented vowels reflect a difference between French and these other languages. The reasons for this compression resistance include the tendency to maintain intrinsic vowel quality in unaccented syllables in spoken French. Further analyses of the data along the guidelines suggested in the above sections are currently being undertaken. Nevertheless we may interpret our present results as evidence of the trade-off proposed by Port (1981) between the mechano-inertial demands of rapid articulation as part of the strategies to increase speaking tempo and phonological conditioning.

References

Butcher, A.R. (1981): Aspects of the speech pause: phonetic correlates and communicative functions. Phd Thesis, Universitaet Keil. Also as: Arbeitsberichte des Instituts fuer Phonetik der Universitaet Kiel 15.

Delattre, P. (1968): An acoustic and articulatory study of vowel reduction in four languages. IRAL VII/4, 295-325.

Gay, T. (1978): Effect of speaking rate on vowel formant movements. JAcoustSocAm 63, 223-230.

Lindblom, B. (1963): Spectrographic study of Vowel reduction. JAcoustSocAm 5, 1773-1781.

Lindblom, B. (1983): Economy of speech gestures. In Macneilage,P.F.(ed.) The Production of Speech. New York, Heidelberg, Berlin: Springer-Verlag.

Miller, J.L., Grosjean, F. and Lomanto, C. (1984): Articulation rate and its variability in spontaneous speech: a reanalysis and some inplications. Phonetica 41, 215-225.

Nooteboom, S.G. (1972): The interaction of some intra-syllabic and extra- syllabic factors acting on syllable nucleus durations. IPO AnnProgRep 7, 30-39

Peterson, G.E. and Lehiste, I. (1960): Duration of syllabic nuclei in English. JAcoustSocAm 32, 693-703.

Pickett, J.M. (1982): The Sounds of Speech Communicatioon: A Primer of Acoustic Phonetics and Speecn Perception. Baltimore: University Park Press.

Port, R.F. (1981): Linguistic timing factors in Combination. JAcoustSocAm 69, 262-274.

Smith, A. (1976): The timing of French, with reflections on syllable timing. Work in Prog Dept Ling Edinburgh Univ 9, 97-108.

The Timing of Voicing In British English Consonant Clusters as a Function of Medial Boundary Status

Gerard J. Docherty

Linguistics Department and

Centre for Speech Technology Research,

Edinburgh University, Scotland

An experiment was performed to examine the timing of voicing during CC sequences, in which C_1 is a stop, and C_2 is either a liquid or a glide. These sequences were studied with different types of medial boundary ranging from a segmental boundary, to a clause boundary. The frequency of ocurrence and the degree of progressive assimilation of voicelessness were measured as a function of boundary type. Progressive assimilation of voicelessness from C_1 to C_2 was found to occur consistently when the cluster was in syllable-internal position, but less frequently when there was a deeper medial boundary. The type of boundary had no significant effect on the degree of assimilation.

1. Introduction

The work reported below is part of a study of the timing of voicing in southern British English (SBE) obstruents. Despite the work which has been carried out on voice onset time, and on physiological aspects of voicing, there are still a number of gaps in our knowledge regarding the way in which voicing is realised, and its relation to underlying phonological categories. This study has three main strands. Firstly it provides descriptive data on the way in which voicing is timed in relation to the supralaryngeal events marking the onset and offset of obstruents in SBE. This will allow experimental evaluation of claims which have been made via auditory analysis with regard to the timing of voicing in SBE. Secondly, the investigation will contribute to the debate surrounding the status of fine phonetic detail within models of vocal performance. An increasing number of reports show that many aspects of the fine detail of articulatory coordination are neither universal or automatic. Therefore, it is difficult to attribute them to mechano-inertial performance constraints. Attempts to provide alternative explanations are hampered by the fact that many models of speech production do not have the means of incorporating such low level controllable behaviour [2,9]. Experimental evidence on the timing of voicing in obstruents will be looked at in this light. Thirdly, the increasing need for high quality speech synthesis in advanced speech output applications has lead to the requirement for very detailed phonetic knowledge bases to be available, incorporating comprehensive detail of systematic segmental and subsegmental variation in a language or accent [6]. The data gathered in this experiment will be exploited in this way in order to improve the allophonic rule base for a SBE text-to-speech system.

An earlier experiment [4,5] has shown that in SBE, voice onset time and the occurrence of voicing during obstruents are affected by the phonological category of the segment concerned, and by the nature of the phonetic environment within which it is embedded. The experiment reported below was designed to investigate whether the timing of voicing in relation to a stop closure was affected by the status of the boundary between the stop and a following sonorant, and to test the strong claims that are made in the auditory phonetic literature about SBE, namely that sonorants which follow voiceless stops will only be devoiced if the sequence occurs word internally, or across a sequences of words which 'forms a close-knit entity – a phrasal word or a rhythmic group' [8].

2. Experimental Procedure

The sequences investigated in this experiment were as follows (sequences commencing with a phonemically VOICELESS segment, will be referred to as VOICELESS sequences, those beginning with a phonemically VOICED segment are referred to as VOICED sequences);

VOICELESS Sequences	VOICED Sequences
/pr/ /pl/	/br/ /bl/
/tr/ /tw/	/dr/ /dw/
/kr/ /kl/ /kw/	/gr/ /gl/

These sequences were examined under six different conditions.

1. CCV word initial position (i.e. syllable-internally).
2. C$C across a syllable boundary which is also a morpheme boundary
3. C$C across a syllable boundary which is also a morpheme and compound boundary (this may or may not involve a word-boundary due to the fact that compound nouns are not consistent in. their structure).
4. C#C across a word boundary phrase-internally (between an adjective and noun)
5. C#C across a word boundary phrase-internally (between the subject of a verb and the verb)
6. C#C Across a word boundary which is also a clause boundary.

In conditions 1, 3, 4, 5, and 6, all the initial clusters occur before a stressed syllable. In condition 2, it was not possible to maintain this control throughout the data due to the large number of unstressed affixes which were used in order to produce the correct sequences.

In addition, tokens of [sm] and [sn] occurring under all the six conditions were obtained from all the subjects. The purpose of including these sequences was to investigate whether the delay in onset of voicing which occurs between the fricative and nasal in these sequences was affected by the different boundary conditions in the same way as the delay in voicing onset in the [stop]-[sonorant] sequences.

The fifteen sequences were read in random order twice each in the six different experimental conditions by seven subjects. The subjects were all male speakers of SBE, aged between 18 and 25. The speech waveform was recorded on channel one of a REVOX A77 tape recorder and the output of a throat microphone (FJ Electronics) attached to the neck at the level of the thyroid cartilage was recorded on channel 2. The signals were digitised and stored on VAX 11/750 computer. The two signals were then aligned (using the ILS signal processing package), and manually controlled cursors were used for segmentation and durational measurements.

The criteria used for extracting the measurements from each token were as follows. The amount of devoicing of the sonorant was measured from the point at which the stop closure was released (visible as a sudden burst of noise in the time-waveform) to the first peak indicating periodicity in the throat microphone signal. It is assumed that at the moment at which the stop is released, the vocal tract would have adopted the configuration required for the following sonorant, therefore the delay in voice onset corresponds to an interval of devoiced sonorant. In the cases in which devoicing was not observed to occur, speakers used a range of different strategies. These were identified as follows. The replacement of the stop by a glottal stop, and glottalisation of the stop were detected auditorily. A pause was taken to have occurred if there was an interval of silence between the release of the stop closure and the onset of voicing for the sonorant. An incomplete stop closure was identified by a continuous noisy signal throughout the period in which the stop was being produced (i.e. there was no silence to mark a complete closure). [1]

Data analysis consisted of two main tests. A CHI^2 test was used to examine whether devoicing occurs with significantly greater frequency in any of the boundary conditions. Analysis of variance was used to see if the degree of devoicing was significantly affected by the boundary condition.

[1]The problems encountered in segmentation, measurement, and instrumentation are discussed in detail in [5]

3. Results

The measurements presented describe the FREQUENCY with which devoicing or partial devoicing of the sonorant took places, and the DEGREE to which it was devoiced.

3.1. Frequency of Occurrence of Progressive Devoicing

The results show that progressive devoicing of sonorants in a stop-sonorant environment occurs with varying frequency according to the status of the boundary which is present between the two components of the CC sequence. Figure 1 shows the number of cases of VOICED and VOICELESS sequences in which devoicing occurred in the six conditions investigated (all of the results presented are from data which has been pooled across all the subjects).

In VOICELESS sequences occurring syllable-initially (i.e. with only an intervening segmental boundary), there is always progressive devoicing of the following VOICED sound. In clusters with a deeper medial boundary (conditions 2 through 6), devoicing of the following VOICED sound does occur, but less frequently. Statistical analysis using the CHI^2 method has revealed no significant effect of boundary condition in the pooled data, or in the individual subject data. There are however, two tendencies which are consistently present across all the subjects and which are noteworthy. Firstly, there is a consistently greater frequency of devoicing when clusters occur in syllable-initial position compared to when they occur in non-syllable-initial position. This is due to the fact that in the former boundary condition, progressive devoicing always occurs, whereas in the other conditions, a range of different realisation strategies are observed, only one of which is progressive devoicing of the following sonorant. The second tendency emerging from the results is that the clause boundary condition produces by far the fewest cases of progressive devoicing.

In the cases in which progressive devoicing did not occur, the most frequent strategies were insertion of a pause between the release of the stop and the onset of voicing for the following sonorant (especially in the clause boundary condition), replacement of the stop by a glottal stop or glottalisation of the stop, and an incomplete stop closure. The relative frequency of occurrence of the various alternative realisation strategies observed is shown in Figure 2.

In syllable-initial VOICED clusters, devoicing of the sonorant occurs almost as reliably as it does in VOICELESS clusters under the same boundary conditions (this is a reflection of the fact that phonemically VOICED stops in SBE are frequently not accompanied by vocal fold vibration throughout their duration [5]). However, it is not the case that devoicing *always* occurs in VOICED clusters in syllable-initial position. When deeper boundaries intervene between the components of the cluster progressive devoicing does occur, but less frequently than in the segment-boundary condition. The main effect to emerge from the CHI^2 analysis is that the frequency of devoicing is significantly greater in VOICED clusters occurring in syllable-initial position than in VOICED clusters occurring in the other five boundary conditions ($p < 0.001$).

In the instances in which progressive devoicing did not occur, the most frequent strategy which was observed in its place was continuation of voicing unbroken right through the closure and release phases of the stop (see Figure 2). This occurred in almost one third of the non-syllable-initial VOICED clusters (it was also subject to large between-speaker differences -- some subjects used this strategy frequently, others hardly at all).

3.2. Extent of devoicing

The amount of devoicing in VOICELESS clusters is affected by the place of articulation of the stop and the identity of the sonorant, thus confirming the results of [5]. In every case, with the exception of /pl/ there is a greater delay in voice onset when the sequence occurs in syllable-initial position compared to when the same sequence occurs across deeper boundaries. There is no consistent trend observable across the remaining boundary conditions. In some cases, the delay in voice onset becomes gradually less as the boundary is deepened, whilst in others the values vary seemingly at random. There is no trend for the shortest values always to occur under the deepest boundary (this is only found in /tw/, /kl/, and /kw/.

The amount of devoicing in VOICED clusters is affected by the place of articulation of the stop and the identity of the sonorant, as established in [5] It is particularly noteworthy that the voice onset times for /dr/ sequences are almost double those for other voiced stop clusters. With regard to the effect of the boundary condition, no clear tendencies emerge from the data. Unlike the findings for VOICELESS clusters, it is not the case that the highest delays in voice onset are observed when the sequence occurs in syllable initial position.

4. Discussion

The results show that fine details of voicing timing in these clusters in SBE are affected by medial boundary status. Specifically, the nature of the boundary between a stop and a following sonorant has an effect on the probability of devoicing of the sonorant taking place. The crucial factor determining this aspect of voicing timing in stop-obstruent sequences seems to be that it should occur syllable-initially; when a syllable boundary intervenes devoicing occurs less frequently. It is instructive to compare these results with data, given in Figure 3, concerning the frequency of occurrence of progressive devoicing in /s/-nasal sequences under the same boundary conditions. The frequency of progressive devoicing is approximately the same under all of the different boundary conditions. This suggests that the devoicing observed in these clusters is a different sort of context-sensitivity to that observed in the stop-sonorant sequences (possibly a lower-level form of coarticulatory activity since it is not inhibited by the different grammatical boundaries).

The fact that devoicing of the sonorants has been shown to occur under all of the boundary conditions investigated provides counter-evidence to claims made in descriptions of the pronunciation of SBE. It was not found that devoicing occurred only within words or between words that formed 'closely-knit units' This backs up the findings of [3] and [1] who both provided a small amount of data on the devoicing of /l/ following stops in English.

A more accurate description of what is taking place in SBE is as follows In VOICELESS sequences, the default CV implementation strategy involves considerable progressive devoicing of the sonorant. In VOICED sequences, the default CV implementation strategy is to commence voicing just after the release of the stop (on the assumption that the stop closure is not accompanied by voicing throughout its duration). Across all the sequences, in the $CV condition (boundary condition 1), the default case is the one which occurs most frequently. In the other conditions (boundary conditions 2-6) a good deal of free variation takes place, and the default case occurs less frequently, the deeper the boundary condition. In both VOICED and VOICELESS sequences a range of different realisation strategies is used, apparently in free variation. In both types of sequences, but especially noticeable in the VOICED sequences, it appears that there are considerable between-speaker differences regarding the likelihood of one particular strategy occurring rather than another one.

The fact that the default case occurs less frequently under deeper boundary conditions may reflect the fact that under those conditions, a greater number of alternative strategies are available to the speaker (e.g. the pause option would presumably only normally be available to boundary conditions 4-6) Alternatively it may reflect a reduction in the perceptual weight which is being carried by the devoicing in the deeper conditions. It will be possible to evaluate whether any perceptual weight is carried by this aspect of timing, by observing subjects' response to synthetic stimuli which are constructed in such a way as to countervene the findings of this experiment.

It is clear that in SBE, this particular feature of interarticulator coordination is not wholly determined by peripheral, mechano-inertial effects. Phonetic implementation is subject to a constraint which is introduced at a more central level of utterance planning. These findings can be added to the ever-increasing body of data (see [7] and [9] for further examples) which suggests that there exists a wide range of complex, fine-grained, language-specific, systematic realisational detail which cannot be attributed to universal effects, and yet which so far have failed to be adequately accounted for in models of speech production.

References

[1] Bladon, R.A.W., & Al-Bamerni, A. (1976) Coarticulation Resistance in English /l/. *Journal of Phonetics,* 4:137-150

[2] Browman, C., & Goldstein, L. (1986) Towards an Articulatory Phonology. *Phonology Yearbook,* 3:219-252

[3] Dent, H. (1984) Coarticulated Devoicing in English Laterals. *Work In Progress, Phonetics Laboratory, Reading University,* 4:111-134

[4] Docherty, G.J. (1985) The Timing of Voicing in Stops and Fricatives in British English. *Edinburgh University Linguistics Department, Work in Progress,* 18:27-41

[5] Docherty, G.J. (1987) An Experimental Study of Voicing in Obstruents in British English Unpublished ms.

[6] Docherty, G.J., & Shockey, L.S. (1987) Speech Synthesis: Weighing Factors of Task Complexity and Linguistic Sophistication. Paper submitted to *Proceedings of the Institution of Electrical Engineers.*

[7] Fourakis, M., and Port, R. (1986) Stop Epenthesis in English. *Research in Phonetics and Computational Linguistics, Indiana University,* 5:38-71

[8] Gimson, A.C. (1980) *An Introduction to the Pronunciation of English* Longman.

[9] Port, R. (1986) Translating Linguistic Symbols into Time. *Research in Phonetics and Computational Linguistics, Indiana University,* 5:153-173

Figure 1

The frequency of progressive devoicing in voiced and voiceless stop - sonorant sequences in the six boundary conditions.

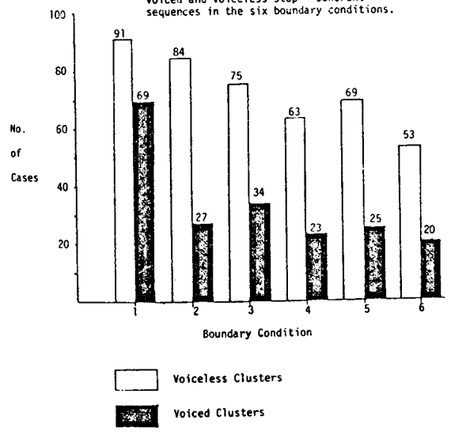


Figure 2

The frequency of occurrence of alternative implementation strategies in the six boundary conditions.

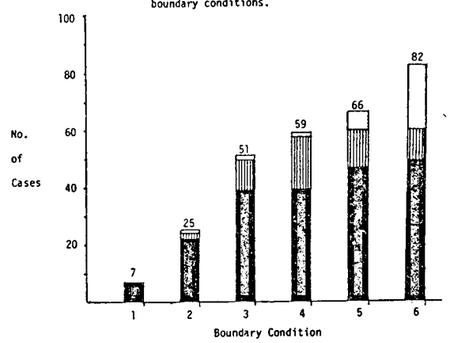
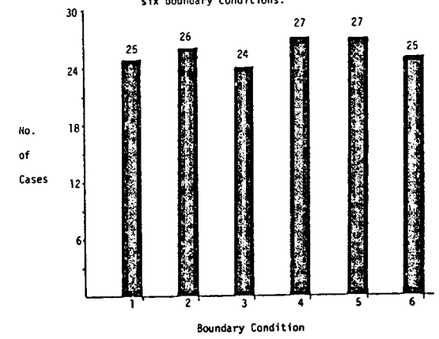


Figure 3

The frequency of occurrence of progressive devoicing in /s/ - nasal sequences in the six boundary conditions.



THE VARIATIONS IN THE WORD PHONETIC STRUCTURE CAUSED BY SPEECH TEMPO VARIATION

LILIA VYGONNAYA

Dept. of Experimental Phonetics
Institute of Linguistics
Minsk, Byelorussia, USSR 220000

ABSTRACT

Observations of the changes of word phonetic structure show that the constitutional language features are less changeble as compared to the intonationally relevant speech characteristics. It is confirmed experimentally that speech tempo evaluation is based not only on length but on other prosodic parameters.

It is important to ascertain the nature of the changes in the phonetical structure of a word in the light of essential dependence of a word phonetics on a phrase prosody and a speech tempo in particular / 1,2,3/.
Changes of the phonetical structure of a word within the bounds of narrative sentences of various length caused by quickening or slowing down a speech tempo were analysed. Speech recordings of the Byelorussian literary language speakers from different dialectal zones served as a base material for the research. The speakers reproduced the experimental text containing 21 phrases, 1981 sounds at a normal, fast and slow tempo. The text included phrases of the type: Невысокі хударлявы чалавек размеранаю роўнаю хадою ідзе лесам. Лес, ахінуўшы свае плечы белаю накідкаю, урачыста захоўвае цішыню і спакой. Такі лагодны спакой бывае толькі ў мінуты мудрага разважання. Зоркай іншага сусвету бліснуў кляновы ліст, што раптам з'явіўся перад вачыма на дарозе. Навальніца -- атмасферная з'ява, пры якой у магутных кучава-дажджавых воблаках і паміж воблакамі і зямлёй пры вялікай напружанасці атмасфернай электрычнасці ўзнікаюць моцныя электрычныя разрады -- маланкі. Зімнія навальніцы бываюць вельмі рэдка. Агульны зарад асобнай маланкі невялікі.

The selected material was recorded in a studio. It was subjested to the acoustic and instrumental analysis and the data were subsequently processed mathematically. The transcription of the text was carried out with the help of a segmenting apparatus (a tape recorder with rotating heads) which permitted to hear repeatedly word segments from 10 to 900 ms long as well as a whole word with extracted segments. The text transcription was controlled visually with the help of the oscillograph. Duration of phrases, psuses, word groups pronounced without pauses as well as of sounds and open syllables was defined. Errors in phrase duration did not exceed ± 0,5 s and could be perceived (≈15%) only when analysing definate short phrases uttered at a normal and fast tempo. In experiment the text was read by 29 men and 11 women at the age of 25-30 years. They were representatives from the north-eastern (14), south-western (16) and mid-Byelorussian dialects. (10).
Determination of the mean tempo of speakers pronunciation showed that they uttered at mean, slow and fast tempo individually /4/.
14 speakers pronounced the text with the mean sound duration from 69 to 80 ms, 18 speakers uttered with the mean duration from 82 to 88 ms. In 8 cases this magnitude exceeded 92 ms.
The comparison of the figures obtained with the data characterizing Russian speakers shows, that the Byelorussian speech tempo is slower, than Russian / 5, 6/.
Regular qualitative and quantitative sound variations due to the changes of speech tempo were analysed. There is no doubt that speeding up or slowing down of articulation causes certain qualitative sound changes. Nevertheless no delimitation of phonetical phenomena caused mainly by the changes in speech tempo or caused by some other reasons in different languages was made although it is of a typological nature. It is obvious that even a considerable quickening and slowing down of speech tempo (by 1,5 - 2 times) does not cause a loss of speech perceptibility and its categorical language features.
The analysis of the transcribed material was carried out with regard to sintagmal division of a phrase, sandhi vocal and consonant speech errors such as false starts, word substitutions, word omissions, hesitation pauses were taken into account and their number of various speeds of utterance was registered. It was defined that the character of speech errors was not directly connected with a certain type of changes in speech tempo. Word omissions, substitutious of short words for long ones were observed at a slow tempo (рэльефы ~ рэльеф, процілеглага ~ працяглага). Word transpositions (яркае свячэнне ~ свячэнне яркае; моцна разграе ~ разграе моцна; забяспечвае зямлю ~ зямлю забяспечвае; лідэр дахо-дзіць ~ даходзіць лідэр; у адваротным кірунку ~ у кірунку адваротным; ударную гукавую хвалю -- пярун ~ пярун -- хвалю), repetitions (што -- што, свячэнне -- свячэнне) and insertions (гэта, яны, ён, той), substitution of long words for short ones (палове дня ~ палавіне дня) were observed at fast tempo. Hesitation pauses were more frequent at a slow tempo, false starts of words (mainly of phrase initial words) were more frequent at a fast tempo (лес ~ леса, села; такі ~ які; цэлы ~ малы; зоркай ~ за зоркай, зойкай, сузолкай; больш-шасць ~ больш; пры гэтым ~ этым, пры тым). In the overwhelming majority of instances a considerable change in a speech tempo caused certain phenomena which are to be found in the speech of announcers while in the state of emotional stress /7/.
The degree of speech segments fusions, quantity, quality and localization of pauses depended on a tempo (mean tempo: Невысокі хударлявы чалавек /120 ms/ размеранаю роўнаю хадою /190 ms/ ідзе лесам /810 ms/ Лес ахінуўшы свае плечы /450 ms/ і спакой /520 ms/. Fast tempo: Невысокі хударлявы чалавек размеранаю роўнаю хадою /80 ms/ ідзе лесам Лес ахінуўшы свае плечы белаю накідкаю /60 ms/ урачыста захоўвае цішыню і спакой /100 ms/ Intonational structure of phrases changed considerably with quickening or slowing down of the speech tempo. At a slow tempo the role of a word stress and separability of each word increased as well as the number of words having a strong stress and the number of cases of laying a secondary stress increased (невысокі, хударлявы, чалавек, старасвецкіх, асірацелыя, атмасферная, распрацаваных, дажджавое, ледавіковае, размеранаю, роўнаю, з павышаных). Laying of independent stress on dynamically unstable words was observed as well (на беразе, не варушачы, на дарозе, пры якой, і паміж воблакамі, на Беларусі, над узвышанымі, да васемнаццаці, на зямным, каля ста, за адну, а дадатныя, да зямнога, са слабым, па якім, да зямлі, ад зямлі, пры гэтым, цякуць у распрацаваных, на інтэнсіўнасць). At a quick tempo the number of word stresses and separable units reduced (свае плечы, такі лагодны, цэлы гай, калі лясы, што забяспечвае, што моцна, так што, палове дня, больш моцны), the loudness increased. Melodic structure of phrases was also subjected to changes dependent on tempo variations.
An increasing significance of speech flow vocalic elements was registered among the most important interword phonetics phenomena, observed in low speed. Their weakening was observed when a speech tempo quickened, the alternations of syllabic and asyllabic elements y:y, i:i, existing in the Byelorussian language, at a slow tempo were mainly in favour of syllabic elements, while at a fast tempo they were in favour of asyllabic ones (Slow tempo: толькі у мінуты; дубы у чорных шапках; купалі у сонцы; размешчаныя у басейне; накідкаю урачыста; паміж воблакамі і зямлёй; высечкамі і прасекамі, азёры і балоты . Fast tempo: толькі ў мінуты; дубы ў чорных шапках; купалі ў сонцы; размешчаныя ў басейне; накідкаю ўрачыста; паміж воблакамі й зямлёй; высечкамі й прасекамі; азёры й балоты).
In the sphere of vocalism a change in vowel correlation according to their duration was fixed. It depended on a tempo change direction and was more significant in regard to low and mid vowels. At a slow tempo the vowels became longer, at a fast tempo they became shorter. Qualitative vowel changes at a fast tempo did not exceed the limits of phenomena observed in the case of reduction (воблака ~ воб/лъ/ка; у воблаках ~ у воб/лъ/ках; дажджавых ~ даж/джъ/вых; найчасцей ~ най/чъ/сцей; чалавек ~ ча/лъ/век; урачыста ~ у/ръ/чыста; разважання ~ /ръзвъ/жання; асірацелыя ~ аси/ръ/целыя, аси/р'/цельи/йь/; палове/пъ/) /8/. Tempo variations did not influence the phenomena of "jakanje" and "akanje".
Changes in duration of consonants are less essential as compared to those typical for vowels. A number of phonetic changes affecting consonants were not connected with tempo variations. For instance the simplification of consonant groups accomponied by weakening or loss of plosive and omission of sonants were observed mainly at a fast tempo (старыя ~ стар/ыъ/; атмасферная ~ атмасферн/аа/; адмоўныя ~ адмоў/ныъ/; дадатныя ~ дадатн/ыъ/; найчасцей ~ начасцей; участкамі ~ учаскамі; частцы ~ часцы; электрычнасці ~ элетрычнасці). At the same time the simplification of consonant clusters caused by vocal insertions and the loss of a word final consonant was registered at all the three speech tempo gradations (магчы-

масць ~ магчымась; ляс істасць ~ ляс істась; напружанасць ~ напружанась; дваццаць ~ /дъв/аццаць; стварэння ~ /сът/варэння; папярэднічае ~ папярэ/дьн/ічае). The speech tempo did not influence the hardness and softness of correlation of consonants. In some cases residual ties of speakers with certain dialectal zones were reflected (раскінуўся ~ ра/с'к'/інуўся; іншы ~ і/н'ш/ы; высечкамі ~ вісечкамі; выклікае ~ віклікае, ліпень ~ ліпен).

The data obtained permit to draw a conclusion that the constitutional features of a language (a dialect) are better presented with the tempo variations than those having predominantly intonational significance. They conform to the existing notion about the stability of certain phonetical features at various speeds of phonation and about the absence of direct ties between changes in a speech tempo and word phonetics /9,10,11/.

A series of perceptional experiments, in which 20 persons evaluated a tempo of sound segments of various length (one to twenty syllables) was carried out. Identical lexical sound units (phrases, word groups, separate words), pronounced by one and the same person at three various tempi were extracted from the experimental text, then assembled in pairs and in triads (the distance between the segments was 0,2; between the pairs and triads - 2s) and was produced to the auditors to indentify the tempo of each speech fragment. Each speed stimulus produced was repeated 72 times.

The analysis of the speakers ability to indentify various prosodie qualities of a phrase and its parts is necessary, because objective characteristics of speed signal and prosodic markers do not correspond to each other directly. While it is easy to single out objectively certain acoustic parametres of a speech signal, it is impossible to state definitely how their perception and evaluation are carried out. The obtained experimental data made it possible to see to what degree the auditors are able to compare different speech segments according to their intonation pattern using mainly one of its acoustic parametres - a tempo.

It is found out that the tempo of a whole phrase was identified better, than the tempo of its separate fragments. The full tempo was evaluated better, than the articulatory one. The degree of deviation of speed segments duration in quick and slow tempo from the same units pronounced at an average tempo was compared to the identification level of the produced segments. The comparison showed that identification was better when the segments produced to the auditors at a fast tempo were 4-7% times shorter and at a slow tempo 10-12% longer than the corresponding segments. Such improvement of tempo identification was observed when segments consisting of 6 to 18 syllables were identified.

The auditors identified better those segments pronounced at a fast tempo which not only differed temporally, but were also characterised by slightly increased loudness and more expressive melidic patterns. A phrase, pronounced in fast tempo was characterised by a special marking of the first word, which didn't change the melodical frame of the whole phrase. Stres syllables of other words were shaped independently. A considerable rising-falling tone movement was observed even in a post-stressed syllable, marked out in a special way. The total quantity of melodical peaks was greater than in slow pronunciation of the phrase. The perception of a tempo by the auditors was based not only on the evaluation of duration but on other prosodic parametres as well /12/. It provides thegrounds to state that the language phenomena defined by the term "Speech tempo" is a complex one from the point of view of a language speaker and duration is one of its constituents.

REFERENCES

/1/ Т.М.Николаева, Фразовая интонация славянских языков, М., 1977.

/2/ Н.Д.Светозарова, Интонационная система русского языка, Л., 1982.

/3/ Дж.Х.Гейтенби, Эластичные слова.-- Исследование речи: Труды Хаскинской лаборатории. Сб. переводов, Новосибирск, 1967.

/4/ Фанетыка слова ў беларускай мове, Мінск, 1983, с. 19--98.

/5/ А.С.Агафонова, Л.В.Бондарко, Л.А. Вербицкая и др., О некоторых характеристиках русской речи в зависимости от разных темпов произношения.-- Слух и речь в норме и патологии, Л., 1979, вып.1.

/6/ Р.Ф.Пауфошима, О темпе речи в некоторых русских говорах.-- Русские говоры, М., 1975.

/7/ Э.Л.Носенко, Особенности речи в состоянии эмоциональной напряженности, Днепропетровск, 1975.

/8/ Галосныя беларускай мовы, Мінск, 1975.

/9/ Л.Р.Зиндер, Влияние темпа речи на образование отдельных звуков.-- Ученые записки ЛГУ, №325. Сер. филол. наук, Вып. 68, Л., 1964, с. 3--28.

/10/ С.С.Высотский, О звуковой структуре слова в русских говорах.-- Исследования по русской диалектологии, М., 1973.

/11/ S.Wood, Speech tempo.--Phonetics Laborary Lund.Univ. Working Papers, 1973, N 9.

/12/ Л.Ц.Выгонная, Успрыманне тэмпу маўлення.-- Беларуская лінгвістыка, Вып. 24, Мінск, 1983.

ON TEMPO DIVERGENCES IN MONGOLIAN LANGUAGES

TAMARA YESENOVA

Philological Department
Kalmyk State University
Elista, Kalmykia, USSR, 358000

ABSTRACT

The phonetic divergences in the structure of modern mongolian languages have long been attracting the attention of scholars. Despite a great interest in Mongolian phonetics, the problem of reasons of sound alteration in Mongolian word is still very actual. There's an opinion, expressed by V.I.Rassadin following B.Y.Vladimirtsov that the modification of phonemic aspect of Mongolian word was influenced by the general weakening of the articulation.

INTRODUCTION

The Mongolian languages, being in a foreign environment, have been greatly influenced by the languages in contact (Russian, Chinese, Turkic, Tunguso-Manchurian). Thus, for example, the change of h<s, the monotony, the slowing down of the speech tempo are regarded to be the result of long-term contacts of Buryats with Evenks. The language of the Mongols of Inner Mongolia, who had lived in the Chinese surrounding for a long time, has undergone profound structural changes. A.D.Rudnev first noted that the language of East Mongols remains Mongolian, but the intonation, its rhythm and tempo, i.e. the external aspect of speech reminds that of Chinese /1/. In the study of the phonetic divergences of mongolian languages they do not take into consideration the prosodic data which, in our opinion, promoted their appearance.

The tempo of speech most frequently distinguished by the scientists is the prosodic difference of Mongolian dialects. A.D. Rudnev was the first to pay attention to it by saying: "of all the Mongolian tribes known to me the Buryats speak most slowly"/1/. Among Mongolian languages the literary Buryat language is singled out as a language where the non-first syllable vowels are the most distinct ones. But the dialectological material shows that the process of reduction involves most of the Buryat dialects, though the degree of reduction is not as strong as in the Kalmyk or Khalkha-Mongolian. The Mongolian scholar E.Vandue who studied the Derbet dialect refers to the slow speech tempo of Derbets, as compared with that of Khalkha /2/. B.Y.Vladimirtsov also pointed out the divergences in the dialects of the Volga Derbets and those of

Kobdoss./3/.

The experimental data showed that the average duration of the consonant in the Buryat language is 150-198 ms; of a short vowel - 100ms; of a long vowel 133-205ms /4/. The average statistic duration of the sound in Kalmyk according to our data is 88ms. Thus, the difference is essential. The Kalmyk dialects are characterized by the most fluent tempo in comparison with the speech tempo of other Mongolian dialects. But they are also discriminated on the basis of this significant prosodic feature which is the most important condition of formation of oral speech. The quick speech tempo is inherent to all Kalmyk dialects, it is illustrated by the data given below on the tempo of articulation of sounds in 2 basic Kalmyk dialects: the average duration of the sound in Torgut dialect is 81,9ms, in Derbet - 87,2ms. Having a general fluent tempo of speech the Torgut dialect displays the smallest average duration of the sound, i.e. the tempo of speech is the highest. The consistent divergence in the average duration of the sound in the dialects is observed in all intonation types of sentences. Thus, the average duration of the sound in the final intonation without accentuation makes 81,4ms within Torgut dialect and 88,3 ms in Derbet; in the final intonation with accentuation - 80,1ms, 81,4ms. The longest duration is due to the falling melody, and the shortest - to the rising melody.

However, the analysis of the distribution of these duration figures points out within the Kalmyk phrase that the tempo alteration is observed throughout the whole sentence. While the tempo of pronunciation of the initial syllables in the phrases pronounced with the falling and rising melody is relatively the same. The final syllables are pronounced with divergences, the rising melody is accompanied by increasing of the sound duration of the final syllables in the phrase. The sound of intermediate syllables of the phrase in the first and the second melody types are characterised by the relatively slow duration i.e. its peculiar of them to quicken the tempo of pronunciation.

Compare: the average relative duration of the sound in the final melody at the beginning, in the middle and at the end of the sentence is 2,01-1,86-1,92; in the rising 2,01-1,85-2,32 respectively.

As it was noted more than once the temporal characteristics singles out the informatively significant components of the phrase. The Kalmyk language, alongside with other Mongolian languages, is characterized by the fixed word-order in the sentence where the predicate is in the final position.

However, in some styles of literary Mongolian language the inversion of parts of the sentences is perceived as a fact of normalisation and legalisation of a more free word order in general /5/. As a rule, parts of phrases singled out by the tempo correspond to the rheme. Informatively less significant parts of the phrase are pronounced in a higher tempo, which results in modification of the phonemic aspect of words and sometimes even in sound elimination. As for the structure of the word itself this tendency embraces all its elements. The modification of phonemic structure of Mongolian word is a thoroughly investigated aspect in literature: the development of long vowels owing to the weakening and disappearance of intervocal consonants, sporadic falling out of sonorants, vowel reduction in a weak position are observed in all Mongolian languages, The Khalkha-Mongolian is closer to the Kalmyk on the degree of reduction. We observe the complete disappearance of short vowels in speech, the degree of reduction depends on the tempo of speech; the higher the tempo, the stronger the reduction.

It is exactly this peculiarity that is responsible for typical phonetic mistakes in the Russian speech of the Mongol speaking people who haven't quite mastered the literary pronunciation. Thus, for example, one of the marked peculiarities is the ellipsis of non-first syllable vowels; машина-машин, касса-касс, парта-парт and etc. It is so consistent that it is frequently reflected in written speech. In the Khalkha-Mongolian language not only separate vowels, but very often the final syllable of the word is reduced completely. If in the Kalmyk and Buryat languages the final consonant "n" which is the morphological feature of the noun is preserved when the preceding vowel falls out, in the Khalkha language it consistently falls out too. The reduction of vowels lowers down the vocalic factor and the consonantal saturation of the word. The quickening of the speech tempo results in carelessness, indistinction in the articulation of sounds. It changes the correlation of occlusives and constrictives to a great extent and frequently leads to spirantisation of occlusives and sometimes to their total disappearance (especially unstable is the mediolingual constrictive <j>), simplification of the groups of consonants due to the falling out of the consonants between the voiced, devoicing of the voiced, the falling out of the constrictives. Thus, in the Buryat dialects the forelingual occlusive <n> changes into medio-lingual constrictive <j> and it's characteristic of Sard-Kalmyks to change the soft constrictive forelingual <l'> by the soft constrictive medio-lingual <j>. The original phonemic structure of Mongol word is changed due to simplification of articulation. The appearance of these changes is sufficiently substantiated by V.Y.Vladimirtsov when he speaks of All-Mongolian tendency of weakening the tension of articulation the basis of which is the phonetic law of economising the articulation efforts. In his "Comparative Grammar of the Mongolian

Written Language and the Khalkha Dialect" he wrote that the long final vowels in Khalkha words are weaker and more open than the long initials and the middle vowels. They are still more weakened at the absolute end of the utterance and before the pauses./3/. The economy of pronouncing efforts is carried out mainly due to falling out of sounds and vowel reduction. And these in their turn account for the reconstruction of the consonantal chain of the words. The vowels of the first syllable turned out to be untouched by the reduction as they are characterized by distinct articulation which is accounted for by their place in the word: the beginning is stronger than the end. Besides, it's required by their role in forming the vocalism of the word as they determine the vocalic structure of a vowel harmonised word. The progressive assimilation according to the vowel serie of the first syllable of the word is characteristic of all Mongolian languages. Based on the accuracy, on the distinction of pronunciation, on the vivid manifestation of all differential features of this vowel, many scholars speak of its accentuated character.

Thus, most modifications in the phonetic aspect of the word are explained by the weakening of tension of organs of speech, sponsouring inaccurate articulation. This tendency is redoubled by the quickening of the speech tempo, which is characteristic to Mongolian languages, but to a greater extent-to Kalmyk language and its dialects. If for the appearance of these modifications the maximum favourable position is the middle of the word, but in a phrase - the communicatively less important elements. However, with the general tendency towards the reduction of the sounds which do not belong to the first syllables in the definite phonetic positions the duration of vowels with inaccurate articulation proves to be high (in some cases - to 70ms).

In revealing the reasons of the appearance of innovations and the presevation of all its general features in the Mongolian languages it's necessary to take into account the prosodic data, such as, for example, the general speeding up of speech, the status of strong and weak elements of the phrase. Due to it the elements which are exactly in these positions undergo these changes.

/1/ A.Rudnev, "Materials of the dialects of East Mongolia", Sant-Peterburg, 1911.

/2/ E. Vandue, "The derbet dialect", Ulan-Bator, 1976.

/3/ B.Vladimirtsov, "Comparative grammar of mongolian written language and khalkha dialect", Leningrad, 1929.

/4/ M.Mokhosoeva, "Intonation of question-answer sentences of the buryat language", Ulan-Ude, 1981.

/5/ G.Purveev,"On innovations in the sentence Syntex of mongolian", Problems of linguistics, 1977, 5.

HOW MANY RISE-FALL-RISE CONTOURS?

Janet B. Pierrehumbert and Shirley A. Steele

AT&T Bell Laboratories,
600 Mountain Ave, Murray Hill, NJ 07974
U.S.A.

ABSTRACT

This paper reports an experimental study of the rise-fall-rise intonation patterns of English. Variants of the rise-fall-rise contour differ from each other in the alignment of the F0 valley and peak with the text. The phonological analysis of this difference is controversial and is part of a wider debate on the function of category and continuum in intonation. Variants of the rise-fall-rise pattern have been viewed as either (1) falling into two categories, each with a distinct pragmatic meaning, or (2) occurring along a continuous dimension of peak delay and pragmatic meaning. This issue was addressed by examining how speakers imitate stimuli varying continuously in the alignment of the F0 rise-fall.

In the stimuli for the experiment, the alignment of the F0 rise-fall was varied in small steps by using LPC coding and resynthesis. Subjects heard the stimuli in randomized order and imitated what they heard. Peak delays in the responses were found to cluster in two groups, thus differing systematically from the peak delays in the stimuli. This result is readily explained by a model with two categories.

1. INTRODUCTION

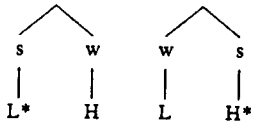
1.1 Topic

In this paper, we report an experimental investigation of the rise-fall-rise intonation patterns of English. Variants of the rise-fall-rise contour differ from each other in the alignment of the F0 valley and peak with the text. Figure 1 exemplifies the F0 contour of this pattern with an early peak and Figure 2 shows a longer peak delay. (In both Figures, a vertical line shows the location of the [m]-release.) The pattern in Figure 2 conveys speaker incredulity or uncertainty (see Ward and Hirschberg [7], [8]), while that shown in Figure 1 typically marks contrast or correction.

The phonological analysis of the rise-fall-rise intonation pattern is controversial and is part of a wider debate on the function of category and continuum in intonation. Peak delay variants of the rise-fall-rise pattern may be viewed as either (1) falling into some number of categories, each with a distinct pragmatic meaning, or (2) occurring along a continuous dimension of peak delay and pragmatic meaning. A formulation of the first view is found in Pierrehumbert [5] and a statement of the second in Gussenhoven [1].

According to Pierrehumbert, there are two different rise-fall-rise patterns, differing in how the low-high pitch accent is aligned with respect to the stressed syllable. In one, the high (H) is aligned with the stressed syllable (Figure 1). In the other, the low (L) falls on the stressed syllable (Figure 2). Using a diacritic "*" to represent alignment with the stress, the two patterns are transcribed L+H* and L*+H and are called "bitonal accents" because two tones are required to describe the accent. In both patterns, the fall-rise after the accent is explained by a L H sequence which marks the end of the phrase.

In Pierrehumbert's theory, the "*" diacritic for tones is analogous to stress for syllables. In metrical stress theory, as laid out in Liberman and Prince [4], the stress pattern of a word like "Peter" has a relatively strong syllable followed by a relatively weak syllable. In "repeat" the strength relation is reversed. Pierrehumbert's bitonal accents are treated in the same way, with a strength relation among the tones, as illustrated below:



The single tone accents have a status corresponding to that of stressed monosyllables, such as "Pete".

This approach contributes to a broader picture in which tones participate in a hierarchical organization, which controls their alignment and phonetic realization. This broad picture is first proposed in Liberman [2] and is amplified and revised in Pierrehumbert and Beckman [6].

Gussenhoven asserts that English does not have a binary distinction, but a continuous dimension of peak delay. He suggests ([1], p. 218) that while there may be a preferred ("ideal") position along this continuum, a wide range is possible. His treatment of peak delay is thus analogous to the treatment of continuously variable overall pitch range in Liberman and Pierrehumbert [3]. Peak delay in pitch accents is treated as a paradigmatic feature, while word stress is viewed as a matter of metrical organization.

Our aim in the present work was to empirically investigate the category/continuum question in the rise-fall-rise intonation contour.

1.2 Method

In our experiment, subjects heard and imitated randomized rise-fall-rise contours constructed along a continuum of peak delay. If subjects perceive a continuum, the response peak delays should be approximately continuous. (A preferred peak position might cause responses to stimuli at the extremes to drift towards the center). If subjects hear categories, responses should cluster into discrete groups.

This experimental method is a variant of the paradigm familiar from studies of categorical perception of speech segments. Our study is, to our knowledge, the first application of such methods to the study of syntagmatic features. We have used an imitation task rather than the more commonly used labelling and discrimination tasks, because we were not concerned with separate analysis of production and perception systems. In most categorical perception studies, the linguistic analysis was relatively uncontroversial (e.g., English /b/ versus /p/); at issue was the status of linguistic description in the psychological system. In our study, we looked for evidence about the system of linguistic analysis; the relationship of any categories to perceptual or articulatory systems is a matter for later research.

2. EXPERIMENTAL PROCEDURES

2.1 Stimuli

The stimuli were versions of the phrase, "only a millionaire," in which the location of the F0 peak was incrementally moved from a relatively early to a relatively late position in the accented word. This was done by recording a natural production of the sentence and using LPC coding and resynthesis to produce a systematic set of variants. In both the original recording and in subjects' imitations of it, main stress fell on the first syllable of the word "millionaire", not the last. This is an acceptable pronounciation in American English and subjects reported no difficulty using it.

The shape of the rise-fall-rise pattern in the stimuli was established by making a piece-wise linear approximation to the rise-fall-rise pattern of the original recording. Peak positions varied between 35 and 315 msec from the end of the [m] in "millionaire", by 20 msec increments. As the peak was shifted, the durations of the rise and fall were kept constant. In the stimulus with the greatest peak delay, the peak occurred just before the end of the [n]. These bounds were established by asking several naive listeners to evaluate the naturalness of the stimuli. Stimuli at the ends of the continuum which the listeners felt to be unacceptable in English were eliminated.

The particular phrase used was chosen for two reasons. First, it is composed entirely of sonorants, thus avoiding devoicing in the F0 contour and minimizing consonantal effects in the stressed syllable. Second, its pragmatic interpretation could be sensibly altered by variation in peak delay. The early peak variant could be used to assert that the speaker does not feel very rich, while the late peak variant would be appropriate for an incredulous rejoinder.

Note that this meaning distinction could be viewed as either categorical (assertion versus incredulity) or continuous (along a dimension of degree of speaker commitment). The meaning difference was not discussed with the subjects; thus, using a sentence with two potential pragmatic interpretations did not prejudice the experimental outcome.

For each subject, data was collected in at least two sessions, using a real time data collection program. Subjects were told that they would hear a series of aural prompts, and that the phrase was always the same but the intonation varied. They were asked to listen carefully to each token and then imitate what they had heard. If subjects were not satisfied with a particular response or wanted to hear the prompt again, they could tell the experimenter, and the token would be repeated. No time limit was imposed for responses.

2.2 Randomization

The 15 versions of the prompt were randomized in blocks; then in each of two sessions, 15 different randomized blocks were presented to the subject, for a total of 225 tokens per session or 450 tokens in all. Thus, each of the 15 tokens was repeated 30 times.

2.3 Subjects

The subjects were five native speakers of American English, two females and three males. Four of the five were naive about the purpose of the experiment. The subjects were: DTT, a software engineer; HLT, a psychology research assistant; RLB, an opto-electronics processing engineer; SAS, one of the authors; and TWB, a high school student.

2.4 Measurement

The measurement of primary interest is peak delay, defined as the difference between the time of the F0 peak and the time of the [m]-release. This was found by examining displays of the F0 contour and waveform of each response. Time points were established for the release of the [m] into the vowel, for the F0 peak, for the implosion of the [n], and for the F0 minimum preceding the peak. The transitions into and out of the nasal consonants could in general be found with great accuracy, due to the abrupt change in the waveform occurring at these points. The F0 peaks were fairly narrow, and thus their location provides a good index of the location of the H tone. In cases where several time points shared the same maximum value, the earliest was selected. A few utterances had to be eliminated from the data set because the F0 at crucial points could not be measured. There were no more than four such utterances per subject. The data summaries for each stimulus represent in every case at least 28 responses.

3. RESULTS

3.1 Predictions about peak delays

In order to make the discussion of the data more transparent, let us first explain some idealized experimental results for different models of the intonation system.

First, consider a continuum model in which peak delay is continuously variable and all possible peak delays are equally preferred. In this model, the subject should, on the average, faithfully reproduce the peak delays in the stimulus. A plot of response peak delay against stimulus peak delay whould thus follow the line $y = x$. The overall distribution of responses, being the sum of the individual distributions, should be broad and unimodal, approaching the rectangular distribution of peak delay in the stimulus set.

Next, consider a continuum model in which peak delay is continuously variable but a central value is preferred. In this model, the responses tend to shift towards the center. Individual distributions are likely to show a dependence of shape on stimulus position in the continuum. The responses to stimuli at the ends of the continuum are likely to include tokens from the center of the continuum. But stimuli in the center (exhibiting the preferred form) would be less likely to elicit responses from the ends. Overall distribution would still be unimodal.

In a model with two categories, there are two preferred values of peak delay. The subject perceives each stimulus as an instance of the pattern with the closest peak delay value. He then produces an instance of that pattern. Thus, the overall distribution in the response data should be bimodal. A graph of response peak delay against stimulus peak delay should show a sigmoid shape. In the idealized case, the transition from the lower to the upper arm is abrupt. Real data however, usually exhibit a transitional region which arises when the stimuli seem ambiguous and the subject vacillates in his response.

3.2 Peak delay data

In general, the data supported the existence of two intonational categories. Data plots for three subjects are displayed in Figure 3. Figure 3A shows plots of data for subject TWB (the high school student). The histogram for the peak delays is obviously bimodal. On the plot of median response peak delay against stimulus peak delay, the diagonal line shows how the medians would behave if the subject had faithfully reproduced what he heard. It is clear that there are substantial deviations between the stimuli and the responses. For the first 9 stimuli, the peak delay values cluster between .1 and .15 seconds, whereas for the last 4 stimuli, they cluster between .2 and .25 seconds. The responses to stimuli number 10 and 11 have intermediate values for the median peak delay.

Figure 3B shows data for subject HBT. The results for this subject, and for subject SAS, were very similar to the results for subject TWB. HBT and SAS differ from TWB in the location of the boundary between the two categories.

Figure 3C shows data for subject RLB. This data set shows the same tendencies that we saw in the data for the other subjects, but less strongly. The second mode of the histogram is less pronounced than for the first three subjects. Also, on the response versus stimulus peak delay plot, there is a greater tendency for the median peak delay to track the values for the stimuli. We believe this tendency arose because of the subject's difficulties with the task. The subject reported that he was hearing more patterns than he could easily reproduce. His pattern of responses to the lower numbered stimuli is quite similar to that of the other subjects, but his responses to the higher numbered stimuli appear to include tokens of both the early peak and the late peak pattern, giving rise to broad distributions. Thus we believe that he was aware of a category difference in the stimuli but did not completely control it in his own speech.

The data for one subject, DTT, did not conform to that for the other subjects. The histogram for all his data is unimodal, and the median peak delay shows little variation. We believe that DTT lacks the L*+H pitch accent. The continuum theory must also make a special case of DTT, since he does not exhibit the substantial range of variation in peak position which is claimed to be possible. Presumably DTT would be described as having an unusually strong preference for his central peak position. Thus, DTT does not provide strong evidence for distinguishing between the two proposals.

3.3 Other characteristics of the data

The early peak and delayed peak variants of the rise-fall-rise showed no significant difference in the F0 minimum value preceding the peak. Thus, we are confident that subjects produced instances of the L+H* and the L*+H, not instances of plain H*, which would exhibit a much higher F0 minimum preceeding the peak. According to several measures, the L tone occurs later (relative to the segments) in the delayed peak variant than in the early peak variant, as predicted by the Pierrehumbert model. We omit supporting graphs, for lack of space.

Individual histograms of responses to each stimulus were examined. In general, these were broader in the transitional region of each sigmoid than on the arms. This result is predicted by a two-category model, since the distribution of responses to an ambiguous stimulus arises as a mixture of sampling from the distributions for the two categories. We are in the process of determining whether individual distributions can be quantitatively modeled in this way.

The relation of the peak delay to duration pattern is also of interest. It is sometimes claimed that syllables bearing L tones are longer than equally stressed syllables bearing H tones. Since the L*+H accent and the L+H* accent place opposite tones on the stressed syllable, there might be an effect of this sort. It is also important to rule out any possibility that the peak delay data might be an artifact of durational differences. In the phrase we used, it was impossible to measure the duration of the stressed syllable per se. The [l] in "millionaire" does not yield a well-defined measurement point; in fact it was often missing, with the syllables separated only by a [y] glide. However, an increase in the duration of the stressed syllable should be reflected in an increase in the total duration of from [m] to [n], which was easily measured.

The [m]-to-[n] duration did not increase with peak delay. The range of variation in duration is about one third the range of variation in peak delay. There is little pattern to it, and what pattern exists is not consistent across subjects. Only SAS shows some evidence for longer durations at longer peak delays. But for her, the effect is still far to small to explain the variation in the peak delays.

4. CONCLUSION

Four out of five subjects support the existence of two categories of rise-fall-rise. The L tone shifted rightward with the peak, as predicted. No significant durational effects were found. Thus, the results support a taxonomy in which alignment functions as a binary linguistic distinction.

REFERENCES

[1] Gussenhoven, C.: (1984) *On the grammar and semantics of sentence accents*. (Foris Publications, Cinnaminson, N.J.).

[2] Liberman, M.Y.: (1975) *The intonational system of English*. MIT Ph.D. diss.; published by Garland, New York, 1979.

[3] Liberman, M.Y.; Pierrehumbert, J.: (1984) Intonational invariants under changes in pitch range and length," Aronoff and Oerle, eds., *Language Sound Structure*. (MIT Press, Cambridge, MA).

[4] Liberman, M.Y.; Prince, A.: (1977) On stress and linguistic rhythm. *Linguistic Inquiry 8*: 249-336.

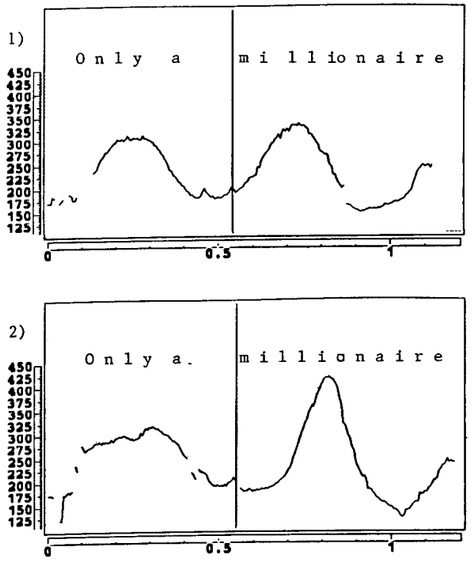
[5] Pierrehumbert, J.: (1980) The phonology and phonetics of English intonation. MIT Ph.D. diss.

[6] Pierrehumbert, J.B.; Beckman, M.E.: (forthcoming) *Japanese tone structure*. Linguistic Inquiry Monograph Series.

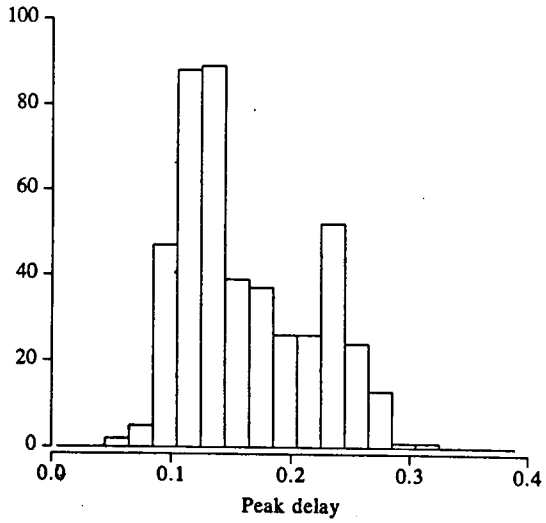
[7] Ward, G.; Hirschberg, J.: (1985) Implicating uncertainty: the pragmatics of fall-rise intonation. *Language* 61:4 747-776.

[8] Ward, G.; Hirschberg, J.: (1986) Reconciling incredulity with uncertainty: a unified account of the L*+H L H% intonational contour. Linguistic Society of America Annual Meeting.

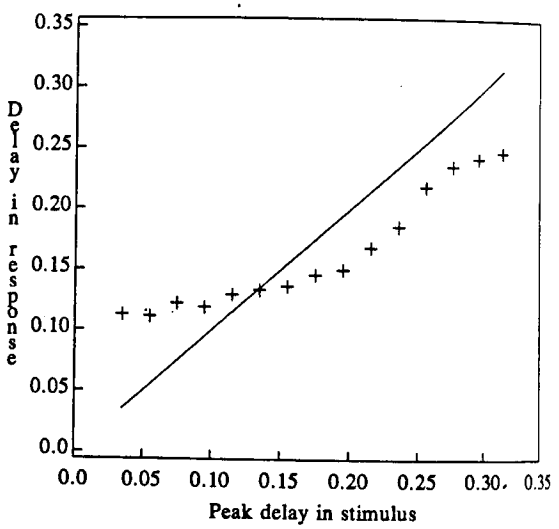
FIGURES



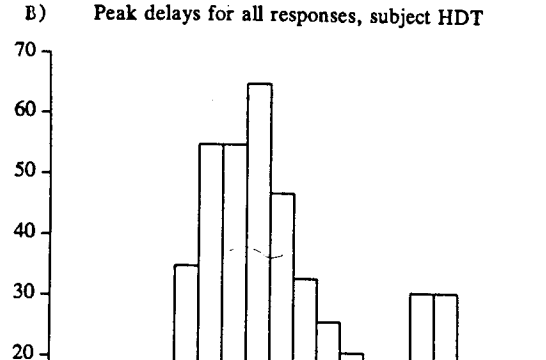
3A) Peak delays for all responses, subject TWB



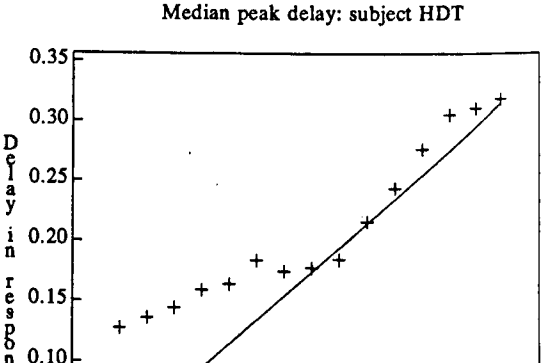
Median peak delay: subject TWB



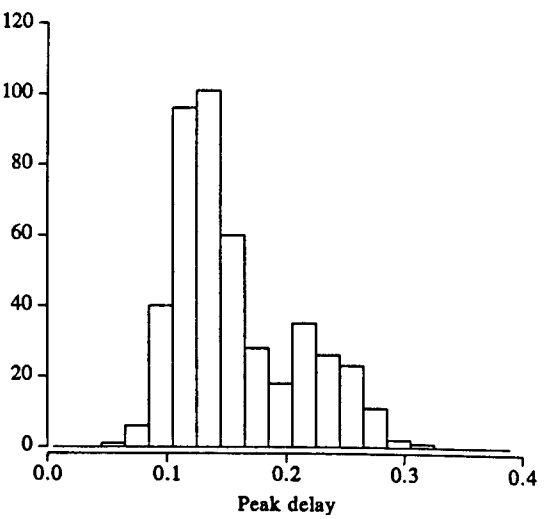
B) Peak delays for all responses, subject HDT



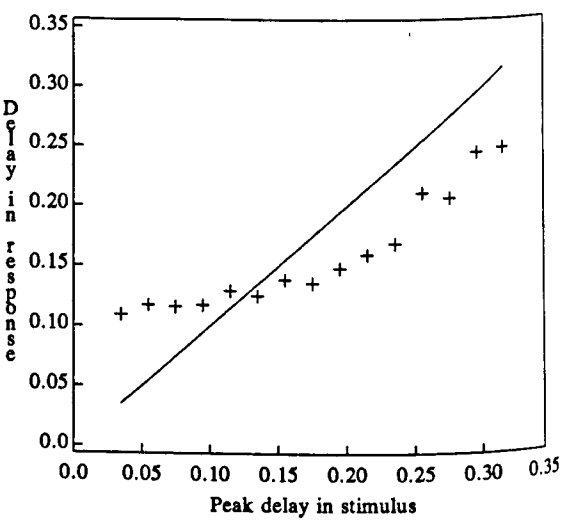
Median peak delay: subject HDT



C) Peak delays for all responses, subject RLB



Median peak delay: subject RLB



THE LINGUISTIC FUNCTIONS OF F0 PEAKS

KLAUS J. KOHLER

Institut für Phonetik und
digitale Sprachverarbeitung
Universität Kiel
2300 Kiel, FRG

ABSTRACT

F0 is an essential acoustic signalling property for both stress and intonation. Although sound duration and intensity are further characteristics of the former, a change in F0 may be sufficient to shift stress from one syllable to another. Taking a German minimal verb pair ("umlagern" with prefix or stem stress) as its point of departure, this paper presents experimental data showing (a) some of the conditions under which F0 is sufficient for a stress shift, and (b) the interaction of the stress and intonation functions of F0.

INTRODUCTION

It has been shown in /1,2/ that within the same word and sentence stress (e.g. the syllable "-lo-" in "Sie hat ja gelogen.") three types of intonation F0 peak positions are possible: early, medial, or late (in the syllable "ge-" or central or late in the syllable "-lo-" of the qoted example), with the corresponding changes of meaning from 'established' to 'new' to 'emphatic'. On the other hand, a shift of the F0 peak position from one syllable to another can change the stress position in a syllable chain. Thus two questions arise:

(a) Under what conditions is an F0 peak shift (without concomitant changes in sound duration and intensity) sufficient to shift stress to a different syllable?
(b) How can the stress and intonation functions of F0 peaks be differentiated, and in what ways do they interact?

To provide answers to these questions two experiments were carried out in German, which offers a good example for testing the issues because it has minimal verb pairs, with either prefix or stem stress, which can occur in the same natural sentence frame, e.g. "Er wird's wohl umlagern." (with stress either on "um-" /'um/ = "verlagern", "He is presumably going to shift it to another place."; or on "-la-" /'la:/ = "belagern", "He is presumably going to besiege it.").

PROCEDURE

Two utterances of this sentence, (a) with stress on "um-" and a medial intonation F0 peak on this syllable, and (b) with stress on "-la-" and an early intonation F0 peak, which actually falls on the syllable "um-", were selected for stimulus construction from a large corpus containing several repetitions of all the 6 combinations of 2 stress and 3 intonation positions, spoken by a trained phonetician (the author). The two tokens were analyzed using the same procedure as in /1/. Figures 1a,b present the waveforms together with their F0 displays. The F0 peak positions in the two uttterances are practically identical in relation to the syllable structures of "umlagern": they occur at more or less the same time interval before the beginning of /l/. The differences between the two are in the shapes of the F0 peak contours and in the syllable durations:

(a) in prefix stress, the F0 rise of the peak contour sets in at the beginning of "um-", in stem stress, however, as early as the beginning of /l/ in "wohl",
(b) in prefix stress, "um-" is much longer than in stem stress (222 ms vs. 135 ms), but "-la-" has very similar durations in both cases (258 ms vs. 268 ms).

In a second step, the F0 peak contours of the two utterances were exchanged and adjusted to the comparable points in the segmental structures. Figures 2a,b show the waveforms of figures 1a,b with the new F0 contours. Finally, the following F0 parameter manipulations were performed:

(1) In the stimulus of figure 1a (original prefix stress), the whole peak contour between the marks A and C was shifted to the right along the time axis in 6

equal steps of 30 ms; the tail of the F0 contour beyond mark C was then time-compressed between the new time position C' and the end of periodicity, and the F0 precursor in "wohl" was time-expanded from its beginning point to the new time position A'. The left branch of the peak contour (AB) was also shifted to the left in 5 equal steps of 30 ms; the right branch of the peak contour was then time-expanded between the new time position B' and the time mark C, and the precursor was time-compressed between its beginning and the new time position A'. When A' fell to the left of the beginning of "wohl" the section of the contour that thus entered the voiceless stretch was masked.

(2) In the stimulus of figure 2b (original stem stress with transferred F0 peak shape), the whole peak contour between the marks A and C was shifted to the left in 8 equal steps of 30 ms; the tail of the F0 contour beyond mark C was then time-expanded between the new time position C' and the end of periodicity. As regards the left-branch adjustment the same procedure was followed as in the left shifts of (1).

(3) In the stimulus of figure 2a (original prefix stress with transferred F0 peak shape), the same F0 peak shifts were carried out as in (1).

(4) In the stimulus of figure 1b (original stem stress), the same procedure was followed as in (2).

From these parameter manipulations, there resulted 12 F0 contours, with peak positions from near the beginning of "um-" to the second half of "-la-", in (1) and (3), and 9 F0 contours, with peak positions from the beginning of "wohl" to near the end of "um-", in (2) and (4). These F0 contours entered into a stimulus synthesis with the LPC-derived formant and volume values of the original prefix-stress utterance in (1) and (3), and with the corresponding data of the original stem-stress utterance in (2) and (4). In each case, two test stimulus sets were thus generated, with a slowly and an abruptly falling F0 peak contour, respectively: (3), (4) vs. (1), (2). In (1) and (3), the F0 peak positions straddle the syllable structures where a change from prefix to stem stress is to be expected if F0 is a sufficient cue. The two sets differ in that the peak shape of (3), but not of (1), approximates the configuration found in the early peak of the original stem-stress utterance (see figure 1b). It is hypothesized, therefore, that if stress is perceptually shifted at all in (1) and (3), there will be a more clear-cut change in (1) because there is a higher probability in (3) that an F0 peak position on "um-" is not only perceived as a medial peak with prefix stress but also as an early peak with stem stress. The same would apply to (4) as against (2).

To check these hypotheses two test tapes were compiled: (I) containing the 12 stimuli of (1) and the 9 of (2), (II) containing the 12 stimuli of (3) and the 9 of (4). (I) was produced in a short version with 5 repetitions of the 21 stimuli, and in a long version with 10 repetitions, with separate randomizations of the 105 and 210 test stimuli, respectively. (II) was only produced in a short version. Each stimulus sentence was preceded by a bleep and followed by a 4 s pause in which subjects were to answer, by ticking the appropriate boxes on prepared response sheets, whether the meaning of the perceived stimulus was "belagern" or "verlagern". 18 subjects did test (I) in its long version, 9 in its short one. 4 of the 18 deviated in their responses by judging the 9 stimuli of (2) exclusively as "verlagern". They were, therefore, dealt with separately and not included in figures 3 and 4. 16 subjects, some of whom had done test (I), took test (II) in later sessions. The subjects listened to the test tapes in several subgroups via a loudspeaker in a sound-treated room of the Kiel Phonetics Institute.

RESULTS AND DISCUSSION

Figures 3 and 4 present the results from these experiments for the 12-stimulus sets (1), (3) and for the 9-stimulus sets (2), (4), respectively. In the shift of the more sharply falling (original) F0 peak contour through the original prefix-stress utterance, there is a clear change from initial to stem stress, in spite of the duration of "um-" pointing to the former. F0 can thus override duration, particularly since the duration of the unstressed "-la-" syllable in the original utterance is very close to its duration under stress. In stimulus 10, which is the first in the ordering from 1 to 12 to yield an unequivocal stem-stress categorization with over 80% positive responses, the F0 peak position is 30 ms into the vowel of the syllable "-la-". This corresponds to the medial intonation peak on the stressed syllable found in /1,2/. The fact that the change from one stress category to the other is gradual rather than categorical can be related to some interaction of the stress and intonation functions of F0 because the more sharply falling F0 peak assumes positions before the beginning of the syllable nucleus /a:/ of "-la-" which can simultaneously function as the medial intonation peak in stressed "um-" and as the early intonation peak related to stressed "-la-". When the more slowly falling F0 peak is substituted the initial-stress category is not clearly represented, the interference from the early intonation peak of the stem stress becomes too strong.

When an F0 peak contour is shifted through the original stem-stress utterance, there is no change between the stress categories: the answers are predominantly in favour of stem stress. In this case, F0 can thus not override the duration cue completely because "um-" is too short in relation to "-la-" to signal initial stress. But there is some effect of F0 when the more sharply falling F0 peak is moved into the syllable "um-": in stimulus 5 the peak is positioned at the end of /l/ in "wohl", with the F0 fall occurring in "um-", in all subsequent stimuli the peak is itself located in "um-". When the characteristic peak contour occurs in the relevant syllable the duration cue is checked by the F0 cue to a certain extent. The 4 subjects that behaved differently from the other 23 were guided by F0 altogether: they only perceived initial stress in all the 9 stimuli, where the F0 peak precedes "-la-". The substitution of the more slowly falling F0 peak reduces the prefix-stress judgements because of the interference of the intonation function of F0.

The hypotheses that led to the experiments discussed in this paper have thus been confirmed, and the questions asked initially can be answered as follows:

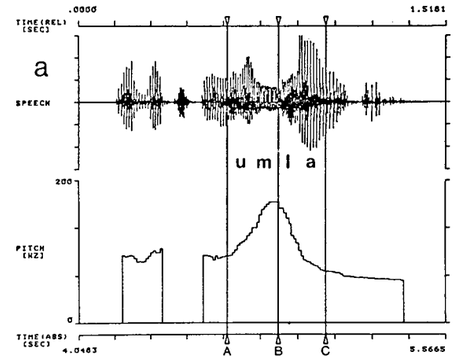
(a) An F0 peak shift by itself is sufficient to bring about a clear change from one stress position to another, provided the duration of the stressed-syllable-to-be is not too short. But even when it is there is a residual F0 effect.

(b) The intonation function of F0 interferes with its stress function if the latter is not supported by duration. This finds its expression in a gradual change from one stress category to another over a stretch of utterance where the positions of a medial intonation peak in one stressed syllable and an early intonation peak related to a stressed syllable following can coincide. This interaction is strengthened when the shape of the F0 peak contour approximates the more slowly falling one of the early intonation peak of a later stress.

REFERENCES

/1/ K.J. Kohler, "Computer synthesis of intonation", Proc. 12th Intern. Congr. Acoustics, A6-6, Toronto, 1986.

/2/ K.J. Kohler, "Categorical pitch perception", Proc. 11th Intern. Congr. Phon. Sc., Tallinn, 1987.



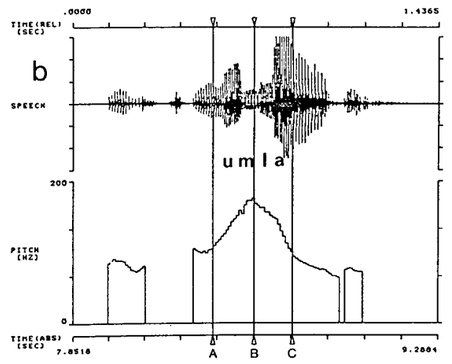
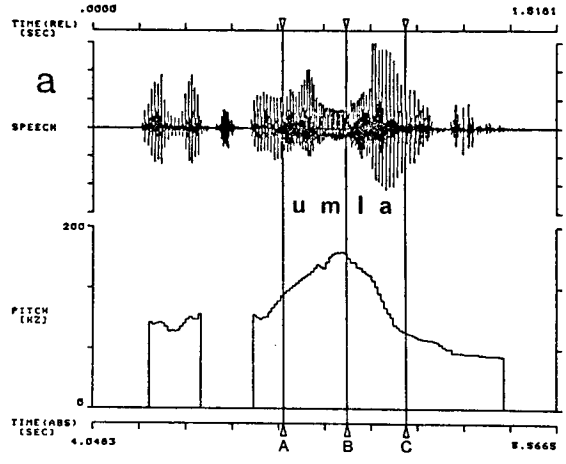


Fig. 1. Waveform and F0 of the original prefix-stress (a) and of the original stem-stress (b) utterance of "Er wird's wohl umlagern."A, B, C mark the base and peak points of the F0 peak contour for the F0 shifts.



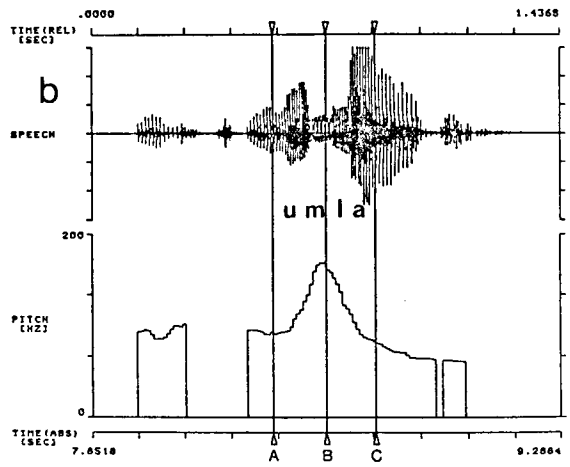


Fig. 2. Waveform of the original prefix-stress (a) and of the original stem-stress (b) utterance of "Er wird's wohl umlagern." with the F0 peak shape transferred from the stem-stress (a) and from the prefix-stress (b) utterance and adjusted to the different timing of the new utterance. A, B, C as in figure 1.

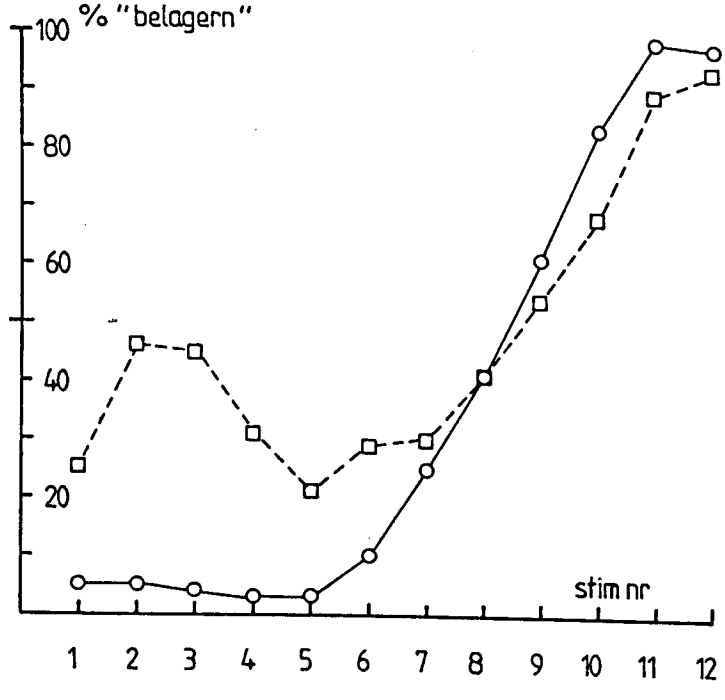


Fig. 3. Percentage stem-stress responses for "umlagern" (="belagern") in the series of 12 F0 peak positions combined with the original prefix-stress utterance of "Er wird's wohl umlagern."; original, sharply falling peak contour (continuous line, at each data point N=14x10+9x5=185), and slowly falling peak contour, transferred from the original stem-stress utterance (broken line, at each data point N=16x5=80).

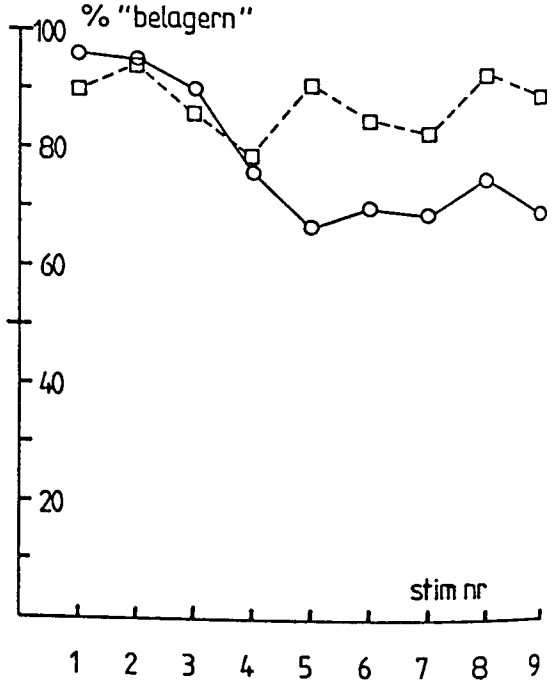


Fig. 4. Percentage stem-stress responses for "umlagern" (="belagern") in the series of 9 F0 peak positions combined with the original stem-stress utterance of "Er wird's wohl umlagern."; original, slowly falling peak contour (broken line, at each data point N=16x5=80), and sharply falling peak contour, transferred from the original prefix-stress utterance (continuous line, at each data point N=14x10+9x5=185).

DECLINATION AND SENTENCE INTONATION IN ITALIAN

CINZIA AVESANI

Scuola Normale Superiore - 56100 Pisa - ITALY

ABSTRACT

This work on Italian intonation was designed to test Cooper and Sorensen's declination model. In their model, one half of the downdrift is accomplished in one fourth of the utterance's temporal extent; all the peaks except the first are related in a linear fashion and the first peak is correlated with sentence length. My analyses of F0 contours do not support the Cooper and Sorensen model. There is no significant correlation between the first peak height and sentence duration. In addition, a linear declination slope fitted to the whole sentence or to the sentence peaks after the first, predicts acuurately only a small percentage of sentence peaks. I suggest that local pitch accents and final pitch fall account for Italian F0 contours better than global realization rules.

INTRODUCTION

According to Bolinger (/1/), the most widely diffused intonational phenomenon is the tendency towards a low pitch at the end of declarative sentences. This "running down" pattern has been given a variety of names, of which the best known is probably declination. Declination is said to arise from the property of speech fundamental frequency to gradually decline over the course of utterances (e.g. Cohen, Collier and 't Hart /2/; Collier /3/; Thorsen /11/; Gelfer /6/), or from a rapid downmotion at the very end of an F0 contour, while non-terminal portions do not necessarily display a downwards slope (Lieberman, Katz, Jongman, Zimmerman and Miller /8/; Umeda /13/).

In the former approach, phrasal intonational properties are described in terms of global falling trends; in the latter, in terms of (final) falling features with a sentence-level parsing function.

Cooper and Sorensen's model (/4/) is an example of the first approach, since it aims to capture single F0 peak values by means of a global rule which generates falling trends within an utterance domain (Top-line Rule). In this model one half of the downdrift is accomplished in one fourth of the utterance temporal extent. The first peak is correlated with sentence length and all remaining peaks are related in a linear fashion. According to Cooper and Sorensen (henceforth C&S), declination peaks are under the talker's voluntary control. The general theoretical claim is that of a statistical relation between utterance length and F0 as evidence of sentence-level global pre-planning.

However, other models of English sentence intonation account for downtrends without invoking pre-planning (e.g. Fujisaki /5/; Pierrehumbert /12/). Interestingly, these are base-line models: Fujisaki's is in terms of global declination trends, where a physiological "baseline" component constitutes the phonetic framework or constraint for local phonological entities (accents) to occur (see also Ladd /7/). In this model an exponential baseline decay is posited, which may be seen as representing damping effects of the speech motor implementation. In a similar "frame of reference" approach proposed by Pierrehumbert, target F0 peaks are scaled on an abstract declining baseline.

The point I want to make here is that the weak rate of decay in F0 peaks following the first one (P1), predicted by an exponential equation, is consistent with a physiological explanation of global declination alternative to C&S's cognitive model (Fujisaki /5/; Gelfer /6/): according to Gelfer, declination is the byproduct of F0 automatic modifications in response to decreasing subglottal pressure; in Fujisaki, declination is the effect of the dynamic properties of the speech control mechanism. Also, it is not inconsistent with the "local feature" approach proposed by Liberman and Pierrehumbert /10/, in terms of exponential declination as the global statistical effect of such different local events as phonetic "final lowering" and a set of phonological "downstepping rules" characterizing the medial portion of sentences.

Moreover, as compared to C&S algorythm, what characterizes no pre-planning models such as Fujisaki's and Liberman and Pierrehumbert's (L&P), is the crucial fact that they do not posit a covariance relation between P1 and sentence length, even if for different reasons.

Therefore, as an attempt to test the C&S hypothesis, this work focuses on the statistical relation between F0 peaks and sentence length, in order to see: first, whether a correlation exists in Italian bet-

ween the FO contour first peak (P1) and sentence length; second, whether the attested downdrift in Italian declarative sentences (e.g. Magno-Caldognetto, Ferrero, Lavagnoli and Vagges /10/) can be adequately described in terms of linear declination trends fitting the peaks which follow the first one as predicted by C&S model.

EXPERIMENT

Three male speakers produced four tokens of different sentences varying in length. Three sets of utterances of increasing length included: one-word sentences with five steps of length variation, i.e. ranging from one to five syllable length (e.g.: "Tom"... "Indicameli"); simple sentences with four steps of length variation, ranging from one-word tofour-words length (e.g.: "'Monica" ..."'Monica compera mobili rustici"); and complex sentences with five steps of length variation. These latter consist of two coordinated clauses including /e/ ("and") and /ma/ ("but") as coordinating conjunctions. Length variations were obtained by increasing both coordinated clauses: three steps of variation derived from a symmetrical increase of both clauses, which produced short-short, medium-medium, long-long durational patterns; two other steps were obtained asymmetrically varying each clause length, in order to get a medium-long and a long-medium durational pattern. Analyses of three out of four repetitions were made using a pitch extraction algorythm in the ILS analysis system at Haskins Laboratories.

To test the first prediction of the C&S model, i.e. that P1 is strongly related to sentence length, I measured the P1 height of all the sentences across all conditions. In order to minimize segmental effects on the first peak height, both one-word and simple sentences were produced also in reiterant versions, using /da/ as the syllable for mimicking natural utterances. Pearson's correlation coefficients were calculated between the first peak height and sentence length computed in number of syllables (see C&S /4/; Fujisaki /6/) for each group of one-word, simple and complex sentences of increasing length. I expect that the P1 height will rise as sentence length increases.

To test the second prediction of the model, i.e. that all the peaks except the first in a contour are related in a linear fashion, I measured all the high values occurring on stressed words for complex sentences only. A linear regression technique for fitting theoretically linear sets of data points was used to derive "best fit" toplines, choosing peak's time of occurrence to represent the independent variable and the peak's height in Hz as the dependent variable. Correlation coefficients provide an adeguacy measure for the linear relation of the data points. Levels of significance were assigned based on r values.

RESULTS AND DISCUSSION

First peak height.

One-word sentences. Table 1 shows that no significant correlation was found betweenword length and the first peak height in both natural and reiterant versions of words of increasing length. Correlation coefficients separately calculated on the production of individual speakers reveal strong intraspeaker differences: for one speaker only (MV) r is significant in both natural and reiterant utterances; for speaker GN r is never significant; for speaker MM there is a significant and positive correlation only as far as reiterant words are concerned.

Simple sentences. Essentially the same picture emerges from the results of the correlation test on simple sentences computed on all talkers where the correlation between P1 of the first word in the utterance and global sentence length does not reach the significance level of .05 in either natural or reiterant versions.

Simple sentences were divided into three subgroups according to the stress pattern of the initial words (e.g. /'Monica; Do'nata; Mari'lu/); as P1 cooccurs with stressed sullable, I tested whether stress position in the initial word had any significant relation with sentence length. Table 2 shows that initial, medial and final peak positions in the initial word actually yield different r values, but that is statistically significant.

Intraspeaker differences again arose from correlations computed separately on each talker. Speakers MV and GN show no significant correlations in both natural and reiterant speech, while r is higly signifivant for the third speaker (MM).

An observation can be made: if a significant correlation between P1 and sentence length is an evidence of a look-ahead strategy by the subject, at the word level speaker MV alone appeared to use it consistently in natural and reiterant speech. However, in the (simple) sentence domain, the same speaker does not show any effect of sentence length on P1; on the other hand, (anticipatory) effects related to sentence length did show up significantly, in a sentence domain,for one of the other speakers (MM) who showed non-significant effects of increasing length on P1 height of natural words.

Complex sentences. These sentences are composed of two coordinated clauses; hence, a P1 raising effect if any, might be induced either by the increasing length of the whole sentence or by the increasing length of the first clause. Table 3 provides correlation scores of statistical analyses computed between P1 and first clause length.

Correlation computed on all the speakers are non significant not only for P1 and whole sentence length but also for P1 and first clause length. Intraspeakers differences are confirmed: MV and GN show non-significant correlation, consistently with results from their simple sentences; talker MM shows a significant r value for correlation between P1 and whole sentence length, and a non significant one for correlation between P1 and first clause length.

These data skow that, when it occurs, the raising of P1 height is due rather to a global sentence length increase than to a first clause length increase.

As a first conclusion, it may be said that my results do not support C&S's predictions of a systematic relation between first peak values and sentence length. Intraspeakers differences were such that only one out of three speakers consistently showed higher P1 values as simple and complex sentence length increased. Moreover, interspeaker variations point out the unstable character of such an effect (see L&P /10/

Topline declination.

To test the hypothesis of a linear declination over the peaks which follow the first one, a simple linear regression was computed on all complex sentences varying in length. In this analysis I took into account all the peaks except the first one, according to C&S's formulation. The relative correlation coefficient was significant ($p < .001$), showing that complex sentences actually display peak downtrend, but the coefficient's low value ($r = -.294$) suggests that only a small percentage of the data peaks (8%) can be fitted by a straight line .

As complex sentence length was substantially varied systematically modifying each clause's length, I checked whether correlation coefficients varied because of the different durational patterns of sentences: regressionlines computed separately on sentences of differnt lengths show that differences in r values ranged from $r = -.230$ (for sentences with a long first clause and a medium second one) to $r = -.691$ (for sentences where both clauses were short).

In presence of complex sentences, a possible interpretation of these outcomes (e.g. C&S) is that topline declination is "reset" at the clause boundary, generating individual toplines for each clause. If this is the case, single linear regression lines separately computed for each clause on all their peaks except the first should give better fits than the global one. My results, however, do not support this hypothesis . First clause: computed on the whole set of sentences, the analysis gives non-significant r values; separately computed on all first clause tokens belonging to different durational patterns, the analysis provides quite different r values, which in itself is not a result that confirms C&S's predictions. Second clause: a linear regression line has a negative value and a significant correlation score ($r = -.238$; $p < .001$). However, it shows a lower r value than that obtained for the entire complex sentence.

These outcomes seem to suggest that the regular falling intonational patterns of the second clause mainly contribute to sentence level topline declination.

P1. based on regression lines computed on all peaks except the initial one on complex sentences, the P1 observed values are not systematically higher than the predicted ones. In one durational pattern only (long-long) the average of observed P1's is 11 Hz higher than the predicted first peak value.

As a consequence, linear regressions computed on all the peaks in complex sentence contours give better fits than regression computed excluding P1: the coefficients for all talkers on all sentences being respectively $r = -.385$ vs. $r = -.294$.

So, the outcome of no substantial drop between the first and the second peak couples with the L&P hypothesis of local final lowering as an effect contributing to declination.

Final lowering effects.

To test final lowering, new regression lines computed on all peaks except the last were calculated. Results give worse fits and flatter slopes compared with all peaks regression, as I would expect if final lowering substantially contributed to declination: on all sentences scores are respectively: $r = -.256$ vs. $r = -.385$; slope = -4.46 vs. -6.81 Hz per sec. Looking at the overall peak regression, it is final peak - as opposed to the first one - that is crucial in determing declination: this might suggest that sentence declination trends are mainly generated by local (initial and final) FO values.

Medial downdrift.

Moreover, excluding both initial and final peaks from linear regression computation, an almost negligible relationship emerges between the peaks' height and their position in an utterance; and the slope of the line reduces drastically to 2.7 Hz per sec. ($r = -.146$; $p = .002$).

CONCLUSION

Results suggest that the height of the first peak in a FO contour is not systematically related to sentence length for all speakers. Moreover, linear regression lines fitted onto all peaks in a contour except the first one give an account of topline declination that captures only a small set of the data.

Despite the fact that P1 is higher than the subsequent peaks, it does not lie well above the line connecting all the others, as predicted by the C&S model.

On the other hand, a weak decay in the medial portion of FO contours and a substantial effect of final peak lowering seem to point to an explanation of topline declination as statistically arising from local (initial and final) FO values, while the attested weak medial downdrift may be consistent with a physiological explanation of declination as the by-product of declining subglottal pressure (Gelfer et al. /6/).

Final lowering may be the result of the physiological relaxation of the articulatory system. As for the shape of the medial downdrift, the experiment was not designed to answer the question whether

local phonological rules (e.g. "downstepping") determine the shape of intermediate FO contours (L&P /9/). This will certainly be a line for future research, as Ladd indicates (/7/). The experiment aimed to collect evidence regarding Italian intonation which might allow to rule out the strong assumption of global preplanning in explaining falling intonational trends.

TABLE 1

One-word sentences

	Reiterant		Natural	
	r	p	r	p
gen.	.341	>.10	.214	>.25
MV	.939	<.025	.962	<.05
GN	.73	>.10	.465	>.25
MM	.8	<.05	.632	>.25

Table 1. Correlations between first peak height and sentence length.Gen.: on all speakers; MM, GN, MV: on single speakers.

TABLE 2

Simple sentences

	Reiterant		Natural	
	r	p	r	p
gen.	.304	<.10	.21	>.10
MV	-.136	>.25	-.03	>.25
GN	.309	>.25	.268	>.25
MM	.905	<.001	.866	<.005
/'---/	.264	>.25	.149	>.25
/-'--/	.16	>.25	.163	>.25
/--'-/	.471	>.10	.318	>.25

Table 2. Correlations between first peak height and sentence length. /---/ schematically represents the different stress patterns of the sentence initial words.

TABLE 3

Complex sentences

	P1/sentence		P1/first clause	
	r	p	r	p
gen.	.226	<.10	.093	>.10
MV	-.341	>.10	-.412	<.10
GN	.216	>.10	.169	>.10
MM	.504	<.05	.410	<.10

Table 3. Correlations between first peak height and sentence length, and between first peak height and the length of the sentence's first clause.

REFERENCES

/1/ D. Bolinger, "Intonation across languages". J. Greenberg, C. Ferguson & E. Mravcsik, eds., Universalsof human language, Standford University Press, 1978.

/2/ A. Cohen, R. Collier & J. 't Hart, "Declination: construct or intrinsic feature of speech pitch?", Phonetica, 39, 1982.

/3/ R. Collier, "Physiological correlates of intonation patterns", J.A.S.A., 58, 1975.

/4/ W. Cooper & J. Sorensen, Fundamental frequency in sentence production, Springer, 1981.

/5/ H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing". F.Ferrero, ed., Proceedings of the Fourth F.A.S.E. Symposium on Acoustics and Speech, Venice, 1981.

/6/ C. Gelfer, K. Harris, R. Collier & T. Baer, "Speculations on the control of fundamental frequency declination", Haskins Laboratories Status Report, 76, 1983.

/7/ R. Ladd, "Declination: a review and some hypotheses". C. Ewen & J. Anderson, eds., Phonology Yearbook, 1, 1984.

/8/ P. Lieberman, W. Katz, A. Jongman, R. Zimmerman & A. Miller, "Measures of the sentence intonation of read and spontaneous speech in American English", J.A.S.A., 77, 1985.

/9/ M. Liberman & J. Pierrehumbert, "Intonational invariance under changes in pitch range and length". M. Aronoff & R. Oehrle, eds., Language sound structure, MIT Press, 1984.

/10/ E. Magno-Caldognetto, F. Ferrero, C. Lavagnoli, & K. Vagges, "FO contours of statements, yes-no questions, and wh-questions of two regional varieties of Italian", J. of Italian Linguistics, 3, 1978.

/11/ N. Thorsen, "Sentence intonation in textual context - Supplmentary data", J.A.S.A., 80, 1986.

/12/ J. Pierrehumbert, "The phonology and phonetics of English intonation", PhD dissertation, MIT, 1980.

/12/ N. Umeda, "FO declination is situation dependent", J. of Phonetics, 10, 1982.

ACKNOWLEDGEMENT

I thank Haskins Laboratories for their friendly hospitality and continuous support during this research. My particular thanks go to Carol Fowler and to Mario Vayra.

IRREGULAR PERIODICITY AS A BOUNDARY CUE BETWEEN PHRASES

Christian Sappok
Seminar für Slavistik

Ruhr Universität Bochum, 4630 BOCHUM, BRD

ABSTRACT

The voiced portions of natural speech exhibit a tendency towards a certain amount of irregularity concerning the F0 values of successive periods. Is this due to the specific functioning of the voicing apparatus, or is it one of the features of the signal with a definit function of delimiting phrasing units? To gain evidence for one of these alternatives a test was constructed containing manipulated F0 declinations and gradual irregularity at the position of virtual boundaries. The results of perceptual experiments confirm the systematic status of both of the phenomena in the structure of the intonational contour.

It is a well known experience in the field of manipulation and synthesis of speech signals that a sequence of identical periods or a sucession of periods with regularly changing parameters are percieved as unnatural. Already introducing a small amount of irregularity, called jitter in the case of FO-deviations, helps to make the signal sound more natural. A tendency "to vary randomly from the general intonational trend line" is characteristic to all instances of voiced speech (HILLER, LAVER and MACKENZIE 1984,59 ff.)

There is some evidence for the hypothesis that irregularity is not only charactaristic of natural utterances, but that it is more often found in specific parts of the connected speech chain than in other parts. CARDOZO and RITSMA (1968) found that irregularity is more readily perceived when it occurs in the center of a stationary signal; this observation leads to the conclusion that the phenomenon is more readily expected in the non-central parts of the signal. Leaving the field of artificial pulse trains aside, we would say that the phenomenon is a substantial part of the speech signal, fulfilling a systematic function in the subdivision of the natural utterance into phrases; the problem of prosodic demarcation of certain constituents and its semantic functions have been systematically described by PEŠKOVSKIJ (1914).

There are two sorts of boundary signals in the phrased natural speech signal: the modified continuation of the parameters valid for the phrase central parts (F0-declination, decreasing intensity, lengthening; this last mentioned phenomenon has been described extensively by LEHISTE 1970 and KRIVNOVA 1983), or in the introduction of a specific signal as the pause which is without function in the central parts of the intonation contour.

Are these means alternative possibilities of realizing an abstract boundary mark between consecutive chains of the speech signal with the status of phrases? Or do they form a specific set of features characteristic for the non-central parts of the well-phrased utterance? If the former alternative is true, we would expect that gradually enforcing the change of prosodic parameters characteristic of a boundary would gradually reinforce the subdivided profile of the respective phrase. The latter alternative which gives these parameters the status of linguistic features with strictly delimited values, motivates the expectation that no such linear dependency existsbetween the strength of deviation between central and peripheral parameters on the one hand, and the interpretation of this deviation as distinctive on the other exists.

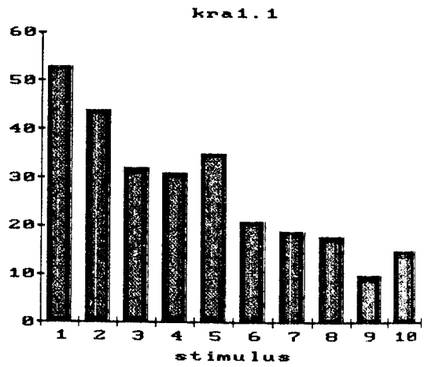
To investigate the possible solution of this alternative modelling of phrasing units, a series of tests was constructed on the basis of natural speech production and the controlled manipulation of pitch periods. No pitch transformations of periods were applied,

rather the periods of the original utterance were segmented and combined anew with the aim of imitating the internal phrasing boundary.

A chain of particles such as Russian "da da da ..." and "nu, nu, nu ...", which allow for different intonational realizations and, according to these differences, more than one pragmatic interpretation, were produced by male native speakers , once as a single chain with reassuring funktion: "nu nu!" (= go ahead), once in a internally phrased form "nu nu! nu nu nu!", thus reduplicating the mentioned function. The resulting utterances were analyzed and manipulated in one of two ways:

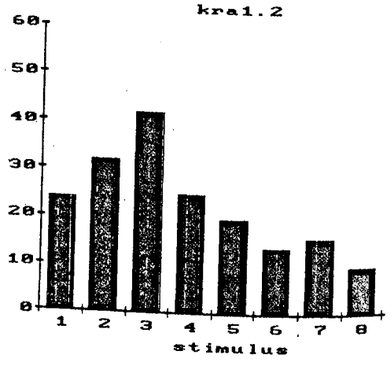
The first type of manipulations imitates at the final portion of the second "nu" of the internally not-phrased version the values of the internally phrased version, with increasing grade of intensity: F0 is gradually lowered or raised, and a lengthening was achieved by reduplicating periods. Alternatively, irregularity is implemented at exactly the same place of the virtual boundary and tested as to its subdividing effect.

In the chart kra1.1 the stimuli 2 to 5 contain a declination of F0 from the level of 208 Hz to 202, 200, 198 and 196 Hz respectively, in each case spread over the last ten periods of the second "nu" of internally non phrased series of five "nu"-syllables. The stimuli were presented in randomized order to native speakers (students of the Moscow State University, Institute of General, Comparative,Historical and Applied Linguistics, directed by Prof. L.V. Zlatoustova, to all of whom I wish to express my sincere gratitude), in a forced choice test, with the instruction to mark each of the perceived stimuli as either non phrased or internally phrased. The reactions to the stimuli with F0 declination show a marked shift in the interpretation, with a decreasing number of positive answers (= + internally phrased), but not in strict linear dependency. The same holds true within the group of stimuli 6 - 9, being identical as to the form of the F0 declination, but with an additional lengthening of 20 ms within the non - modified part of the last syllable. Note that even the unphrased original stimulus was not percieved as unphrased in 100 % of the reactions. In general, the use of natural speech production with the task of identifying communicative speech functions is much more complicated than

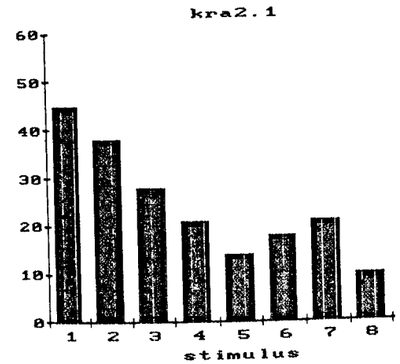


the use of meaningless pulse trains, or at least differ to a large degree; HESS 1983,79 reports suggestions that the difference limen for the audibility of F0 changes is higher in natural speech.

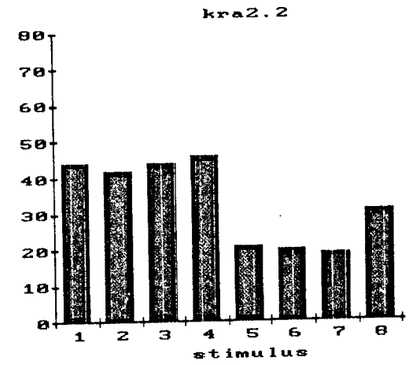
In the next series of manipulations (kra1.2, stimuli 1 to 4) we realized the same F0-declination as in the former case, i.e. in kra 1.1, but in upward direction. The first and the last step of this manipulation leads to a considerable increment of positive reactions (i.e., plus internal-phrasing interpretations in 36 cases out of 60); the subdividing effect is weakend in the case of stimulus 2 and 3, a fact that is not easy to explain. Maybe the over-all contour of the unmanipulated natural signal has a slight internal contour within each of the syllables. This being the case, changing the F0-values in upward direction has a levelling effect with respect to this internal contour, thus reducing the uncertainty of interpretation. The addition of 20 ms, again, results in an increasing number of YES - interpretations.



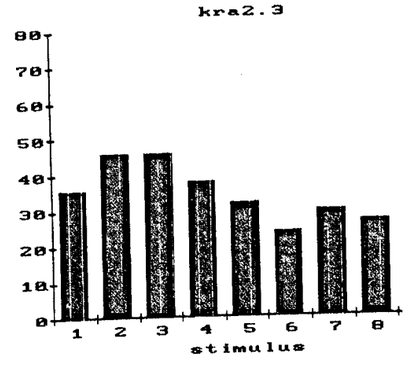
To what extent are these results comparable to similar manipulations in the field of irregularity, provided that the location and the graduation of the manipulated stimulus parts remain the same? To answer this question, deviating periods were inserted in the chain of 208 Hz periods in randomized order, beginning with 204 hz and increasing this deviation by 2-Hz steps. The reactions show an increasing tendency towards the phrased interpretation, reinforced by an additional lengthening of 20 ms, this time without a regular shift, cp. kra2.1.



This kind of manipulation was repeated in kra2.2, this time with a upward deviation throughout, but with an identical randomized sequence. This type of irregularity shows a marked rise in the positive reactions as to the presence of an internal boundary, which again is reinforced to a remarkable degree in combination with lengthening; but in this case without any clear differentiation within the sequence of different degrees of deviation. This can be interpreted as evidence for the fact that jitter is interpreted as a boundary signal irrespective of the distance between the deviant periods from the surrounding level.



In the series kra2.3 the periods near the virtual boundary were combined in the following manner: Up- and downward deviations alternate in randomized succession, the distance between the highest and the lowest deviation corresponding in stimulus 1 to 4 to the unilateral deviation of the respective sequence in kra2.1 and kra2.1. There is a difference between the responses to the unlengthened stimuli and the lengthened (compare the responses to stimuli 5 to 8), but again no marked influence caused the prominence of the deviation. If we take it for granted that this type of irregularity comes closest to the jitter of natural speech production, we can conclude that the function of irregularity in the process of boundary marking is relatively independent of reinforcing influences of other boundary signals.



The tentative conclusions on the basis of these results may be formulated as follows:

1. The stimuli under consideration correspond to simpl, yet natural utterances, revealing a clearly marked pragmatic function, with the possibility of reduplication of contour and function, the latter case showing an internal phrase boundary.

2. The decision of interpretation in unclear cases, being the result of minimal manipulations, is influenced to the same degree by F0 - declination and by irregularity of subsequent pitch periods. This effect is in both cases reinforced by additional lengthening of the already manipulated syllable.

3. In the case of irregularity there is a remarkable absence of interdependence between the degree of deviation and the bias of interpretation within one and the same type of boundary signalling.

Mainly this last observation gives evidence for the hypothesis that jitter does indeed havea fixed place in the system of boundary signals in natural speech. It reveals its function in position near the phrase boundary and interacts in a systematic way with the other boundary cues. Further experiments have to show whether this is a characteristic function only in positions adjacent to the boundary, or whether also fullfledged lexical items on the periphery of intonation contours, bearing phrase accent ("frazovoe udarenie" according to NIKOLAEVA 1982), but not phrasal stress ("akcentnoe vydelenie" according to NIKOLAEVA ibid.), can be marked by irregularity.

Literature:

CARDOZO B.L., RITSMA R.J. (1968): On the Perception of Imperfect Periodicity. IEEE Transactions 16 (2), 159-164.

HESS W. (1983): Pitch Determination of the Speech Signals. Algorithms and Devices. Berlin et al.

HILLER S., LAVER J., MACKENZIE J. (1984): Durational Aspects of Long-Term Measurements of Fundamental Frequency Perturbations in Connected Speech. Work in Progress Nr. 17, Dept. of Linguistics, Edinburgh University.

KRIVNOVA O.F. (1983): Dinamika tempa v sintagme. Fonetika-83. Materialy k X. meždunarodnomu kongressu foneticeskich nauk (avgust 1983 god, Utrecht, Niderlandy), 102-114.

LEHISTE I. (1970): Suprasegmentals. Cambridge, Mass.

NIKOLAEVA T.M. (1982): Semantika akcentnogo vydelenija. Moskva.

PEŠKOVSKIJ A.M. (1914): Russkij sintaksis v naučnom osveščenii. Moskva.

VARIABILITY OF PHONEMES IN SPOKEN RUSSIAN

Geilman Natalja Iosifovna

Chair of Phonetics, Dpt. of Philology,
Leningrad State University,
Universitetskaya nab, II, 199164, USSR

Abstract.

The segmental characteristics of Conversational Literary Russian are reviewed. Various modifications of different classes of phonemes caused by the loosening of their articulation and their acoustic correlates are described. The positions favourable to the modifications are analysed. Modifications of phonemes in 3 varieties of Spoken Russian /territorial dialects, urban popular speech and Conversational Russian/ are compared. There exists in them a number of spontaneity-due modifications common for these varieties. Thus we conclude that the difference between Conversational Literary Russian and Standard /Literary/ Russian proper is caused by spontaneous character of the former and not by the difference in their phonetic systems.

Introduction.

There existed 3 main varieties of Spoken Russian /SpR/ up to now used as spontaneous communication means: 1/ substandard territorial dialects, 2/ substandard urban popular speech, 3/ Standard /Literary/ Conversational Russian /CR/. Nowadays the latter is rapidly expanding owing to general secondary education, the expansion of mass-media means and the growth of the cultural level of various strata of society. That is why CR should be studied properly and with great care. Firstly, such a study could help us to see the main tendencies of the Russian language development because it is through CR that dialectal and urban popular speech forms penetrate into Standard/Literary/ Russian /SR/. Secondly, the comparison of CR with the two other varieties of SpR on one hand, and with SR on the other, would enable us to dicover spontaneous traits proper and to decide whether the system of linguistic units of CR is independent on that of SR. The phonetic study and the study of segmental units in particular may be of great importance because it is these units that when pronounced are minimally controlled by a speaker and it is in the field of segmental phonetics that substandard and common spontaneous traits can be most apparent.

The phonetic systems of territorial dialects have been explored for decades. There recently appeared a number of works on urban popular speech. As to the phonetic characteristics of CR, they remain beyond the scope of experimental studies. That is why the present investigation has been undertaken.

Discussion of investigation.

The inventory of linguistic materials consisted of 3 sections. 1/ 15 fully transcribed spontaneous dialogues which made up 3 hours /60,000 phonemes/ used for the investigation of various modifications of segmental units in CR. 2/ Oscillogrammes and spectrogrammes of the fragments of the dialogues /37-min. duration, more than 10.000 phomenes/ containing the most distinctly pronounced modifications used for the study of their acoustic correlates. 3/ 16 fully transcribed monologue fragments /1-1,5 min.each/ extracted from the dialogues performed by the speakers of all the 3 varieties of SpR used to discover spontaneous features common to them.

The results of the investigation of the segmental characteristics of CR described earlier /1/ showed that 18% of consonants and 8% of vowels /in the most frequently used words of CR 30 and 20% respectively/ are subjected to various modifications, i.e. pronounced not as specified by SR pronunciation rules. 5% of consonants and 3% of vowels /18 and 7% in frequent words/ are elided.[1]

The analysis of stability of various phonemes and classes of phonemes has shown that more "lax" voiced consonants are modified and elided more often than the voiceless ones, soft or palatalized /marked/ more often than hard or velarized, more frequently used non-sonorants more often than sonorants. Affricates and liquids are the two most unstable

[1] Note that the terms "modification", "deviation", "ellipsis" are used here for the sake of conveniency as it is SR that is used for reference, CR having pronunciation rules of its own.

groups, 48 and 40% of them modified respectively in CR. Nasals and vibrants are the two most stable groups, 12 and 9%. Plosives are modified more often than fricatives but the latter are more frequently elided. Dorsals are modified more often than labials and velars, palatal /j/ is most frequently elided.

There exists a certain correlation between stability, information load and perception of consonant classes: the more informative classes are more stable in speech. The more stable classes tend to be better perceived in noise /2/.

There exists a certain correlation between stability of sounds and frequency of occurrence: the phonemes /a/, /i/, /j/ are the most frequent ones in Russian /3/ - the former two being most unstable among vowels and the latter among consonants. The high frequency of occurrence of a morpheme or a word containing a given sound is also a favourable condition of its modification or ellipsis.

We believe that stability of segmental units is hierarchically organized: more frequent linguistic units /phonemes, morphemes, words/ have less information load due to their frequent usage in speech. They are modified or elided more often by a speaker without any consequence for speech understanding. When the process of speech perception is carried out in impaired conditions /CR perception can be regarded as such due to the large quantity of modified and elided segments causing considerable distortion of segmental structure of the word/ the rarer units /classes of units/ are more stable in realisation and thus they are better percieved by a listener which still increases their information load.

The analysis of reasons of sound changes in CR has shown that the majority of them are caused by lax articulation which explaines 66% of the consonant changes and 46% of them for vowels. Lax articulation leads to spirantization of stops /10% of stops in the dialogues are subjected to partial and entire spirantization/, weakening of nasal resonance of nasals /m,m',n,n'/ and their substitution for oral nasalized[b̃,b̃', d̃,d̃'], lisping of the spirants but for /v-v'/, /f-f'/, vocalization of liquids and vibrants and /v-v'/, ellipsis of the weakest /f-f'/. Voicing of voiceless, i.e. articulation of "lax" instead of "tense" consonants can also be attributed to lax articulation. For vowels lax articulation leads to greater qualitative reduction of /a/ /pronunciation of [ɤ] instead of [ʌ] or [ə]/, the appearance of qualitatively reduced /u/ and /ɞ/ in the unstressed syllables /pronunciation of [ɤ] and [ɞ] instead of [u] and [ɤ] instead of [ɞʊ]/.[1] Besides, there exists a shift of vowels towards the center: instead of more close and more front [i,e] more open and more back[ɪ,ɛ] are pronounced; more close and more back[u,o] are replaced by more open and more front[ʏ,ɔ].

Analysis of positions favourable to articulation-loosening has shown that the majority of the cases occur in the intervocalic position where a consonant is necessarily weakened /vocalized, spirantized, voiced etc./ as a result of it's articulation adjustment to that of the surrounding vowels. The mid-word position is also favourable for the process. The positions favourable to lax vowel-articulation have been studied elsewhere /4/.

Analysis of acoustic correlates of the consonants' modifications has shown that the acoustic changes can be grouped as follows: F-structure changes, noise components' changes, tonal components' changes, changes of duration. For spirintization the appearance of high-frequency noise at the place of stop is characteristic. Spirantized consonants in CR differ from the analogous consonants in SR. In SR they occur in intervocalic position only and have spirant-phase duration up to 50% of their entire duration /5/. In CR they may occur in any position and mostly have no stop at all: 58% of spirantized plosives and 85% of spirantized /č/ turn into full spirants in CR.

Noise resonance weakening is reflected in the formant structure, i.e. in the weakening of nasal formant.

Misp pronunciation of spirants is realized as a substitution of /s,s',z,z'/ with a round opening for flat-opening consonants. Lisped /š,ž/ have a loosened second /back/ focus parallelled by very strong lip-rounding. Acoustically lisping is manifested through the lowering of upper noise-limit or high intensity noise-limit as compared to that of normally realized consonants in analogous positions.

When vocalized, /l-l', v-v'/ differ from non-vocalized by peak intensity shift from the consonant to the neighbouring vowel: peak intensity of the vocalized consonants is in most cases 0.76-1.00, of the "normal" ones - 0.25-0.50 of that of the neighbouring vowel. The paired comparison of peak intensity of vocalized and normal consonants in analogous phonetic positions has demonstrated that /l-l'/ and /v-v'/, when vocalized, have in statistically significant number of cases lesser peak intensity shift to the neighbouring vowel; with /r-r'/ this regularity also shows but shifts in the opposite direction. This can be explained by the presence of several high intensity peaks

[1] Traditionally in Russian phonetics [ɤ] is used for very short lax mid-open back-retracted vowel, [ɞ] - for very short high front vowel, [ɞʊ] - for a diphtongoid with central back-retracted initial and front middle and final stages.

corresponding to the number of flaps with normal vibrants while vocalized consonants have more smooth picture.

Spectral analyses of vowels[1] has shown that three allophones of /a/ /the stressed [a] and two unstressed ones - [ʌ] and [ɤ]/ can be described as follows: [ɤ] has FI lower and FII higher than [ɑ], [ʌ] has FI lower than [ɑ] and higher than [ɤ], FII higher than [ɑ] and lower than [ɤ] which testifies that it is more close and more front than [ɑ].

The vowel pronounced in the unstressed syllables where "o" is spelled is "o" - like vowel: FI and FII are lower than for [ɤ] or [ʌ] which are pronounced here in SR. The higher position of FI and FII for [ɪ] shows that the vowel is more open than [i] and somewhat back-retracted.

The comparison of phonemes' modifications in CR and SR shows that the inventory of these modifications coincides almost entirely. The fact that such hesitation-due modifications as elongation, vibration, i.e. repetition of one and the same sound many times, and such emphasis-due modifications as intensive pronunciation etc. are mentioned nowhere is caused evidently not by their absence in SR but by the absence of the phonetic studies of the spoken form of SR. CR and SR thus differ only in the inventory of phonemes udergoing this or that modification and in the inventory of positions where this or that modification occurs. Both inventories are usually more versatile in CR. Besides, modifications of phonemes in CR are more clearly pronounced: in CR consonants are fully, in SR - only partially voiced, vocalized, spirantized etc. All this leads to the extending of variability limits in CR.

The comparison of phonemes' modifications in different types of SpR has shown that out of 15 the most essential dialectal vocalic specific features and 15 consonantal ones /6/ 10 and 9 respectively were also registered in CR. Out of 54 pronunciation deviations found in the 3 variaties of SpR 14 were registered for the statistically significant number of speakers.[2] For consonants these are lisping of the spirants, "unnormal" voicing and devoicing, spirantization of stops, vocalization of liquids, vibrants and /v-v'/, pronunciation of not enough palatalized /č/, pronunciation of not soft enough consonants before front vowels. For vowels these are "unnormal" allophones of stressed vowels /[ʊ,ɔ,ɪ,ɤ] instead of [u,o,i,a]/ and those of the unstressed ones /[ɤ,ɪ] instead of [ʌ,i]/, complete

[1] The spectral analysis was made by E.I.Oshujko.
[2] This work was carried out together with E.V.Andrjushchenko.

vowel reduction, pronunciation of [o] in formwords when unstressed. Weak consonants and vowels and reduced forms of frequent words should also be mentioned.

If for dialectal speech the presence of only some more or less regularly realized traits is characteristic, CR is marked by a large number of pronunciation peculiarities each appearing sporadically. Being of common spontaneous origin, they can coincide with the most versatile dialects' traits. But the phonetic system of CR being based on that of SR, these peculiarities can not be realized consistently. The appearance of common spontaneous traits in all the varieties of SpR gives the way for dialectal traits to penetrate into CR /either directly or through urban popular speech/ and then into Literary SR proper. That is how spirantized /š/ pronounced now more and more often as [s'] /7/ penetrated into SR.

Conclusion.

The analysis of segmental units' modifications in CR has demonstrated that the difference between CR and SR manifesting itself through the extension of the limits of allophonic variance in the former should be explained not by their systematic difference but by the spontaneous character of CR. This assertion is testified both by the identity of the inventory of allophones and mainly by the inventory of the functional units, i.e. phonemes, in CR and SR. This contradicts to the opinion that "CR can be regarded as a particular language" because "in is a particular s y s t e m having the specified inventory of units and the specified laws of their functioning; this system is opposed to that of SR within the limits of Literary Russian" /8/. The study of the spoken form of SR where the appearance of the same phoneme modifications, however minor in quantity, could be expected, would help to ascertain our conclusion.

Modification of phonemes in CR leads to the indefinitness, uncertainty of the characteristics of the sound units which is caused by the high speech tempo and becomes possible thanks to the great role of context /9/. That's why there appear in CR the allophones of different phonemes coinciding in their sound form, for example [ɸ] can appear both as a result of /t/ - spirantization and of /s/ - lisping, [ɤ] can be an unstressed allophone of /a/, /u/ and /ɞ/ etc. How can a listener accomplish phoneme identification of these segments which should be attributed to "non-full type of pro-

nunciation segments"[1] Generally, there is no necessity for a listener to produce their phoneme identification at all because "using redundancy he can recognize a word by its very general contour that is created by its rhythmic structure and by some cue sounds of its entire sound structure... A word can be even fully reconstructed from the context". Therefore "it is a profound error to suppose that each segment which can be singled out of the word should be directly attributed to a definite phoneme" /11/. The recognition of the word performed, the phoneme identification, if necessary, is easily accomplished by a listener because "a human being when working as a recognizing mechanism can identify punctually one and the same sound stimulus as different phonemes and various sound stimuli - as one phoneme" /12/. The impossibility to recognize a strongly "destructed" word, i.e. the word with distorted rhythmic structure or the word with the stressed vowel reduction etc., leads to mishearing or asking for repetition. These two can also be caused by homophony that can not be solved by the context, a word having become homophonous to another word as a result of distortion. Asking for repetition appears only when a communicatively significant word can not be recognized. Both the "distortion" of non-informative words and their homophony are paid no attention to by a listener. The cases when communicatively significant words cannot be recognized are extremely rare: in the 10 dialogues analysed from that point of view there were found only one mishearing and nor askings for repetition. This shows that in spontaneous speech the loosened control of a speaker over the outer form of expression is differential: strongly distorted are uninformative /frequent/ units /phonemes, morphemes, words/. As for the number and degree of distortions of communicatively significant parts of the text, they should stay within definite limits which a speaker would never trespass in fear of disturbing successful communication.

[1] "The full type of pronunciation provides a possibility a l w a y s to determine the phoneme structure of a word. Non-full type vice versa requires for this purpose proper context or situation" /10/.

References.

1. V.M.Beljavskij, N.I.Geilman, L.P.Scherbakova. Style, tempo and segmental characteristics of speech. - In: Experimental phonetic analysis of speech. Problems and methods. No.1. Leningrad, 1984, p.39-50. /In Russian/; N.I.Geilman. Consonantism. Ellipsis., N.I.Geilman, E.I.Osujko. Vocalism. - In: The Phonetics of Spontaneous Speech. Leningrad, 1987. /In Russian/.
2. N.I.Geilman. Stability, informativeness and perception. - In: Hearing and speech in norm and pathology. No.5, Leningrad, 1982, p.15-20. /In Russian/.
3. L.V.Bondarko, L.R.Zinder, A.S.Stern. Statistical characteristics of Russian speech. - In: Hearing and speech in norm and pathology. No.2, Leningrad, 1977, p.3-16. /In Russian/.
4. Russian Conversational Speech. - Moscow, 1973, p.41-61. /In Russian/.
5. L.V.Bondarko, L.A.Verbitskaja, L.R.Zinder. The dependance of temporal characteristics of consonants upon their phonetic position. - Problems of radioelectronycs. Series XI. Conductor communication technique. 1960. N.3, p.126-127. /In Russian/.
6. L.A.Verbitskaja, L.V.Ignatkina, N.F.Litvachouk et all. Regional traits of Russian speech /on the phonetic level/. Leningrad University Report /Vestnik LGU. 1984, No.8, p.71-80.
7. R.F.Paufoshima. About pronunciation of /č/ in Literary Russian. - In: The Development of Phonetics of Contemporary Russian. Phonological Subsystems. - Moscow, 1971, p.268-269. /In Russian/.
8. Russian Conversational Speech. - Moscow, 1973, p.22-23.
9. L.V.Bondarko. The Phonetic Description of Language and Phonological Description of Speech. - Leningrad, 1981, p. 144. /In Russian/.
10. L.V.Bondarko, L.A.Verbitskaja, M. V.Gordina et all. Styles of pronunciation and types of pronouncement. Problems of linguistics /Voprosy Jazykoznanija/, 1974, No.2, p.65. /In Russian/.
11. L.R.Zinder. The real flow of speech and the reconstruction of phonemic structure of a word. - In: Language Theory. Methods of Language Teaching and Language Studying. - Leningrad, 1981, p.105. /In Russian/.
12. L.V.Bondarko. Useful features and hierarchical organisation of phonemic classification. - In: Sound Structure of the Language. - Moscow, 1979, p.24-25. /In Russian/.

COMPUTER-ASSISTED EXAMINATION OF THE PHONETIC VARIANCY OF SPONTANEOUS SPEECH /A PRELIMINARY REPORT/

LILIANA MADELSKA

Adam Mickiewicz University
Institute of Linguistics
Poland, Poznań, Marchlewskiego 124/126

ABSTRACT

Tape- recorded texts have been transcribed by means of the sequential transcription of phonetic variancy. The transcription, adjusted to the computer keyboard, includes information on: 1. a phonetic transcription of a tape-recorded textes, 2. the pronunciation model adopted for the purposes of this work, 3. certain orthographic phenomena, 4. some other paralinguistic phenomena.

On the basis of the adopted transcription and assisted by the IBM compatible we have been able to develop a dictionary of phonetic realizations of all the text words and tables of phonetic realizations of allophones and allophonic clusters.

DESCRIPTION OF THE MATERIAL

The material examined consisted of tape recordings of spontaneous, casual conversations with students at Adam Mickiewicz University. The recordings were made under good acoustic conditions. Texts of conversations with 30 speakers were analysed. Each text was about 2000 words long. In total, over 66 000 words were analysed.

PHONETIC TRANSCRIPTION

All the material to be analysed was transcribed phonetically. The accuracy of the transcription of particular realizations depended on the auditory sensitivity of the transcribers. In our transcriptions we used over 100 different signs /symbols, letters/ to denote allophones discriminated in the speakers pronunciation.

The term allophone as used here means a set of homophonic sounds; i.e. they differ only with respect to insignificant phonetic features. With respect to specific phonetic realizations, allophones are at the first level of abstraction, and phonemes are at the second level of abstraction. We are aware of the fact that even the most accurate phonetic transcription cannot truly reflect the reality but only interpret it. The interpretation depends not only on the predispositions of the phonetician, his/her hearing habits, but also on the adopted manner of notation - this is mentioned by numerous linguists(D.Abercrombie/1/)In the texts under analysis we came across fragments which differed with respect to distinctness - sometimes entire phrases were pronounced very indistinctly, rapidly, "mumblingly" - hence it was not possible to transcribe all the material by means of typical phonetic symbols. Only fragments which were pronounced in a relatively careful manner were included in the analysis of phonetic variancy. In the remainder of the analysed material two degrees of "obliteration" or articulatory carelessness were distinguished. Such fragments /1. indistinct, 2. incomprehensible/ were also analysed, yet since they were treated as separate entities they will not be discussed here.

The generally accepted principles of phonetic transcription were used to account for fragments considered distinct.

THE PRONUNCIATION MODEL

By "phonetic variancy" we mean different phonetic realizations that occur in phonetic phrase (e.g. between phonetic pauses;M.Steffen-Batogowa/4,p. 27 /).

In order to examine phonetic variancy, it appeared necessary to adopt a uniform pronunciation model as a point of departure for the description of specific pronunciations of the informants. /In principle, it agrees with the Polish orthophonic norm, cf. Madejowa, M./3/. When different realizations are possible, however, only one had to be selected.

The model is uniform within phonetic phrases; therefore particular words can have some collaterant forms, that agree with the pronunciation model e.g. [jɛst sam] but [jɛzd̪ gɔ̃a] or [jɛst; irɛk]. The source of "differentness" is a specific phonetic position. Such realisations, compatible with the model, are not regarded as indications of phonetic variancy.

THE SEQUENTIAL TRANSCRIPTION

Before the computer could be used for phonetic analysis, the phonetic transcription had to be modified to suit the QWERTY

keyboard. We used letters and other symbols occuring in the EBCDIC code, the standard code of all computers.

Some of the following principles were observed while developing the sequential transcription:

1. Letters and diacritical marks occur in one line of the text.
2. Since the letters available on the keyboard are fewer than the number of allophones used in the transcriptions of phonetic realizations, it was necessary to introduce certain diacritical marks.
3. One allophone is represented by one symbol; the symbol begins with a letter of the alphabet followed by diacritical marks.
4. The transcription we propose whenever possible resembles the Slavonic, international or orthographic transcriptions so that it is easy to read the text.
5. For technical reasons only upper case letter are used.
6. Other symbols used denote paralinguistic and other phenomena /features/ which accompany pronunciation such as laughter, stuttering, indistinct pronunciation, incomprehensible pronunciation. These phenomena will not be discussed in this article.

Table 1 includes the set of 70 allophones of the pronunciation model transcribed both in the international alphabet and the sequential alphabet.

The phonemes are marked by numbers that indicate the alphabetical order, adopted in "The Dictionary ..." and "The Tables of phonetic realization ...".

TABLE 1. The international and the sequential transcription of allophones of pronunciation model.

	/API/	{seq.}	
1.	ə	A	
2.	b	B	
	bʲ	B,	
3.	t͡s	C	
	t͡sʲ	C,	
4.	t͡ɕ	C"	
5.	t͡ʃ	C.	
	t͡ʃʲ	C.,	1
6.	d	D	
	dʲ	D,	
	d	D.	
7.	d͡z	Z1	
	d͡zʲ	Z1,	
8.	d͡ʑ	Z1"	
9.	d͡ʒ	Z1.	
	d͡ʒʲ	Z1.,	
10.	ɛ	E	
11.	-ɛ	E-	2
12.	f	F	
	fʲ	F,	
13.	g	G	
	ɟ	G,	
14.	x	X	
	ɕ	X,	
	ɦ	H	
	ɦʲ	H,	
15.	i	I	
16.	j	J	
17.	ɲ	J=	
18.	k	K	
	c	K,	
19.	l	L	
	ɫ	L*	
	ʎ	L,	
20.	w	W	
	ʷ	W*	
21.	ẘ	W=	3
22.	m	M	
	mˠ	M*	
	mʲ	M,	
23.	n	N	
	nˠ	N*	
	nʲ	N,	
	n	N.	
24.	ŋ	N4	4
25.	ɲ	N"	
	ɲ̟	N"*	
26.	o	O	
27.	p	P	
	pʲ	P,	
28.	r	R	
	rˠ	R*	
	rʲ	R,	
29.	s	S	
	sʲ	S,	
30.	ɕ	S"	
31.	ʃ	S.	
	ʃʲ	S.,	
32.	t	T	
	tˠ	T*	
	tʲ	T,	
33.	u	U	
34.	v	V	
	vʲ	V,	
35.	ɫ	Y	
36.	z	Z	
	zʲ	Z,	
37.	ʑ	Z"	
38.	ʒ	Z.	
	ʒʲ	Z.,	

Remarks on Table 1.

1. Some allophones, e.g. [t͡ʲ], [d͡ʒʲ] could occur in between words or in words of foreign origin, e.g. [d͡ʒʲi͡ʃsy].
2. The oral pronunciation of {E-}, e.g. "idę"-{IDĘ-} has been adopted as an equivalent of the orthographic final "-ę". The "minus" sign after {E} signals merely that we have an allophone, an equivalent of the orthographic "-ę". Thanks to this notation it was easy to calculate whether in this position we had the oral or nasal pronunciation. This also permitted the discrimination of such pairs as e.g. "chce" and "chcę" in the "Dictionary ...".
3. Since in the transcriptions of phonetic realizations nasal vowels were written asynchronically, making a distinction between hard and soft nasal resonance, the symbol {W=} was introduced into the phonological system to denote a hard nasal element and the symbol {J=} to denote a soft nasal element, e.g. "koński" was transcribed as {KOJ=SK,I} and "kąski" as {KOW=SK,I}.
4. The retention of phonemes [n] and [ŋ] in the system permits discriminating between such pairs as "błonka"-[bwɔnka] and "błąka"-[bwɔŋka].

DIACRITICAL MARKS WHICH EXPAND THE ALPHABET OF THE PRONUNCIATION MODEL

The number of allophones in the pronunciation model was strictly defined while the number of allophones which occurred in the transcriptions of specific phonetic realizations was much greater and could not be predicted. For example, while transcribing our test material we heard several times a sound described as bilabial "w" - this articulation does not belong to the pronunciation model and is rarely described in works devoted to the phonetics of modern Polish. On the other hand, we have reserved a symbol for [ʔ] - yet no occurrence of this allophone has been detected. In Table 2 we give a set of diacritical marks /with examples of their use/ which permit expanding the alphabet of the pronunciation model. They denote certain phenomena or phonetic features and with the exception of [ə] and [ʔ] constitute a complete whole together with the preceding letter. Owing to this solution we get an open set of symbols which can be freely modified to serve various purposes.

TABLE 2. Symbols expanding the alphabet of the pronunciation model.

Symbol	Meaning
ə	shwa
ʔ	laryngeal occlusion,
+	nasalplosion
4	a change of the place of articulation, e.g. {V4} - a bilabial fricative /the typical Polish [v] is labio-dental/
-	minus after a consonant denotes aspiration, e.g. {T-AKA}
:	prolonged articulation, e.g. {TA:KA}
;	an element of geminate, e.g. {POT; TYM}
%	a vocalic consonant/related to the elision of a vowel/
?	weakened articulation, e.g. {XC"AW?A}
,	softening, e.g. {M,JAW, IS"C"}
#	devoicing, e.g. {FS.YSTKO#}
5	equivalent of "complex" notation; combines two preceding symbols, thus expanding the vowel set, e.g. [ɛ̯̃] - {AE5}.

SEQUENTIAL TRANSCRIPTION INCORPORATING PHONETIC VARIANCY

Texts transcribed phonetically were entered in the computer memory by means of the sequential transcription incorporating phenomena of phonetic variancy. Thanks to their incorporation all differences between phonetic realizations and the adopted pronunciation model were indicated. Three types of phonetic changes were taken into account: 1. the disappearance /elision/ of allophones or allophonic clusters, which was denoted with a slant following the letter, 2. the addition of allophones which was denoted by figure "6", 3. qualitative changes which were denoted by means of angle brackets. The "<...>" brackets embrace the pronunciation model and the ">...>" brackets embrace the phonetic realization. This can be illustrated as follows:

a/ Phonetic transcription: [t͡ʃɛã̯ ɕɛ nawut͡ʃɨt͡ɕ]
b/ Pronunciation model: [t͡ʃɛba ɕɛ- naut͡ʃɨt͡ɕ]
c/ Sequential transcription of phonetic variancy /denoted by braces/:
{<T.S.>C.>EB/A S"E- NAW6UC.YC"}

Depending on the aim of work the manner of sequential transcription proposed here can be easily modified, e.g. by changing the combination of symbols or by assigning them a different meaning. With a few modifications introduced into the program it is possible to analyse dialectal pronunciation, progress in the mastery of foreign language phonetics, development of child s speech and numerous other phonetic problems.

The symbols used to denote disappearance /elision/, addition and substitution of allophones as well as other symbols used in the transcription of phonetic variancy helped to compare phonetic transcriptions of the realizations with the adopted pronunciation model. Those other symbols for comparing the pronunciation model and the phonetic realization will not be discussed.

THE PHONOLOGICAL SYSTEM

The pronunciation model was uniform for a speech sequence and not for words pronounced in isolation. Therefore, particular word entries could have several optional forms compatible with the model. In order to arrange the material for the sake of the computer program we introduced a phonological system developed for the purposes of this work. Phoneme is treated as an element representing a class of allophones, and it is capable of discriminating meanings - hence we adopted a semantic definition of the phoneme. Thus, both in the "Dictionary of phonetic realizations" and "Tables of phonetic realizations of allophones and allophonic clusters" we got an adequate classification of the texts under analysis /adequate = meeting the expectations/. It is worth adding here that over 100 allophones occurred in the transcriptions of specific realisations, 70 allophones were present in the adopted pronunciation model, and 38 phonemes constituted the phonological system. Thus, the adopted phonological system proved very economical at different stages of data processing.

DICTIONARY OF PHONETIC REALIZATIONS

On the basis of the texts we have developed a "Dictionary of phonetic realizations". In the transcriptions cited in the previous part of this paper, words are separated by a space. We have adopted the graphic definition of the word, and segmented the text in accordance with the principles of Polish orthography. Hence, a word as it appears in the text and not merely its base form is the basic unit. Texts 66 398 words long produced 8 720 different entries. Entries written phonologically are arranged alphabetically /cf. Table 1/. /Unfortunately, the orthographic forms of the words had to be handwritten/. The structure of entries is illustrated below:

Phonological transcription	Orthographic transcription		
Allophonic transcription of specific realization in the rank order within the entry		b	c
/VJEZ1"AWAM/	"wiedziałam"	26	100.0
1. V,JEZ1"AAM		21	80.8
2* V,JEZ1"AWAM		2	7.7
3. V,IZ1"AAM		1	3.8
4. V,JEZ1"AAW=		1	3.8
5. V,JEZ1"AAM,		1	3.8

Remarks on the structure of entries.

a/ Asterisk denotes optional forms, compatible with the pronunciation model and different with respect to the phonetic context.

b/ The number of occurrences of a given realization in the texts analyzed.

c/ The percentage of occurrences of a given realization within a given entry.

The "Dictionary ..." is supplemented with the "Rank index". Entries transcribed phonologically have been arranged in their rank order; their rank number and the number of occurrences in the material tested have been given.

TABLES OF PHONETIC REALIZATIONS OF ALLOPHONES AND ALLOPHONIC CLUSTERS

In linguistic investigations, depending on the aim of the work or the researchers interests, morpheme or syllable are regarded as units whose order is higher than that of the phoneme. In some phonetic works a cluster of phonemes is a unit whose order is higher than that of the phoneme. For example, in B. Dunaj s monograph[2] the phonetic variancy in the pronunciation of consonant clusters of modern spoken Polish is studied. Text segmentation into consonants and consonant clusters and, by analogy, into vowels and vowel clusters /diphthongs and triphthongs/ proved very useful for our purposes. Below we illustrate the segmentation of a text prepared for the sake of the "Tables" transcribed phonologically by means of the sequential transcription /the phonetic variancy was not taken into account to avoid complicating the notation/. The unit-boundaries are marked by vertical lines:

{|J|E|ZD V; V,J|EJ=|Z"|E|N"|U|}

The "Tables of phonetic realizations" are arranged similarly to the "Dictionary of phonetic realizations". The entry consists of a speech sound or a cluster written phonologically, the rank order and the number of occurrences in the material analysed. Next, in the rank order, the following information is given: 1. phonetic transcription of specific realizations, 2. the number of their occurrences in the text, 3. percentage of occurrences of a given realization in relation to the entire group.

CONCLUDING REMARKS

An analysis of the "Tables" permits formulating conclusions on the phonetic variancy occurring in the analysed texts. Comparing data in the "Tables ..." and those in the "Dictionary of phonetic realizations" we can evaluate which of the phonetic changes are phonetically conditioned and which are lexically motivated. For example we have observed that the elision of [w] in the intervocalic position occurs most frequently in the third person of verb forms in the past tense, e.g. "chciała"-[xtɕaa̯], "myślała"- [mɨɕlaa]. In other word forms, in a similar phonetic surrounding, e.g. [mama], [kɔwo] elision is also present but much less frequent.

On the basis of the sequential transcription of phonetic variancy developed with the help of a computer program we can examine numerous phenomena which have not been mentioned here. Apart from typical phonostylistic examinations we can also obtain data on the distinctness of pronunciation of various speakers, paralinguistic phenomena etc.

It is purposeful to use the proposed method for analyses of a large set of texts and when we want to analyse various phonetic phenomena simultaneously. If short texts are available or if only certain phonetic problems are examined, there is no point in using the computer.

REFERENCES

/1/. Abercrombie, D.: "The recording of Dialect Material". in: "Studies in phonetics and linguistics". Oxford University Press: London 1965.

/2/. Dunaj, B.: "Grupy spółgłoskowe współczesnej polszczyzny mówionej /w języku mieszkańców Krakowa/", Warszawa -Kraków: PWN, 1985.

/3/. Madejowa, M.: "Współczesna polszczyzna mówiona. Zagadnienie normy językowej na przykładzie wybranych zjawisk fonetycznych". Kraków 1980. /Ph.D. disertation: UJ/.

/4/. Steffen-Batogowa, M.: "Analiza struktury przebiegu melodii polskiego języka ogólnego". Poznań 1963. /Ph. D. dissertation: UAM/.

ACKNOWLEDGMENT

The computer program was written by L. Mierzyński, of the Computer Centre, A. Mickiewicz University, Poznań.

PHONOLOGICAL PLANNING FOR SPEECH PRODUCTION
SPEECH ERROR EVIDENCE FOR WORD-BASED VS. SYLLABLE-BASED STRUCTURE

Stefanie Shattuck-Hufnagel

Speech Communications Laboratory, Research Laboratory of Electronics
Massachusetts Institute of Technology, Cambridge, Mass. USA 02139

ABSTRACT

Single-segment speech errors are more likely to occur between two word-onset consonants than between a word-onset consonant and one in word-medial position. Errors elicited from speakers of American English using tongue twister stimuli show that this onset-position similarity predominates even when the two consonants are followed by vowel nuclei with different degrees of lexical prominence or stress. This onset similarity constraint holds for tongue twisters in both word-list and phrasal form, suggesting that, for this language, word and morpheme structure forms a part of the representation that is in force at the point in production planning when segmental interaction errors occur.

INTRODUCTION

Phonological planning for speech production presumably involves the representation of organizational units larger than the individual segment. Candidates include elements which have proven useful in the statement of phonological, phonetic and metrical regularities, such as the morpheme, the word, the syntactic phrase, the syllable, the foot and the prosodic phrase, etc. Phonological errors in speech production can shed some light on the question of which of these larger units play a part in the planning process, because of the following characteristic: when two segments of an utterance interact in an error like an exchange, they tend to come from corresponding structural positions. For example, an initial consonant tends to interact with another initial consonant, rather than with a final consonant, giving e.g. "bate of dirth" for "date of birth", but not "thate of bird" [1]. Similarly, final consonants tend to interact with each other and not with initial consonants, giving "noth wort knowing" for "not worth knowing", but not "thot wern knowing."

By examining the nature of this position-similarity constraint on segmental interaction errors, we can begin to determine the identity of the larger elements that play a role in segmental processing. In this way, we can distinguish between proposed models of the cognitive process of production planning that rely on different representational elements. For example, one type of model for segmental processing might postulate that words and morphemes are part of the phonological planning frame, while a second type of model might hypothesize that by the time segmental errors occur, the word boundaries have been erased, leaving an organizational framework that consists only of prosodic structures like the syllable, foot, etc. A model of Type 1 provides for the possibility that two target segments that interact in an error tend to be from similar positions in their respective words, while a model of Type 2 cannot easily account for such a word-based position-similarity constraint. Instead, Type 2 models might predict a constraint based on position in the syllable or in the foot, or perhaps on syllable prominence.

What does the pattern of segmental speech errors suggest with respect to this issue? Segmental interaction errors collected from spontaneous American English speech [2] do not resolve the question clearly , because so many of the errors occur in monosyllabic words, or in words with lexical stress on the first syllable, so that both word-based and prosody-based models predict the same result [3]. We need to know the error distribution for consonants in pairs ofwords that do not both have lexical stress on the first syllable. By comparing the number of /p/-/f/ interaction errors in sequences like (a) "parade fad" and (b) "repeat fad", we can determine whether more errors occur (a) when the two target segments share position in their

word onsets (but not before the stressed vowel), or (b) when the two target sements share position before the stressed vowel (but not in their word onsets.)

Shattuck-Hufnagel [4] has described an elicitation experiment that used tongue twisters of these two types. For both a reading and a recall task, results show substantially more interaction errors between two word-onset consonants than between one word-onset and one word-medial consonant, i.e. more /p/-/f/ errors for stimuli like "parade fad foot parole" than for stimuli like "repeat fad foot repair." This pattern supports Type 1 models, in which words and morphemes are still part of the planning representation at the point where segmental errors occur.

THE EXPERIMENT

One question that might be asked about the generalizability of the tongue-twister results to models of spontaneous speech planning concerns the list-like nature of the stimuli. Since each stimulus consisted of four lexical items presented visually, with spaces between them, it might be argued that the task somehow emphasized the first letter of each item and that this special emphasis turned the corresponding four word-onset segments of each twister list into a mutually-confusable set. The solution to this problem is to embed the four words of each twister in a larger string of words, so that the set of post-space segments does not isolate the four targets as a particularly-confusable set. In the experiment to be reported here, this embedding was accomplished by positioning each of the original twister words in a short phrase, so that "parade fad foot parole" became "The parade was a fad and the foot got parole," etc.

The decision to use four short locally well-formed phrases as the embedding material was motivated by the results of a different experiment, in which the effect of list-like vs. phrase-like stimulus structure was investigated for the initial and final consonants of CVC words [5]. In that elicitation experiment, changing from lists to phrases had a substantially larger effect on final consonants than on initial consonants. When word-list twisters like "peal tone pan tool" and "leap note nap lute" were turned into phrasal twisters like "From the leap of the note to the nap of the lute", initial consonant interaction errors declined only 16% but final consonant errors declined 78%. The resulting low rare of final-consonant errors elicited by phrasal twisters more closely resembles the pattern found in collections of spontaneous speech errors, suggesting that phrasal twisters may be more appropriate than word-list twisters for invoking the representations used in spontaneous speech planning. This provides further motivation for re-testing the word-onset similarity constraint in a phrasal context.

Stimuli

The original word-list experiment employed 24 sets of tongue twisters, each set constructed around a pair of consonantal segments which had a high error rate in a corpus collected from spontaneous speech [6]. For the present experiment, the words of these twisters were embedded in locally-grammatical phrases. Each set contained four twisters, developed according to the principles illustrated here for the target pair /p/-/f/.

Condition 1: Share Word Onset Position
Both members of target pair occur in word onset; only one occurs in position before stressed vowel
e.g. The parade was a fad
and the foot got parole.

Condition 2: Share Stress Position
Both members of target pair occur before stressed vowel; only one occurs in word onset
e.g. To repeat was a fad
and the foot could repair.

Condition 3: Share Both Positions
Both members of target pair occur in word onset and both occur before stressed vowel
e.g. To his peril the fad
was the foot of a parrot.

Condition 4: Share Neither Position
Only one member of target pair occurs in word onset and before stressed vowel
e.g. For a ripple a fad
has a foot and is rapid.

The twenty-four sets of twisters were built around 12 different pairs of target consonants: /f-p, r-l, b-g, l-y, b-p, m-n, r-w, d-g, p-k, l-n, d-t, j-d/. In addition, each twister contained a third "filler" consonant, like /r/ in the set above, which was phonologically dissimilar.

Speakers

Twenty MIT undergraduates participated as speakers; thirteen were male and seven were female. All were native speakers of American English with no known speech or hearing deficits; all were right-handed. Speakers were paid a nominal amount for their participation.

Presentation

The 24 sets of stimuli were divided into two groups of twelve sets each. Speakers were divided into two groups of 10 each. One group of speakers produced half the stimuli: Conditions 1 and 3 from the first 12 sets, and Conditions 2 and 4 from the remaining 12 sets. The second group of speakers produced the other half of the stimuli: Conditions 2 and 4 from the first 12 sets, and Conditions 1 and 3 from the remainder. Thus every speaker produced two of the stimuli from each of the 24 sets, or 48 stimuli in all. The 48 stimuli were presented in two sessions of 24 stimuli each; the two half-hour sessions were counterbalanced for order of presentation.

The tongue twisters were typed on 3 x 5 cards and presented visually to the speaker, who read each one aloud three times and then turned the card over and recited it three times from memory. After each card the speaker generated a sentence as part of another experiment, so that the twisters were separated from each other by a 15-60 second period and by a different kind of activity. Speech was tape-recorded for later transcription.

Scoring

Utterances were transcribed by ear with repeated listening, and scored for errors. Only segmental interaction errors between the two members of the target segment pair were scored, which means that the following error types were not included: substitution errors with no apparent source in the utterance, omission and addition errors, and errors that involved units larger than a single segment, such as a CV or VC. In addition, interactions between one of the two target segments and the third "filler" segment were scored separately. Results for the reading and recitation tasks were also scored separately.

Results

Like the findings for word-list twisters described earlier, the results for phrasal twisters show a stronger tendency for word-onset consonants to interact with each other than for word-onset consonants to interact with medial consonants, even when the onset-medial pairs shared pre-stress position. These results can be summarized as follows.

Condition	Number of Errors		
	Reading	Recall	Total
1) Share Word-Onset	49	85	134
2) Share Prestress	17	42	59
3) Share Both	58	144	202
4) Share Neither	5	10	15

Three aspects of the data in particular illustrate the predominance of onset-onset errors:

1. **Conditions in which the target consonants both occurred in the word onset provoked many more errors than conditions where one of the two target consonants appeared elsewhere in the word.** The two onset-onset conditions (1 and 3) elicited a total of 336 interactions, and the two onset-nononset conditions (2 and 4) only 74.

2. **The two conditions which directly pit word-onset similarity against pre-stress similarity show that shared word onset induces twice as many errors as shared prestress position.** That is, Condition 1 stimuli elicited 134 errors while Condition 2 stimuli elicited only 59. In Condition 1, the two target segments share word onset position but differ in the stress of the following vowel, while in Condition 2, the two targets share prestress position but differ in word position, providing a direct test of the strength of these two position similarity constraints. Under these circumstances, word onset similarity clearly can overpower dissimilarity in stress more often than stress similarity can overpower dissimilarity in onset position.

3. **Unexpected interactions between a member of the target consonant pair and the phonologically less-similar third "filler" consonant also favor interactions between word onsets.** The third "filler" consonant in the bisyllabic words of each twister was less similar to the target consonant in the monosyllabic word than the two targets were to each other, as measured by the number of shared features. On a simple categorization as same or different in Voicing, Manner or Place, the mean difference between the two members of the target pairs was 1.2 features out of 3, while the mean for the target-filler pairs was

2.3 features. Despite this feature dissimilarity, speakers produced 32 unexpected interactions between onset-onset pairs (like /r/-/f/ in "To repeat was a fad that the foot could repair"), and only 3 for onset-nononset pairs (like /r/-/f/ in "The parade was a fad and the foot got parole".) Although the numbers are much smaller, the unexpected errors clearly support the claim that shared word-onset status is more conducive to interactions than is shared prestress position.

DISCUSSION

The distribution of errors in the elicitation experiments described above suggests that the predominance of word-onset errors in corpora gathered from spontaneous speech is not the result of some unnoticed factor that is coincident with word-onset position, but rather can be interpreted as evidence that lexicomorphemic structure is part of the processing representation that speakers make use of as they plan utterances for production. This claim alone would certainly not be surprising; the new and further information here is concentrated in two points:

1) Word structure is part of the processing representation at the point when segmental errors occur. This observation rules out a class of models in which the segmental processing mechanism that is susceptible to serial ordering errors operates on a representation in which morphemic structure has been erased, and supports the class of models in which that structure is preserved. There are several ways in which this claim for the preservation of morphemic structure could be realized in a particular model, including (a) as part of the organizational framework of serially-ordered slots to be associated with target segments, or (b) as part of the lookup mechanism for retrieving the phonological forms of morphemes from long-term storage. For futher discussion of this issue, see Shattuck-Hufnagel [7].

2) The second summary point is that prosodic structure also plays a role in this processing representation. Despite the fact that shared word-onset position can overwhelm both stress dissimilarity and feature dissimilarity to provoke an interaction error, the experimental results provide evidence that prosodic factors are also at work. To see this, compare the results for Condition 2 with Condition 4. Suppose that lexical stress similarity and syllable position affiliation similarity played no role in conditioning errors. Then we might expect a similar rate of /p/-/f/ errors for Condition 2 and Condition 4. That is, Condition 2 stimuli like "To repeat was a fad that the foot could repair" (where /p/ and /f/ are both pre-stressed and syllable-initial), and Condition 4 stimuli like "For a ripple a fad has a foot and is rapid" (where only /f/ is prestressed and the syllable position affiliation of /p/ is unclear), could be expected to provoke about the same number of errors if syllable affiliation and lexical prominence play no role in constraining the error process. Instead, Condition 2 elicits substantially more errors than Condition 4. This suggests that both lexicomorphemic structure and prosodic structure play a role in the phonological planning representation at the point in that process where segmental errors occur. Further elicitation experiments are in progress to determine in more detail the nature of the prosodic component of this representation.

REFERENCES

[1] V.A. Fromkin, Speech Errors as Linguistic Evidence, The Hague: Mouton, 1973.

[2] M.F. Garrett, The analysis of sentence production, in G. Bower (Ed.), The Psychology of Learning and Motivation, Vol. 9, New York: Academic Press, 1975.

[3] S. R. Shattuck-Hufnagel, Sublexical units and suprasegmental structure in speech production planning, in P. MacNeilage (Ed.),The Production of Speech, New York: Springer-Verlag, 1983.

[4] S. R. Shattuck-Hufnagel, Context similarity constraints on segmental speech errors, in J.L. Lauter (Ed.), Proceedings of the Conference on the Planning and Production of Speech in Normal and Hearing-Impaired Individuals, American Speech-Language-Hearing Association ASHA Report No. 15, 1985.

[5] S. R. Shattuck-Hufnagel, Position of errors in tongue twisters and spontaneous speech, in Speech Group Working Papers, Research Laboratory of Electronics, MIT, Vol. 1, 1982.

[6] S. Shattuck-Hufnagel and D.H. Klatt, The limited use of distinctive features and markedness in speech production. Journal of Verbal Learning and Verbal Behavior 18, 41-55, 1979.

[7] S. Shattuck-Hufnagel, The role of word-onset consonants in speech production planning, in E. Keller and M. Gopnik (Eds.), Motor and Sensory Processes of Language, Hillsdale, N.J.: Lawrence Erlbaum Associates, in press.

THE ORGANIZATION OF CONSTRAINTS ON PHONOLOGICAL SPEECH ERRORS

EMANUELA MAGNO CALDOGNETTO, LIVIA TONELLI, KYRIAKI VAGGES, PIERO COSI

Centro di Studio per le Ricerche di Fonetica del C.N.R.
Via G. Oberdan, 10
35122 Padova, Italia

ABSTRACT

The intrinsic and extrinsic constraints are taken into account in the analysis of 455 consonantal errors - part of a corpus containing about 1500 Italian lapses. Parameters governing both types of constraints, and the hierarchical organization of the extrinsic ones are discussed. Our results provide evidence for some phonological properties of Italian.

1. Since the pioneering paper by Victoria Fromkin. [1] the list of those working on speech errors has come to include - as the rich literature on the topic shows [2,3,4,5,6,7,8,9,10] - linguists, psycolinguists, cognitive psycologists, and neurologists, as well as phoneticians, in the common effort to shed light on the organization of language performance. Research has centered mainly on two related questions : a) to what extent the grammatical units and structures represented in the human mind match the processing representations b) to what extent principles and rules which govern grammatical knowledge serve the process of speech production planning.
This paper is concerned with the second topic, and, in particular, with the parameters governing the phonological intrinsic and extrinsic constraints on the occurrence of speech errors.
An analysis of the Italian data was carried out along the lines of van den Broecke & Goldstein's [11] and Shattuck-Hufnagel's work [12,13]. The difference between our results and those obtained for the English data seem to reflect specific properties of the phonological system of Italian.
2. Our analysis is based on a corpus of 455 spontaneous speech errors that involve consonantal phonemes collected at the Centro di Studio per le Ricerche di Fonetica [14] as part of a larger project on speech production. The errors were classified by means of the now-classic superficial typology which includes exchanges : ma è senza senso --> ma è sensa senzo (but it is without any sense), contextual substitutions : insufficienza mentale --> insufficienta mentale (mental insufficiency), i dati bibliografici --> i dati bibliogratici (the bibliographic data) and non-contextual substitutions : questi succhi di frutta finiscono subito --> questi ciucchi di frutta finiscono subito (these fruit juices sell out immediately).

Table 1. Confusion matrix of speech errors.

		intrusion																				
		p	b	t	d	k	g	f	v	s	z	ʃ	ts	dz	tʃ	dʒ	m	n	ɲ	l	ʎ	r	j
	p		6	7	1	18		4	1	6					2		3					1	
	b	3				1	1	1	2			1			1	1	2			1		1	
	t	8	1		4	15	1	4	1	6	2		3	1	7	1	2	3	1	2			
	d	1		5			2	1	1							6	1	1		2			
	k	17	2	15			1	1	1	5		1			1	1	2	1		1		1	
	g	1	1	2	1	1											1			1			
	f	7	1	2	1	2			2	3													
	v	1	1		2			2		7	1				1		2	1		4		1	
	s	7		3	2	6		2	8		1	1					1	2				2	
	z								3									5		2		1	
	ʃ		1				1	1	1	4													
	ts			1						3					3								
	dz															3		2					
	tʃ	2	1	6		3			1	4		5					1	2		1			
	dʒ			1	6										1			2		1			
	m	2	1	1	1	1	1	1	4	1							6			1		3	
	n			3	1				1	1	4	1	2	2		3		2		2	2	2	
	ɲ																						
	l		1	4		1	1		3	1	2				1		1	1				12	
	ʎ																		1				
	r	1	1	1		2					1				1		3	2		9			3
	j																			1		3	

For the purpose of this analysis, insertions, deletions and shifts were not considered.
3. The evaluation of the intrinsic constraints, i.e. the restrictions on the occurrence of speech errors on the paradigmatic axis, is based on the inspection of the confusion matrix given in Tab. 1. The symmetry ($X^2=12.40$ $p<.5$) of the matrix leads to the same results obtained for English by Shattuck-Hufnagel and Klatt [15]: for any given

pair of phonemes there is no preference for one of them to act as intrusion or as target, i.e. the behaviour of speech errors is not governed by the parameter "dominance".
The next analysis was carried out in order to evaluate the weight of a second parameter ; the influence of "similarity" as a function of the intrinsic phonetic/ phonological characteristics of phonemes on speech errors. For this purpose we basically followed the methodological approach proposed by van den Broecke and Goldstein [11]. We employed a "behavioral" feature system obtained a posteriori by means of hierarchical clustering and multidimensional scaling analyses of the substitution patterns.

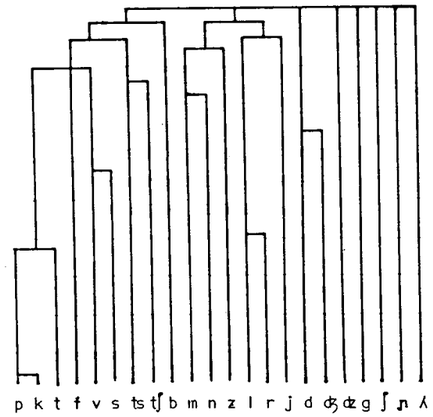


Figure 1. Hierarchical clustering representation.

The dendrogram in Fig. 1. shows a clustering involving the voiceless stops /p,t,k/. At a lower similarity level, a cluster is formed by the voiced phonemes separated from the voiceless phonemes with the exceptions of /v,b,ʃ/. The configurations obtained by the multidimensional scaling technique yield more information, since the groupings derived from the clustering analysis are not sufficient to represent the structure underlying all the phonemes examined. In fact, in Fig.2a voiced and voiceless consonants form two well distinguished groups; resonants constitute another separate group, with the exception of /v/; a fourth group is made up by the stop consonants. Finally, Fig. 2b shows four separate groups which correspond to the places of articulation: labial, dental-alveolar, palatal, velar, with the possible exception of /l/.
The groupings observed above were incorporated in the following a posteriori matrix, in which the features <fricative> and <lateral> were added in order to distinguish the consonants considered in an unambiguous way.
The matrix in Tab. 3 , compared with the matrix elaborated by van den Broecke and Goldstein [11] on

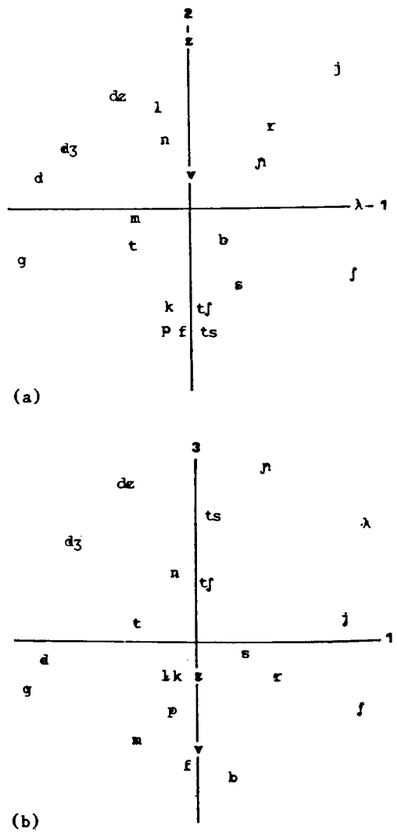


Figure 2. a) dimensions 1 and 2, b) dimensions 1 and 3 of the three-dimensional MDS configuration.

the basis of two different English corpora and a German corpus, shows a first difference concerning the specification of the feature <stop> and a second difference regarding the inclusion of the feature <resonant>. Both properties are evident in the configuration of our data and both of them are part of the phonological representation : in Italian a segment-structure rule operates in order to eliminate palatal articulation within stop consonants, and a syllable structure rule eliminates non-resonant consonants from syllable coda position.
As the matrix in Tab. 3 illustrates, the phonetic/ phonological similarity of phonemes constrains

their interaction in speech errors [15,16,11]; in fact, the degree of involvememt of phonemes in lapses is inversely proportional to their differences expressed in number of features :

Table 3. A-posteriori feature matrix for consonantal speech errors.

	p	b	t	d	k	g	f	v	s	z	ʃ	ts	dz	tʃ	dʒ	m	n	ɲ	l	ʎ	r	j
voice	-	+	-	+	-	+	-	+	-	+	-	-	+	-	+	+	+	+	+	+	+	+
reson	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+
nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-
stop	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
place	l	l	a	a	v	v	l	l	a	a	p	a	a	p	p	l	a	p	a	p	a	p
cont.	-	-	-	-	-	-	+	+	+	+	-	-	-	+	+	+	+	+	-	+		
later	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-

34.5%, 32.1%, 18.7%, 5.5%, 7.0% and 2.2% of the errors occur respectively betweem phonemes differing in 1, 2, 3, 4, 5, and 6 features.
4. The second set of constraints is constituted by the extrinsic constraints, i.e. the structural and contextual factors which influence the occurrence of speech errors on the syntagmatic axis; for this study we analyzed the following factors:
- the role of syllabic structure;
- the positional role of segments within words;
- the influence of lexical stress;
- the influence of the phonological segmental context;
In order to evaluate the role of syllable structure, we excluded from our computation all non-contextual substitutions and ambiguous errors, as well as all errors involving geminates. As far as the geminates are concerned, there is still some theoretical disagreement about their mono- versus biphonematic status, and, hence, about their syllabification [14]. Consequently, the analysis was carried out on a corpus of 240 errors.
A high rate (93%) of the errors concern the interactions between phonemes in syllable onset position. This finding has been used in the literature on speech errors as an argument in favour of the break down of the syllables in two groups: the onset and the rhyme [17,18,7]. This argument has already been challenged by Vennemann [19] and Davis [20]; our data furnish some more evidence for a different structural configuration of the syllabic unit. In fact, 1% of the errors involve interaction between nuclei and codas: il pulman --> il pluman (the bus); 3% involve interaction between onsets and codas: bisogna stimolarlo --> bisogna stimorarlo (it must be stimulated); and 3% of errors involve interactions between onsets and nuclei: lavati --> alvati (wash yourself). The parameter which constrains errors to occur in syllable initial position with such high rate is constituted, in our opinion, by the phonological strength: segments in syllable onsets, i.e. in strong positions [21,22,19] are more available and hence more prone to interact.
The parameter "similarity of strength" seems to govern, also, the next two constraints, i.e. lexical stress and word position, both operating within the word domain.
For their evaluation, only between-word errors were considered (N=123); in fact, within-word errors would, by their nature, limit the array of possible interactions [12,13]. Futhermore, we removed from the computation the compounds and the monosyllabic words, which, however, constituted only the 3% of the whole corpus.
The fact that 48% of the errors concern phoneme pairs in word initial position and the finding that 42% of the errors involve interactions between phonemic segments in stressed syllables, strongly supports the influence of both "word position constraint" and "lexical stress constraint" [23,24].
In order to evaluate which constraint has greater strength, a separate analysis was carried out. We considered only the errors (N=52) in which all the interacting phonemes occurred in stressed syllables but were located in different word positions.

Table 4. Distribution of errors in stressed position in different word position.

Initial initial	Medial Medial	Final Final	Different Position
67.30%	19.83%	1.92%	11.53%

As Tab. 4 illustrates, the results furnish some evidence for a hierarchical organization of both constraints: word onsets influence the occurrence of speech errors more strongly than lexical stress. As far as the last constraint is concerned, i.e. the phonological context, there is no way to establish, in a straightforward manner, its influence on the occurrence of speech errors. The interactions between phonemes followed by different syllable nuclei make up 44% of the errors, 38% concern identical syllable nuclei and 18% is constituted by CV sequences for which there is no way to establish whether whole syllables are involved in errors or simply consonantal phonemes followed by identical vowels.
5. Within a model of language production, our analysis starts at Garrett's [25,26] positional level, on which "superficial phrasal geometry" has been established, and it aims "to delineate underlying representations from which superficial phonetic phenomena can be derived" [27]. The

results fit with the slot-and-filler framework proposed by Shattuck-Hufnagel [28,12,13]: the likelihood for two phonemes to interact in speech errors depends on their degree of similarity. Our data show that the factors by which intrinsic similarity is established reflect the phonological organization of Italian. On the paradigmatic axis the dimensions <manner> <place> <voice> <sonority> correspond to the categories on which phonological processes depend; on the syntagmatic axis the hierarchical organisation of the different domains on which "similarity of strength" works, reflect the fact that, in Italian, underlying syllable structure is substantially neither altered within the word domain nor it is influenced by stress.

REFERENCES

[1] V.A. Fromkin, "The non-Anomalous Nature of Anomalous Utterances", Language 47, 1971, 27-52.

[2] V.A. Fromkin (ed.), Speech Errors as Linguistic Evidence, Mouton, The Hague, 1973.

[3] V.A. Fromkin (ed.), Errors in Linguistic Performance. Slips of the Tongue, Ear, Pen and Hand, Academic Press, London, 1980.

[4] A. Cutler (ed.), Slips of the Tongue and Language Production, Mouton, The Hague, 1982.

[5] H.W. Buckingham, "On Correlating Aphasic Errors with Slips-of-the-Tongue", in Applied Psycholinguistics 1, 1980, 199-220.

[6] E. Söderpalm Talo, Speech Errors in Normal and Pathological Speech, Gleerup, Malmö, 1979.

[7] J.P. Stemberger, "Speech Errors and Theoretical Phonology: A Review", Indiana University Linguistic Club, Bloomington, 1983.

[8] M. Kilani-Schoch, Processus Phonologiques, Processus Morphologiques et Lapsus dans un Corpus Aphasique, Lang, Bern, 1982.

[9] B. Butterworth, "Some Constraints on Models of Language Production", in B. Butterworth (ed.), Language Production, Vol. 1, Speech and Talk, Academic Press, London, 1980, 423-459.

[10] W.U. Dressler, E. Magno Caldognetto, L. Tonelli, "Phonologische Fehlleistungen und Paraphasien im Deutschen und Italienischen", Grazer Linguistische Studien, 26, 1985, 46-56.

[11] M.P.R. van den Broecke, L. Goldstein, "Consonant Features in Speech Errors", in V. A. Fromkin (ed.), Errors in Linguistic Performance. Slips of the Tongue, Ear, Pen and Hand, Academic Press, London, 1980, 213-230.

[12] S. Shattuck-Hufnagel, "Sublexical Units and Suprasegmental Structure in Speech Production Planning", in P.F. Mac Neilage (ed.), The Production of Speech, Springer-Verlag, Berlin, 1983, 109-136.

[13] S. Shattuck-Hufnagel, "The Representation of Phonological Information", in Phonology Yearbook 3, 1986, 117-149.

[14] E. Magno Caldognetto, L. Tonelli, "Syllabic Constraints on Phonological Speech Errors in Italian", in W. Dressler, L. Tonelli (eds.), Natural Phonology from Eistenstadt, CLESP, Padova, 1985, 73-88.

[15] S. Shattuck-Hufnagel, D.H. Klatt, "How Single Phoneme Error Data Rule out two Models of Error Generation", in V.A. Fromkin (ed.), Errors in Linguistic Performance, Academic Press, New York, 1980, 35-46.

[16] S.G. Nooteboom, "Some Regularities in Phonemic Speech Errors", Institute for Perception Research, Eindhoven, Annual Progress Report, 2, 1967, 65-70.

[17] D.G. Mackay "The Structure of Words and Syllables: Evidence from Errors in Speech", Cognitive Psycology 3, 1972, 210-227.

[18] A. Crompton, "Syllables and Segments in Speech Production", in A. Cutler (ed.), Slips of the Tongue and Language Production, Mouton, New York, 1982, 109-162.

[19] Th. Vennemann, Neuere Entwicklungen in der Phonologie. Mouton de Gruyter, Berlin, 1986.

[20] S. Davis, "Topics in Syllabic Geometry", Ph.Diss., The University of Arizona, 1985.

[21] P. Donegan, D. Stampe, "The Syllable in Phonological and Prosodic Structure", in A. Bell and J.B. Hooper (eds.), Syllables and Segments, North-Holland, Amsterdam, 1978.

[22] W.U. Dressler, Morphonology. The Dynamics of Derivation, Karoma, Ann Arbor, 1985.

[23] D.A Fay and A. Cutler, "Malapropisms and the Structure of the Mental Lexicon", Linguistic Inquiry, 8, 1977, 505-520.

[24] A. Cutler, "The Reliability of Speech Error Data", in A. Cutler (ed.), Slips of the Tongue and Language Production, Mouton, New York, 1982, 7-28.

[25] M.F. Garrett, "Syntactic Processes in Sentence Production", in R. Wales, E.C.T. Walker (eds.), New Approaches to Language Mechanisms, North-Holland Press, Amsterdam, 1976, 231-255.

[26] M.F. Garrett, "Levels of Processes in Sentence Production", in B. Butterworth (ed.), Language Production, Vol. 1, Speech and Talk, Academic Press, New York, 1980, 177-220.

[27] W.J.M. Levelt, "Spontaneous Self-Repair in Speech: Processes and Representations", in M.P.R. van den Broecke and A. Cohen (eds.), Proceedings of the Tenth International Congress of Phonetic Sciences, Foris Publications, Holland, 1984, 105-117.

[28] S. Shattuck-Hufnagel, "Speech Errors as Evidence for a Serial Ordering Mechanism in Sentence Production", in W.E. Cooper, E. Walker (eds.), Sentence Processing: Psycholinguistic Studies Presented to M. Garrett, Erlbaum, Hillsdale, 1980, 295-342.

STRATEGIE CONVERSATIONNELLE
DONNER ET PRENDRE LA PAROLE

MARIA RAQUEL DELGADO MARTINS

FACULDADE DE LETRAS
LABORATORIO DE FONETICA
UNIVERSIDADE DE LISBOA:PORTUGAL

RÉSUMÉ

Le présent travail expose les résultats d'une analyse réalisée sur un discours conversationnel entre deux interlocuteurs,constitué par 105 tours de parole,de façon spontanée et authentique.L'analyse expérimentale porte sur certains marqueurs qui permettent l'alternance des tours de parole:le type d'éxécution de chaque fin de tour et la durée de la pause qui le sépare de la prise de parole de l'interlocuteur.Les résultats permettent de dégager certains indices de régulation,d'établir des stratégies pour donner et prendre la parole.Les résultats viennent également confirmer certains principes déjà établis comme celui de la relevance conditionnelle,déterminée dans ce cas par la spécificité du rôle social des interlocuteurs étudiés.

INTRODUCTION

Nos travaux sur l'intonation (1)montrant la non-systématicité de l'organisation des indices au niveau de la phrase lue, nous ont conduits à étudier la parole 'spontanée' en vue de déterminer des stratégies d'utilisation de ces indices comme marqueurs interactionnels, dans le sens où l'éxécution de chaque tour de parole impose un retour de l'interlocuteur, structurant ainsi le discours.
Vouloir trouver une régularité des indices peut paraître une 'illusion'comme pour certains auteurs comme Goldman-Eisler (2) pour qui la parole spontanée est 'une activité fortement fragmentée et discontinue'.Cependant, les travaux portant sur les erreurs de production de la langue maternelle(3) ou en langue seconde (4)montrent que certaines hésitations ou certaines pauses peuvent être des preuves d'une programmation d'éxécution à différents niveaux du discours du locuteur et de l'interlocuteur.
Nous n'utiliserons pas ici les notions dacte de langage comme pour Searle (5), mais nous considérerons plutôt les frontières entre les'actes' et leurs rapports réciproques qui constituent, selon la définition de Labov et Franshel (6) 'une matrice de réalisations et d'actions liées entre elles par un réseau..'

Notre étude portera donc, sur une suite de 'tours de parole' considérés comme unités indépendemment de leur structure syntaxique ou grammaticale et de leur durée, chaque unité élocutionnelle pouvant être constituée par une syllabe , une phrase ou une suite de phrases produites par un locuteur. Les marqueurs d'interaction seront donc considérés aux frontières d'un tour de parole à l'autre.

ANALYSE DE LA CONVERSATION

Pour atteindre les objectifs énoncés , nous avons travaillé sur une conversation 'spontanée' entre deux locuteurs (R) et (C)a qui il a été demandé de parler librement pendant quelque temps sans détermination d'un thème. La conversation a duré'naturellement' aproximativement 15 minutes et s'est constituée en 105 tours de parole, portant sur un sujet scientifique du domaine commun des sujets.Bien qu'étant enregistrée en Chambre isolée acoustiquement, les conditions permettent de considérer ce discours comme totalement oral,spontané et authentique.
Il est important de préciser qu'il existe une relation d'hierarchie institutionnelle entre les sujets (R) étant supérieur à (C) et que la conversation a anticipé une discussion publique sur le travail produit par le sujet (C).Malgré cette relation le lien existant entre les deux sujets a permis un registre très familer au long de l'enregistrement.
Le discours s'est organisé en cinq grandes parties selon les thèmes et les tours de parole suivants:
1ère. partie: du début du tour 1 à la fin du tour 4.Introduction sur les conditions d'enregistrement.
2ème. partie: de la fin du tour 4 à la fin tour 31.Le sujet (R) questionne le sujet(C) sur les conditions générales de l'élaboration du travail de (C).
3ème. partie:de la fin du tour 31 de (C) à la fin du tour 76 de (R). Le sujet (C) exige de (R) une opinion directe sur le travail en soi. (R) met fin volontairement à ce sujet.
4éme. partie: du début du tour 77 de (C) à la fin du tour 95 de (C).Le sujet (C) ques-

tionne(R) sur les objectifs de l'enregistrement qu'ils réalisent.
5ème. partie:du début du tour 95 de (C)à la fin du dernier tour 105 de (R).Le sujet (C) introduit un nouveau thème commentant les circonstances externes à l'espace exterieur de l'enregistrement.Le sujet (R) met fin volontairement à ce thème et à la conversation.En fonction de ces différentes parties, nous présentons la distribution des diverses fins de tour et les durées des pauses qui les séparent du début du tour suivant.

ANALYSE EXPÉRIMENTALE

L'enregistrement obtenu a été analysé par oscillographie donnant la représentation simultanée a) du signal global, b) de la fréquence fondamentale, c) de l'intensité et d) du signal jusqu'à 1500 Hz.La vitesse d'analyse, dans ce cas, a été de 50cs par cm
L'analyse ainsi obtenue de ces paramètres sur quatre lignes oscillographiques a permis une précision de segmentation et une quantification précise du signal sonore, (seulement cinq cas ont été techniquement impossibles de segmenter).
Nous ne présentons ici que les résultats de l'analyse faite selon le type d'interaction déterminé par la fin de chaque tour de rôle sur le début du tour suivant et la variable temporelle relative à la pause qui les sépare. Le type d'interaction est défini par le degré d'imposition que le locuteur exerce sur l'autre quant au moment de prendre de prendre la parole.
Ainsi nous considérons que les fins de tour peuvent être caractérisées par des marques de contrainte qui obligent l'interlocuteur à répondre. Celles-ci peuvent être:

- -des questions explicites gramaticalement(Q).
- -des déclarations complètes (Dc),qui marquent explicitement la fin du tour.
- -les appuis (A) qui sont les expressions linguistiques qui permettent au locuteur de reprendre la parole.

Nous considérons, en outre, les fins de tour caractérisées par des marques d'insécurité. Celles-ci peuvent permettre la prise de parole malgré la volonté du locuteur.Ainsi nous utilisons plusieurs degrés de marques d'insécurité:

- -0 :absence d'insécurité
- -1 :insécurité implicite (hésitation, pause après déclaration incomplète -Di-)
- -2 :insécurité explicite(expressions comme'je ne sais pas').

Nous considérons que les fins de tour du premier groupe sont des stratégies fortes de donner la parole; les fins de tour à degré d'insécurité 2 (explicite) sont des stratégies moyennes) et finalement les fins de tour marquées d'indice d'insécurité 1 sont des stratégies faibles de donner la parole, selon la classification utilisée par Fœrch et Kasper (4). Nous présentons également les valeurs moyennes de durée des pauses entre les tours de parole,quand elles sont constituées de silence, absolu ou accompagné d'éléments vocaux paralinguistiques tels que rire, toux ou respiration bruyante.
Nous tiendrons ici compte ici des tours qui se superposent(supp) et des coupures (Di 0)

ANALYSE DES RÉSULTATS

Si nous observons les fins de tour de chaque sujet qui ont permis à l'autre de prendre la parole nous obtenons en pourcentage les données du tableau 1.

fin de tour	sujet R	sujet C	total
Di 1	16%	22%	38%
Q	8%	16%	24%
Dc 0	11%	3%	14%
A	7%	6%	13%
Supp	3%	3%	6%
Di 0	3%	2%	5%

Tableau 1- Définition des fin de tour

Nous pouvons conclure que les fins de tour qui ont permis le plus souvent la prise de parole de l'interlocuteur sont les déclarations incomplètes avec marqueurs d'insécurité implicite.Ceux-ci sont plus nombreux pour (R) que pour (C).Viennent ensuite les questions (24%) celles de (C) étant le double de celles de (R).
Les déclarations finalisées sans incertitude occupent 14% des cas,le sujet (R) utilisant 11% pour seulement 3% des fins du sujet (C).Les formes d'appui représentent 13% des cas et sont utilisées presque également par les deux sujets.
Les prises de parole par superposition ou par coupure de déclarations non porteuses d'insécurité ne représentent que 6% et 5% des cas respectivement, ce qui vient confirmer une des règles considérées générales de l'interaction verbale qui est la basse acceptabilité de superposition de parole.
Nous devons cependant tenir compte de la division du discours déjà proposée pour vérifier comment ces fins de tour ssont distribués au long de ces parties.
ous présentons également les données des valeurs moyennes des pauses pour les deux sens de l'interaction verbale:

- R-C : représente la valeur moyenne de la pause quand C prend la parole après R
- C-R : représente la valeur moyenne de la pause quand R prend la parole les durées étant indiquées en centième de seconde.

L'ensemble des donnés est présenté au Tab.2.

Partie	Q		Dc0		A		Dil		pause	
	R	C	R	C	R	C	R	C	R-C	C-R
1ère	1	1					1	1	50	85
2ème	6	5	3	1	3	5	2	6	187	84
3ème	2	6	5	2	3	2	9	5	93	74
4ème		3	4	1	1	3	3	2	106	66
5ème		2	2			1		2	50	90

Tableau 2-Distribution des fins de tour par partie du discours.Durée en cs.

Si nous observons la distribution des fins de tour pour les parties qui constituent la conversation en soi,soit les parties 1,2 et 3, nous pouvons dire que les questions de R se situent surtout dans la partie 2,alors que celles de C se trouvent dans 2 et 3 qui portent sur son propre travail.Les déclarations complètes se trouvent pour R distribuées au long des trois parties déterminant une constante de son discours.
Les déclarations incomplètes marquées d'insécurité implicite et qui ont permis la prise de parole malgré le locuteur se situent pour le sujet R le plus significativement dans la partie 3 (9 sur 14), partie où il lui est demandé une opinion.Par contre,le sujet C a ce type de fin de tour presque également dans les parties 2 et 3,celles qui portent sur son travail.
Si, d'autre part, nous observons les durées moyennes des pauses C-R, celles où C répond à R nous pouvons affirmer qu'il y a une tendance générale de régularité de ces durées(entre 66 et 90 cs).Le temps moyen des pauses diminue également,pour R au long des parties (85-84-74-66), mais la durée la plus longue se trouve dans la séquence finale (90 cs).
Le sujet C marque une plus grande variation (entre 50 et 187 cs) et la moyenne de durée est plus élevée que celle de R.Nous pouvons également remarquer que les temps de la partie iniciale et de la partie finale (1-5).
Cependant ces temps sont le s plus court du sujet C et les plus longs du sujet R.Ce fait peut amener à considérer que le temps de pause de la conversation sans thème est spécifique et constant pour chacun des sujets.
C' est dans la discussion que les différences se marquent pour les sujets.Ainsi ,dans la deuxième partie les temps moyens qui séparent les réponses de C de la fin de celles de R sont de 187 cs, alors que celles de R viennent 84 cs après celles de C .L'allongement des pauses est considéré comme un marqueur de difficulté à trouver la réponse.Si nous associons cette différence entre les pauses des deux sujets et le type de fin de tour déjà déterminné pour chacun d'eux,nous pouvons conclure que C traduit une grande insécurité manifestée par des tratégies fortes de passer la parole et par les longs temps de pause qui précèdent ses réponses.
Le sujet R démontre une certaine régularité d'emploi de stratégies surtout marquées par les déclarations complètes comme forme de passer volontairement la parole et par des temps de réponse à C très courts et réguliers.Ces caractéristiques peuvent se vérifier dans les différentes parties et pour chaque sujet.

CONCLUSION

Les données de l'analyse montrent, donc,qu'il est possible de systématiser les indices d'éxécution quant à la conduction des tours de parole dans une conversation spontanée.
Ces indices sont ceux qui caractérisent les fins de tour leur permettant des stratégies différentes de donner la parole à l'autre, volontairement ou involontairement.
D'autre part les indices temporels qui séparent les tours de parole marquent également des stratégies différentes quant à la forme de prendre la parole.
Ces indices sont également révélateurs de la relevance conditionnelle,comme définie par Sacks (7) qui rend compte de l'influence de rôles sociaux des interlocuteurs dans la régulation de l'interaction verbale.

RÉFÉRENCES

(1) Delgado Martins,M.R. Sept Etudes sur la Perception. INIC, Lisboa. 1986
(2) Goldman-Eisler,F. Psycholoinguistics:Experiments in Spontaneous Speech.Academic Press.London-New-York. 1968
(3) Fromkin,V. Speech Errors as Linguistic Evidence. Mouton. The Hague. 1973
(4) Fœrch,C.,Kaspar,G. "On Identifying Communication Strategies in Interlanguage Production".Strategies and Interlanguage Communication, 211-248.Longman. 1983
(5) Searle,J. Speech Acts. Cambridge Press 1970
(6) Labov W.,Fanshel,D. Therapeutic Discourse:Psychoterapy as Conversation.Academic Press.New-York. 1977
(7) Sacks,H."An inicial Investigation of the Usability of Conversational Data for Doing Sociology" Studies in Social Interaction, 31-74. Free Press.New-York.1972

A QUANTITATIVE SURVEY OF NUCLEAR TONE VARIATION IN ENGLISH

TERTTU NEVALAINEN

University of Helsinki, Department of English
Hallituskatu 11, SF-00100 Helsinki, Finland

ABSTRACT

Nuclear tone frequencies were analysed in the London-Lund Corpus of Spoken British English, and compared with five previous RP studies and some regional data. Discourse specific variation can be detected in RP, especially in the distribution of the fall as opposed to the rise and the fall-rise. Some regional differences also appear to be influenced by discourse type.

INTRODUCTION

It is widely held that regional differences in English can be encoded in the distribution of nuclear tones and hence be detected in their relative frequencies in discourse [1]. It seems, however, that more information of the range of variation in standard English is needed before a satisfactory comparison can be made interdialectally. That our generalizations are all too often made on a rather narrow data base is shown by the account of past work on standard British English (RP) in section 2, below.

This study approaches nuclear tone distribution from the point of view of discourse type variation. In order to find out how consistent the distribution of nuclear tones remains in a selection of different discourse types, a quantitative analysis was carried out of the nuclear tones in the main text categories of the London-Lund Corpus of Spoken English (LLC). The corpus, which comprises c. 435,000 running words, is described in more detail in section 3.

The results of this study are discussed in section 4. They show significant variation e.g. in the frequencies of the simple falling and rising tones between the different discourse types. Bearing this internal variation in mind, a comparison is made in section 5 of the average distribution of nuclear tones in spontaneous conversation in RP and nuclear tone data on Tyneside, Shetland, and General American English.

PAST WORK ON TONE DISTRIBUTION IN R.P.

Table 1 presents the results of five previous studies on nuclear tone distribution in RP. It is arranged according to the frequency of the simple falling tone. Although all the studies in Table 1 are based on a head-plus-nucleus analysis of tone units, the figures are not ideally comparable because in most cases no mention is made of the treatment of subordinate tone units in the study. Nevertheless, all studies distinguish a nearly identical inventory of nuclear tones: three simple tones (fall, rise and level), two complex tones (fall-rise and rise-fall) and two compound ones (fall-plus-rise and rise-plus-fall). When other compounds are distinguished (e.g. fall-plus-level) they are included under the category of <u>others</u> in Table 1.

TABLE 1. Relative frequencies of nuclear tone types in five R.P. studies.

Source/Tone (%)	Fall \	Rise /	Fall-Rise \/	Fall+Rise \+/	Level -	Rise-Fall /\	Rise+Fall /+\	Other
Davy (1968 C)	58.7	16.1	7.4	5.1	8.0	4.2	0.4	?
Iivonen (1984)	55.0	13.3	14.6	7.0	4.0	4.4	0.7	1.0
Crystal (1969)	51.2	20.8	8.5	7.7	4.9	5.2	1.7	-
Quirk (1964)	51.0	24.0	6.7	9.0	2.0	3.8	0.6	2.9
Davy (1968 R)	50.2	24.6	11.1	5.5	5.5	2.1	0.6	?
Altenberg (1987)	47.2	26.6	9.8	9.3	4.9	0.8	0.3	1.1

The individual studies in Table 1 can be briefly described as follows. Davy (1968 C) is based on a sample of conversation, but the sample size is not specified [2]. Iivonen <u>et al</u>. (1984) present the average distribution of nuclear tones in the first two text categories of the London-Lund Corpus (c. 140,000 words), which consist of conversations between intimates and distants [3]. Subordinate and incomplete tone units are however excluded from the analysis. The figures given in Crystal (1969) average over c. 30,000 words of conversation [4], while those in Quirk <u>et al</u>. (1964) derive from two panel discussions of about 5,000 words each [5]. The category <u>others</u> in this case also includes doubtful instances. The sample size of the reading data analysed in Davy (1968 R) is not further specified [2]. Lastly, the figures in Altenberg (1987) represent the nuclear tone distribution in a popular lecture (text 12.6 in the LLC) which consists of 4,877 words [6]. Although the distributional differences in Table 1 are not very great, conversations seem to show a higher proportion of the simple fall than the monologue texts. Conversely, the simple rise is slightly favoured in the two monologues studied in Davy (1968 R) and Altenberg (1987).

COMPOSITION OF THE CORPUS

The London-Lund Corpus of Spoken English is a collection of c. 435,000 running words of educated British English (RP) in orthographic transcription with prosodic analysis. It was produced at the Survey of English Usage, University College London, and the computer tape version used in the present study was compiled at the Survey of Spoken English, University of Lund. The prosodic analysis distinguishes seven basic nuclear tones: fall (\), rise (/), fall-rise (\/), rise-fall (/\), level (-), fall-plus-rise (\+/) and rise-plus-fall (/+\). Pitch range variation is encoded separately in the booster system [7].

The corpus contains twelve text categories which consist of a varying number of individual texts (altogether 87 texts of about 5,000 words each). The present study examines nuclear tone distributions primarily at the main category level. Categories S.5 and S.10 are, however, further subdivided into two parts because there is reason to believe that their tone distributions significantly covary with the subdivisions. Further internal variation may, of course, also occur but a detailed analysis of the individual texts falls outside the scope of the present study. Table 2 presents a brief description of the fourteen text categories examined.

NUCLEAR TONES IN THE CORPUS

The distributions of nuclear tones in the main text categories of the LLC are shown in Table 3. The results were obtained using a computer program which identifies tone unit patterns on the basis of the LLC transcription. The figures presented in Table 3 also include the tones in subordinate tone units, which are counted as independent, as well as those in incomplete tone units and vocalizations (e.g. [\m], [/mhm]).

Table 3 indicates a fairly large range of variation in the distribution of the main tones, the simple fall and the simple rise. The maximum frequency of the falling tone (>70% of all tone tokens) occurs in private

TABLE 2. Composition of the London-Lund Corpus of Spoken English.

LLC Text Category/	Characterization and approximate word count
S.1:	Surreptitiously recorded spontaneous face-to-face conversations between intimates and distants; 70,000 words
S.2:	Same as in S.1, 70,000 words
S.3:	Same as in S.1, 30,000 words
S.4:	Mostly non-surreptitious conversations between intimates and equals; 35,000 words
S.5.1-7:	Non-surreptitious public discussions between equals; 35,000 words
S.5.8-11:	Non-surreptitious private conversations between equals; 20,000 words
S.6:	Non-surreptitious conversations between disparates; 30,000 words
S.7:	Surreptitious telephone conversations between personal friends; 15,000 words
S.8:	Surreptitious telephone conversations between business associates; 20,000 words
S.9:	Surreptitious telephone conversations between disparates; 15,000 words
S.10.1-4:	Surreptitious telephone conversations between disparates; 15,000 words
S.10.5-8:	Spontaneous sports commentary (cricket, football, boxing, horse racing); 20,000 words
S.11:	Other spontaneous commentary (e.g. a royal wedding, state funeral, launching of a ship, physics demonstration); 20,000 words
S.12:	Spontaneous oration (e.g. a case in court, dinner speech, recordings in the House of Commons); 25,000 words
	Prepared but unscripted oration (e.g. sermons, university lectures, and political speeches); 30,000 words

TABLE 3. Distribution of nuclear tones in the London-Lund Corpus.

LLC Text/Tone (%) Category	Fall \	Rise /	Fall-Rise \/	Fall+Rise \+/	Level -	Rise-Fall /\	Rise+Fall /+\	Other
S.5.8-11	70.9	7.8	8.8	4.8	4.3	2.4	0.4	0.6
S.4	62.9	11.1	11.7	4.7	4.2	4.2	0.3	0.9
S.3	60.3	11.6	13.8	6.1	4.8	2.5	0.3	0.6
S.7	59.6	17.3	11.8	5.4	2.4	2.7	0.3	0.5
S.2	57.1	12.1	15.4	6.0	3.5	4.4	0.6	0.9
S.1	56.3	13.9	12.5	7.0	4.2	4.4	0.7	1.0
S.5.1-7	56.3	20.9	6.7	8.8	3.4	3.0	0.7	0.2
S.12	53.3	23.0	8.5	6.8	4.7	2.1	1.0	0.6
S.6	52.7	14.9	15.3	7.3	5.3	3.5	0.3	0.7
S.11	51.9	16.6	14.2	9.1	4.3	2.0	0.8	1.1
S.8	50.5	19.9	15.8	6.8	2.6	3.0	0.4	1.0
S.10.5-8	49.6	23.5	13.2	4.5	6.8	0.9	0.4	1.1
S.9	46.0	18.5	20.0	7.0	4.5	2.4	0.3	1.3
S.10.1-4	40.0	32.9	12.1	7.0	5.8	0.8	0.5	0.9

conversations between intimates and equals. The fall is slightly less frequent (c. 55%-60%) in the conversations involving both intimates and distants. Its incidence falls below 50% in radio and TV commentary, and telephone conversations between distants, reaching its lowest point (40%) in sports commentary. Conversely, simple rising tones are most frequent in all kinds of spontaneous commentary, prepared oration and public debates (c. 20%-33%).

With one exception all the broadcasters and commentators in S.10.1-8 are male (N = 23). No women appear in S.12, and male speakers also dominate in the public discussions in S.5.1-7. Hence the data do not support any purely sex-determined motivation for the high frequency of the simple rising tone. This does not, of course, mean that tonal distinctions could not be used to reinforce sexual stereotypes e.g. in comedies [8].

Usually the distribution of a nuclear tone cannot be directly matched with its functions on the basis of mere frequency data. It may however be argued that, in radio and TV commentaries, rising tones are probably used to promote textual cohesion [9]. Hence the rising tone types may also cooccur with statements and new information. A similar strategy would seem to apply to the lecture examined in Altenberg (1987), and reading intonation. In telephone conversations between disparates (e.g. S.9) the rather high proportion of rising tone types could, on the other hand, mark polarity questions and formal politeness.

REGIONAL VARIATION

It has been pointed out in the literature that regional differences can be detected, for instance, in the relative distributions of the rising and level tones [10]. However, any systematic comparison of RP with other varieties of English is complicated by the scarcity and disparity of the quantitative material available. Some highly tentative comparisons are, nevertheless, possible. In Table 4, the RP distribution represents the average of the first three text categories in the LLC (S.1-3), including subordinate tone units and vocalizations (cf. Table 3).

TABLE 4. A comparison of nuclear tone type frequencies in four varieties of English.

Variety/Tone (%)	Fall \	Rise /	Fall-Rise \/	Fall+Rise \+/	Level -	Rise-Fall /\	Rise+Fall /+\	Other
RP Standard C	57.3	12.8	13.9	6.4	4.0	4.1	0.6	0.9
GA Dialogue R	53.7	9.6	15.8	-	20.9	0.0	-	-
Shetland C	52.1	26.5	1.5	5.2	11.3	3.0	0.3	0.1
Shetland N	34.9	33.0	3.1	9.9	10.4	7.0	0.6	1.1
Tyneside I	28.0	17.0	7.0	11.0	21.0	15.0	1.0	-

(C = conversation, R = reading, N = narrative, I = interview)

Oreström's prosodically transcribed corpus (1985) was used to calculate the frequency distribution for Shetland English. The data were divided into two parts because conversation (= C; 2 speakers, c. 770 tone units) differed considerably from narration (= N; 20 speakers, about 3,960 tone units) [11]. The Tyneside figures show the average tone distribution of the data given in Pellowe and Jones (1978) for two sets of interviews (= I; 20 speakers, 4,066 tone units) [12]. The Shetland and Tyneside transcriptions basically follow the LLC notation system, which will facilitate their comparison (but also conceal possible varietal differences in prosodic realization). By contrast, the General American (GA) figures quoted from Pike (1945) are based on a form of notation that does not recognize, for instance, the British notion of compound tones. Pike's figures derive from a detective story dialogue (= R; 1 speaker, 804 contours) [13].

Space only permits some brief comments on Table 4. The simple fall appears to be most infrequently represented in the Tyneside data (28%), but the complex rise-fall most frequently (15%). To what extent they could be functionally equivalent remains an open question. The low frequency of the fall in the Shetland narratives may be partly due to the cohesive function of the rising tone types. In this respect the Shetland conversation sample differs radically from the narrative monologues.

The level tone is remarkably frequent in the regional data, especially in Tyneside and General American. In the American data it could perform a cohesive function: 163 out of the 168 instances of the tone in the sample are found in the middle of sentences [13]. Another factor that might covary with nuclear tone distribution is tone unit length. In Oreström's data, the average length is 5.2 words, in the LLC 4.3 words, and in Pike's General American sample about 3 words. How far the differences are purely discourse specific in a given variety can only be settled when more comparative data become available.

ACKNOWLEDGEMENTS

I would like to thank Visa Rauste and Hannu Hartikka from the Helsinki University Computing Centre for their assistance in the computer processing of the LLC data. Thanks are also due to Bengt Altenberg for kindly making his study of Text 12.6 available to me before it was published.

REFERENCES

[1] J. Wells (1982), Accents of English, 1, An Introduction. Cambridge: Cambridge University Press, p. 90-91.

[2] D. Davy (1968), A Study of Intonation and Analogous Features as Exponents of Stylistic Variation, with Special Reference to a Comparison of Conversation with Written English Read Aloud. Unpublished MA Thesis, University of London, cited in [4].

[3] A. Iivonen, R. Aulanko, H. Kaskinen, T. Nevalainen (1984), Intonaatioteorioista (Publications of the Helsinki University Department of Phonetics, 34b). Helsinki, p. 116.

[4] D. Crystal (1969), Prosodic Systems and Intonation in English. Cambridge: Cambridge University Press, p. 225.

[5] R. Quirk, A. Duckworth, J. Svartvik, J. Rusiecki, A. Colin (1964), "Studies in the Correspondence of Prosodic to Grammatical Features in English." Proceedings of the Ninth International Congress of Linguists, ed. H. G. Lunt. The Hague: Mouton, 679-691, p. 681.

[6] B. Altenberg (1987), Prosodic Patterns in Spoken English, Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion (Lund Studies in English, 76). Lund: Lund University Press.

[7] J. Svartvik, R. Quirk (1980), A Corpus of English Conversation (Lund Studies in English, 56). Lund: CWK Gleerup, p. 21-25.

[8] C. Gussenhoven (1986), "The Intonation of 'George and Mildred': Post-Nuclear Generalisations." Intonation in Discourse, ed. C. Johns-Lewis. London: Croom Helm, 77-123, p. 82-84.

[9] M. A. K. Halliday, R. Hasan (1976), Cohesion in English. London: Longman, p. 271-273.

[10] A. Cruttenden (1986), Intonation. Cambridge: Cambridge University Press, p. 138-144.

[11] B. Oreström (1985), A Corpus of Shetland English (Stockholm Studies in English, 64). Stockholm: Almqvist & Wiksell International.

[12] J. Pellowe, V. Jones (1978), "On Intonational Variability in Tyneside Speech." Sociolinguistic Patterns in British English ed. P. Trudgill. London: Edward Arnold, 101-121, p. 109.

[13] K. Pike (1945), The Intonation of American English (University of Michigan Publications, Linguistics, 1). Ann Arbor: University of Michigan Press, p. 155.

ANTONIO MANGIA LA ZUPPA INGLESE
PHONETIC AND PHONOLOGICAL ASPECTS OF ITALIAN SENTENCE INTONATION

OLGA PROFILI

St.Hugh's College
Oxford, OX2 6LE
GREAT BRITAIN.

PHILIPPE MARTIN

Experimental Phonetics Laboratory
300 Huron Str., Toronto, Ontario,
CANADA M5S 2X6

ABSTRACT

Within the framework of a phonosyntactic model of Italian sentence intonation, pitch movements of stressed syllables can be predicted from the syntactic structure (if congruence is assumed between syntax and prosody). New intriguing data seems to contradict some theoretical predictions. It is shown here that the observed facts can be better understood using the principle of eurhythmicity.

INTRODUCTION

The description of sentence intonation, as shown by the already existing mass of literature in the domain, has been the aim of many theoretical approaches. In the present paper we will be dealing with the phonosyntactic model of sentence intonation which operates on the specific relations of dependency existing between the syntactic and the prosodic structure of the sentence.

The model [1],[2] is based on the fact that stressed syllables are perceptually the most prominent. This reduces the continuum of Fo, intensity and duration to sequences of prosodic contours located only on stressed syllables. It is important to notice that the model deals with pitch changes and not differences in pitch levels.

Each prosodic contour can be described phonologically by means of specific phonological features, which have been postulated as follows [3]:

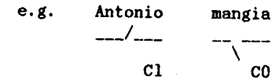
[± Extreme] : the contour attains an extremely low (in the case of statements) or an extremely high (in questions) frequency level as compared to the other contours.

[± Rising] : when the fundamental frequency rises of falls.

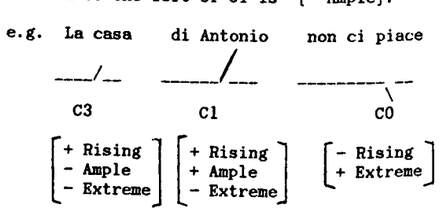
[± Ample] : when the melodic variation is large (or restrained) as compared to the variation of similarly rising or falling contours.

The phonosyntactic model of sentence intonation has also been applied to Italian [4], and two rules are used to determine the prosodic structure of an utterance :

- Rule A indicates a contrast in slope. If the final contour (denoted C0) is falling, then the contour located to its left and at the same level in the prosodic structure (denoted C1) is rising.



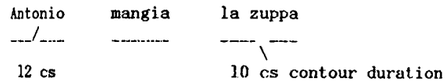
- Rule B indicates a difference in the amplitude of melodic variation. This second rule differentiates the melodic contours of two prosodic words, which are at different levels in the structure. In other words, if C0 is falling and C1 is rising, than C3 equally rising and located to the left of C1 is [- Ample].



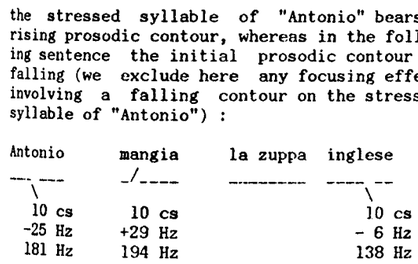
In more recent work [5] possible variations concerning the prosodic contours of Italian sentence intonation were reported. However, this paper deals with new intriguing data which has not been observed previously, and seemed at first puzzling. By comparing two Italian sentences having the same subject noun phrase composed of only one word, but different object noun phrases, we have noticed that the initial prosodic contour located on the stressed syllable of the subject noun phrase was rising

in the first case and falling in the second.

In other words in a sentence such as :



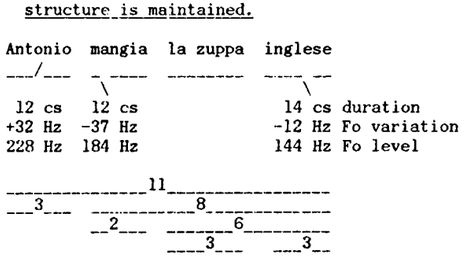
the stressed syllable of "Antonio" bears a rising prosodic contour, whereas in the following sentence the initial prosodic contour is falling (we exclude here any focusing effect involving a falling contour on the stressed syllable of "Antonio") :

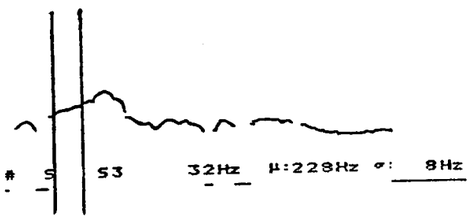


Within the framework of the phonosyntactic model of intonation, pitch movements of stressed syllables can be predicted from the syntactic structure of the sentence, if congruence between the syntactic and the prosodic structure is assumed. However, it has been suggested recently [6] that speakers sometimes prefer a prosodic structure with a rhythmically balanced division of the prosodic words to another prosodic structure, congruent to the syntactic structure but rhythmically unbalanced. In other words, if we release the constraint of congruence, a reasonable criterion for choosing a specific prosodic structure from among all the possible patterns could be based on eurhythmicity.

The hypothesis of eurhythmicity favours prosodic structures that balance the number of syllables of the prosodic words within a sentence. The sentence mentioned above can thus be pronounced by using two different prosodic structures (we indicate the number of syllables below each prosodic group) :

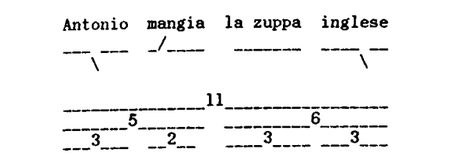
a. Congruence between syntax and the prosodic structure is maintained.





If congruence between the two structures is maintained, the result is a rhythmically unbalanced division of the prosodic words at the first level in the structure (8-3=5 syllables difference).

b. A eurhythmic prosodic structure is preferred.

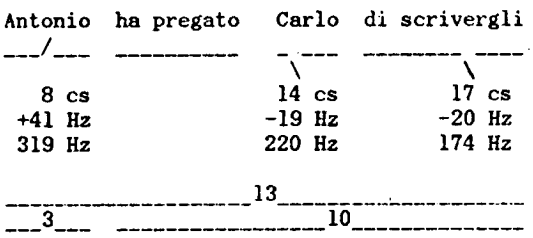


If the constraint of congruence is released and the principle of eurhythmicity is applied, the number of syllables of the prosodic words at the first level in the structure can be balanced by choosing a different prosodic structure (6-5=1 syllable difference).
In order to test the principle of eurhythmicity, we have examined another sentence with a longer verb phrase (square brackets are used to indicate the prosodic structure):

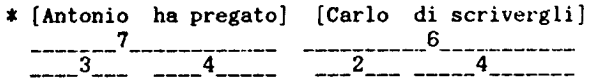
[Antonio] [ha pregato Carlo di scrivergli]

We have observed that in the case of this sentence congruence between the syntactic structure and the prosodic structure is not easily maintained. This might be due to the striking unbalanced

rhythmical effect (10-3=7 syllables difference):

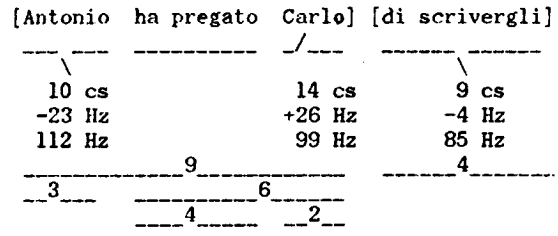


The most eurhythmic prosodic structure of the above sentence would be the following (7-6=1 syllable difference) :



However, as the asterisk shows, this is not possible in Italian.
As it also occurs in French [6] some prosodic structures are unacceptable, because they contradict the syntactic structure of the sentence at the lowest level of the syntactic division. In such case a lowest level syntactic clash (LLSC) appears. When a LLSC occurs a eurhythmic prosodic structure cannot be chosen, because it contradicts the syntactic structure to such a degree that the intonation pattern becomes unacceptable.

We have noticed that speakers tend to choose the following
eurhythmic prosodic structure (9-4=5 syllables difference) :



This rhythmic division is more balanced than the one used when congruence between the syntactic structure and the prosodic structure is maintained.

Our observations show that the prosodic structures of Italian sentences can be independent from syntax, provided that their choice is based on a principle of eurhythmicity that divides the sentence in a rhythmically balanced number of syllables. However, eurhythmic prosodic structures generate acceptable prosodic contours only when the LLSC condition is not violated.

REFERENCES

[1] Ph. Martin, Analyse phonologique de la phrase française, Linguistics, 146: 35-68, 1975.

[2] Ph. Martin, Pour une théorie de l'intonation: l'intonation est-elle une structure congruente à la syntaxe?, M. Rossi et al., L'Intonation de l'Acoustique à la Sémantique, Paris, Klinksieck, 234-271, 1981.

[3] Ph. Martin, Phonetic realizations of prosodic contours in French, Speech Communication, 1: 283-294, 1982.

[4] Ph. Martin, L'intonation de la phrase en italien, Studi di Grammatica Italiana, VIII: 395-417, 1978.

[5] O. Profili, Acoustic investigation of intonation in two regional varieties of Italian: preliminary results, PROPH, II (in print April 1987).

[6] Ph. Martin, Prosodic and Rhythmic Structures in French, Linguistics, (in print 1987).

INTONATION PATTERNS IN GREEK

Th. MALAVAKIS

Institut de la Communication Parlée
L. A. CNRS 368
Institut de Phonétique
Université de Grenoble III
38040 GRENOBLE CEDEX
FRANCE

ABSTRACT

Prosodic contours are constant according to the syntactic function of the prosodic words; the two essential contours (affirmative and interrogative modalities) are dependent of the stressed syllable and the place of the stress in the inner constituent elements.

1. INTRODUCTION

The aim of this work is to present the relationships existing between intonation and syntax in Greek. Except for the three interrogative pronouns :

[τί] "what, which"
[ποι-ός, -ά, -ό] "who"
[πόσ-ος, -η, -ο] "how much, how many"

(introducing a kind of wh-questions), Greek does not possess any particular morphological form for interrogative modality. In addition to that, the syntactic form of a statement remains unchanged in all modality transformations. Thus, only intonation can account for the modality of the statement.

Ph. Martin [1] proposes the features of the contours for French as follows:

[± Long] : the length of the contour

[± Rising] : the direction of the fundamental frequency.

[± Ample] or

[± Restrained] : the amplitude of the melodic variation (large or restrained).

Three different kinds of contours can be found in Greek : final contour in a declarative sentence, final contour in an interrogative sentence, and prefinal intermediate contours, denoted C0a, C0i and C (C1, C2, C3 etc) respectively. But, as was discovered [2], the final interrogative contour is to be sub-divided according to the place of the stressed syllable (oxytones and paroxytones words).

2. PROCEDURE

It has been proved that prosodic and syntactic structures are two autonomous systems but linked to each other. Semantics could also be added to them since it is not considered as an organized system with hierarchies and structures.

The corpus contains assertions and yes-no questions. Declarative sentences are completed with expansions on the right (e.g. SV -> SVO etc.). The questions have the same syntactic structure as the associated declarative sentences. The place of the stress and the number of syllables of each word had been taken into account as well.

The words chosen for the analysis, had the following acoustic structures :

oxytones : C1 V1 C2 <u>V1</u>

paroxytones : C1 <u>V2</u> C2 V2

proparoxytones : <u>V1</u> C1 V2 C3 V3

C1, C2 : unvoiced consonants
C3 : /r/

The sentences were read by Athenian speakers, in order to avoid regional varieties. F0, intensity and duration were measured on the oscillograms and computer processing permitted us to extract the pattern features.

3.RESULTS

Some preliminary results [2] shown that in Greek the final contours have the following features :

C0a [- Rising, ± Ample, + Long]

C0i paroxytones words [- Rising, - Ample, + Long]

C0i oxytones words [+ Rising, ± Ample, + Long]

The same corpus was used in this analysis, enriched with some supplementary sentences.

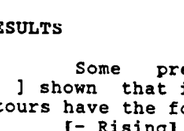
Two remarks could be made :
a/ There are two rules governing the contours slopes, and
b/ The penultimate stressed word plays a particular part in the questions.

a/ The contour slope rules.

Just like in Italian (which is also a mobile stress language), the prosodic structure of an utterance is influenced by the final contour.

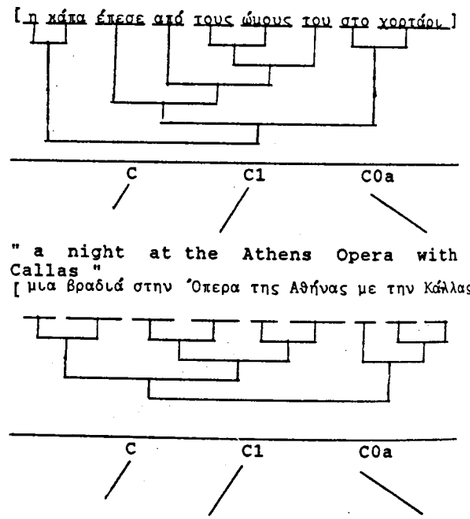
Rule 1 (for prosodic words belonging to the same level in the prosodic structure) : if the final contour is [- Rising], then the preceded contour is [+ Rising] and viceversa.

Example :
" the coat has fallen down "



Rule 2 (for prosodic words located on different levels): if the final contour is [+ Rising] and the preceded contour (C1) is [+ Rising], then the contour on the left of C1 has to be [+ Rising] but [- Ample].

Example :
" the coat has fallen from his shoulders unto the grass "

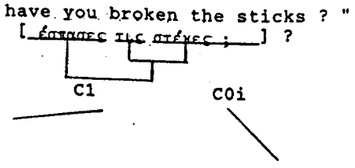


b/ The penultimate stressed words in a question.

The observed behaviour of the final contour is exactly the opposite of what we were expecting. As a matter of fact, instead of having a final contour [+ Rising] we have a C0i

[- Rising]
[+ Ample]
[+ Long]

Example :
" have you broken the sticks ? "



4. CONCLUSION

The goal of this paper was to present the intonation patterns of the two main modalities (affirmative and interrogative).

The results of the analysis provided evidence that the final contour exerts a strong influence on the preceded contours.

The case of the penultimate stressed words in a yes-no question is very interesting : as the C0i falls, the C1 keeps a relative rising (according to the rule 1); the latter does not start from the same low frequency level, although it attains the same high level, without being as inclined as the other C1 contours. This particular form of the final contour illustrates the interaction between the prosodic and the acoustic structure.

REFERENCES

[1] MARTIN Ph. "Résumé d'une théorie de l'intonation", Bulletin de l'Institut de Phonétique de Grenoble, Vol. VI, 57-87, 1977.

[2] MALAVAKIS Th."Συνταχτικά και επιτονικά φαινόμενα" (Syntactic and intonative phenomena), Proceedings of the VIth Congress of Greek Linguistics, Thessaloniki, 41-56, 1985.

TON SYLLABIQUE ET INTONATION DE PHRASE EN VIETNAMIEN

M.V.GORDINA

Chaire de phonétique
Université de Leningrad, Leningrad, URSS 199034

ABSTRACT

The study deals with the characteristics of 6 Vietnamese tones in open syllables and those ending in /i̯,/,/u̯/ or any nasal. The syllables occurred in medial and terminal positins. The data on melodic ranges, average fundamental frequency, duration and intervals of melodic contour of the 6 tones are presented. Throughout the utterances two--octave ranges occur; in the separate utterances, however, the range generally does not exceed one octave.There is great paradigmatic variation of tonal characteristics while syntagmatic tonal relations have a tendency to stability. Changes in duration are used as intonation boundary signal.

En vietnamien, langue syllabique et tonale, où la syllabe-morphème est l'unité de base du système de la langue /1,2/, il existe, dans les dialectes du Nord, six tons, qui sont réalisés en syllabes non--fermées (terminées par une voyelle, une semi-voyelle /i̯/ ou /u̯/ ou bien par une occlusive nasale). Les syllabes fermées (terminées par une occlusive sourde) sont prononcées avec l'un des deux tons seulement (le 5e ou le 6e). Dans la présente étude il ne s'agit que des six tons des syllabes non-fermées.

Dans les travaux précédents /3,4/ il a été établi que les caractéristiques tonales ne s'étendent pas à la consonne initiale de la syllabe; elles sont rattachées à la finale syllabique toute entière. Celle-ci peut comporter trois éléments au maximum; un noyau vocalique représenté par une voyelle simple ou une diphtongue et, optionnellement, la semi--voyelle /u̯/ devant le noyau vocalique et une consonne ou semi-consonne terminant la syllabe. Ce sont les qualités acoustiques de la finale qui font l'objet de la présente étude.

En syllabe-morphème isolée les six tons du vietnamien diffèrent par leur niveau mélodique, leur contour, leur intensité, la présence (ou l'absence) du coup de glotte, le timbre et la longueur (fig. 1). Celle-ci va de pair avec le contour mélodique ou avec le coup de glotte: le 4e ton, au contour complexe descendant-montant, est le plus long; le 6e ton, terminé par un coup de glotte, est le plus bref. La longueur doit donc être considérée comme une caractéristique non-pertinante /5/. Ainsi, les marques suprasegmentales qui en langues non-tonales relèvent de l'intonation de phrase appartiennent en vietnamien, ainsi que dans les autres langues du même type, au niveau des monèmes. L'intonation de phrase ne peut donc être que le résultat de variations de caractéristiques tonales propres aux monosyllabes formant la phrase.

La présente étude a été faite sur un corpus de 60 phrases isolées (propositions énonciatives et interrogatives) lues, dans un ordre aléatoire, par trois locuteurs originaires du Nord. Les phrases avaient la longueur de 4 à 12 syllabes; la distribution des tons était analoque à celle qu'on trouve dans les textes /6/.

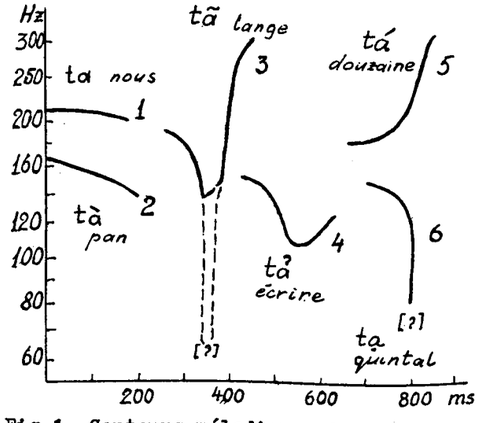


Fig.1. Contours mélodiques des six tons du vietnamien en syllabe isolée.

Les tons syllabiques tels qu'on les observe en syllabe isolée n'apparaissent, pour la plupart des cas, qu'en fin de phrase (ou de groupe de sens), c'est-à-dire à la pause; dans d'autres positions, ils subissent différentes modifications.

A la pause, le ton (la finale de syllabe) est sensiblement plus long qu'à l'intérieur du groupe de sens (tableau 1). De façon générale on peut dire que la frontière intonative entre deux groupes de sens (ou syntagmes) est marquée par un contraste temporel entre une syllabe longue finale du groupe et une syllabe plus brève au début du groupe suivant. A l'intérieur de la phrase (groupe de sens) la finale de syllabe est plus ou moins longue en fonction de la valeur sémantique du monème. Les mots-outils (prépositions, verbes semi-auxiliaires etc.) sont plus brefs que les mots "pleins" (tableau 2). Il en résulte que des groupes se forment comprenant deux, trois ou quatre syllabes dont l'une (la dernière pour la plupart des cas, mais quelquefois la première) est la plus longue. L'alternance des syllabes longues et brèves donne à chaque

Ton	Position	Loc.1	Loc.2	Loc.3
1	finale	170 ± 8	207 ± 8	191 ± 9
	non-fin.	112 ± 6	109 ± 5	123 ± 4
2	finale	203 ± 11	198 ± 7	195 ± 9
	non-fin.	125 ± 11	107 ± 14	124 ± 8
3	finale	202 ± 10	216 ± 18	180 ± 16
	non-fin.	120 ± 17	100 ± 9	102 ± 11
4	finale	199 ± 20	157 ± 16	170 ± 14
	non-fin.	109 ± 8	79 ± 10	110 ± 11
5	finale	193 ± 6	239 ± 12	184 ± 12
	non-fin.	96 ± 8	102 ± 8	106 ± 11
6	finale	149 ± 10	156 ± 15	156 ± 21
	non-fin.	129 ± 10	117 ± 8	143 ± 7

Tableau 1. Durée des tons (finales de syllabe) en ms: valeurs moyennes et limites de confiance (P = 0.05).

Mots-outils	vẽ/ve2/sur, de	của/kuo4/de
Locuteur 1	60 ± 4	81 ± 10
Locuteur 2	57 ± 15	50 ± 15

Tableau 2. Durée des tons dans les prépositions (ms): valeurs moyennes et limites de confiance (P = 0.05).

phrase un certain rythme:

Anh ấy học ở Anh 'Il étudie en Angleterre'

Bà ấy đi chợ lúc mấy giờ? 'Cette dame va (au) marché moment quelle heure?'

Ainsi, l'unité temporelle de la phrase vient de son organisation rythmique. L'unité mélodique de la phrase est due à son diapason mélodique. Le diapason de la voix déterminé à partir de réalisations des tons en syllabe isolée peut dépasser deux octaves /3,4/; ce résultat est confirmé par l'analyse des tons dans les phrases. Cependant, le diapason d'une phrase, même si elle contient des tons à valeurs extrêmes de fréquences, est généralement beaucoup plus restreint (fig.2, tableau 3). Dans le corpus étudié plus de 2/3 de phrases des locuteurs 1, 3 et près de 95% des phrases du locuteur 2 ont un diapason de 1 à 11 demi-tons.

Ce diapason étroit résulte des variations des tons à l'intérieur de la phrase. Les modifications sont plus ou moins importantes selon la durée de la syllabe. C'est la courbe mélodique des tons modulés (3e, 4e, 6e et dans une certaine mesure 5e) qui est surtout atteinte. Leurs contours mélodiques intégraux sont réalisés en fin de phrase (ou groupe de sens) et quelquefois dans les syllabes qui n'ont subi qu'une petite réduction de durée. Dans d'autres positions leurs contours sont plus ou moins déformés.

Parmi les trois locuteurs c'est le locuteur 1 qui a le débit le plus lent et qui garde le mieux les caractéristiques tonales au milieu du groupe de sens: le

Diapason		Loc.1	Loc.2	Loc.3
de la voix		30	22	25
dans la phrase	moyen	10	7	11
	I.V.	2—22	1—15,5	2,5—18,5

Tableau 3. Diapasons de la Fo (demi--tons); I.V. - intervalle de variation.

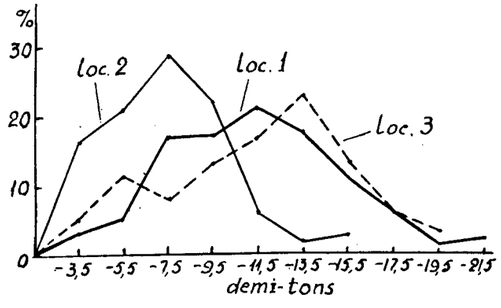


Figure 2. Distribution des diapasons mélodiques de phrase dans le corpus analysé.

coup de glotte au 6e ton est souvent conservé, au 4e ton on observe quelquefois le début de la partie montante du contour; le contour du 5e ton commence par un plateau mélodique suivi d'une brusque montée. Les locuteurs 2 et 3 (ainsi que le locuteur 1 en cas d'une plus grande réduction de longueur) modifient les contours tonals bien davantage. Au 3e et au 6e ton le coup de glotte disparaît complètement, il n'en reste (et pas toujours) qu'une brève descente mélodique dans un très petit intervalle; au 5e et au 3e ton l'intervalle ascendant est fortement réduit, la mélodie du 4e ton ne représente qu'une légère descente (fig.3).

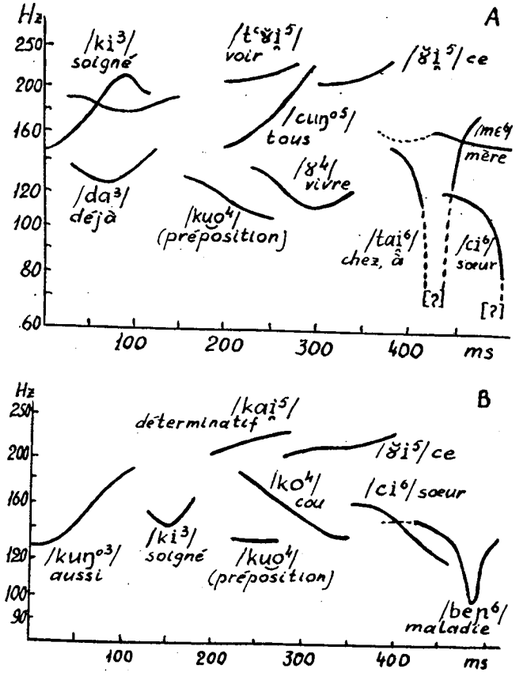


Figure 3. Variantes des tons modulés au milieu de la phrase
A - locuteur 1, B - locuteur 3.

Les modifications des tons 1 et 2, qu'on qualifie de statiques, sont moins importantes. Le 2e ton, au lieu d'une légère descente mélodique, peut présenter un contour plat. Au 1er ton, qui en syllabe isolée a un niveau mélodique stable, on observe, en fin de phrase, une courbe descendante ou quelquefois, dans les questions, montante. Le contour le plus caractéristique du 1er ton est donc réalisé généralement non pas à la fin, mais au milieu de la phrase, ce qui le distingue des autres tons du vietnamien(fig.4).

Ainsi, le contour mélodique de tous les tons, sauf le 1er, est plus ou moins réduit au milieu du groupe de sens; les intervalles de fréquence de la Fo y sont moins grands qu'avant la pause (tableau 4). Le maximum de la réduction tonale est atteint au cas où la finale de syllabe est la plus brève, c'est-à-dire dans les mots-outils, p. ex. cửa /kuo⁴/ (marque de l'appartenance), sẽ /sɛ³/ (marque du futur) etc. Dans ces mots, la mélodie caractéristique du ton est souvent remplacée par un contour plat, il n'y a que le niveau mélodique qui est conservé.

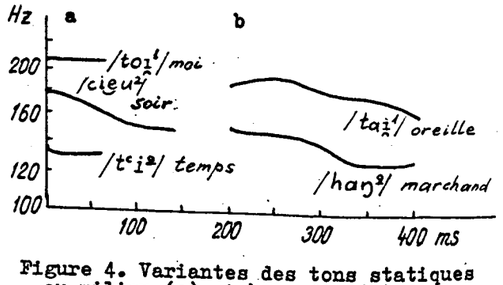


Figure 4. Variantes des tons statiques au milieu (a) et à la fin (b) de la phrase, locuteur 1.

Ton	Position	Loc.1	Loc.2	Loc.3
1	finale	1,4±0,6	2,6±0,4	3,8±1,1
	non-fin.	0,8±0,2	0,4±0,2	0,7±0,3
2	finale	2,9±0,8	4,0±0,7	4,5±1,2
	non-fin.	2,0±0,4	1,0±0,3	1,6±0,4
3	finale	11,0±3,0	3,1±1,0	4,4±1,2
	non-fin.	3,8±0,2	2,0±0,8	4,0±1,2
4	finale	5,6±0,8	5,0±0,9	7,7±2,0
	non-fin.	2,8±0,5	2,4±0,5	3,9±0,6
5	finale	7,0±0,8	5,7±1,2	5,5±0,8
	non-fin.	13,0±0,4	1,6±0,4	3,6±0,7
6	finale	8,3±1,3	7,1±1,5	9,6±2,2
	non-fin.	44,8±1,3	1,4±0,3	4,4±0,6

Tableau 4. Intervalles de la Fo: valeurs moyennes et limites de confiance en demi-tons (P=0,05). Pour le 3e et le 6e ton on a tenu compte de la chute mélodique avant le coup de glotte.

La réduction des contours rend les oppositions tonales moins nettes. Des contours semblables peuvent apparaître pour deux tons distincts appartenant tous les deux soit au même groupe de tons (modulés ou statiques), soit à des groupes différents: tons 1 et 2, 3 et 5, 4 et 6 ou bien 2e et 4e ton, 2e et 6e ton. Cependant,la réalisation des tons distincts avec des contours identiques est un cas rare (on le voit d'après les limites de confiance pour les intervalles de la Fo des tons susceptibles d'être confondus). Au cas où les intervalles de la Fo des deux tons sont les mêmes, ces tons restent distincts grâce aux différences de l'intensité (laquelle n'est pas analysée dans la présente étude) - tons 3 et 5 ou bien grâce à la forme du contour et au différent niveau mélodique - tons 2 et 4.

Le niveau mélodique des tons (déterminé comme la Fo moyenne de la finale de syllabe) n'est pas constant, les tons hauts et les tons bas ayant quelquefois la même Fo moyenne. La Fo moyenne varie à cause de l'aplatissement du contour des tons modulés, à cause de l'influence réciproque des tons de registres différents /5/, en fonction du registre plus ou moins élevé de toute la phrase. L'opposition des deux registres est néanmoins conservée; les limites de confiance sont loin de se recouper (tableau 5). Le dessin mélodique de la phrase toute entière est assez stable et à peu près le même chez les trois locuteurs (figure 5). Malgré les variations considérables, les oppositions tonales se maintiennent grâce à la compensation réciproque de différentes qualités acoustiques.

Ton	Fo	Loc.1	Loc.2	Loc.3
1	Fo m.	204 ± 8	216 ± 4	216 ± 6
	I.V. Hz	147-267	180-248	140-295
	dt	8,5	5,5	12,5
2	Fo m.	153 ± 9	180 ± 8	170 ± 7
	I.V. Hz	120-199	145-219	141-206
	dt	8,5	7,0	6,5
3	Fo m.	203 ± 12	199 ± 5	214 ± 15
	I.V. Hz	158-295	156-256	157-310
	dt	11,0	9,0	11,5
4	Fo m.	129 ± 9	160 ± 7	146 ± 11
	I.V. Hz	93-165	139-174	128-177
	dt	9,5	4,0	4,5
5	Fo m.	209 ± 18	209 ± 9	213 ± 13
	I.V. Hz	127-306	179-239	162-249
	dt	15,5	5,0	7,0
6	Fo m.	143 ± 12	174 ± 17	154 ± 10
	I.V. Hz	110-176	120-208	123-179
	dt	8,5	10,5	5,5

Tableau 5. Niveau mélodique des tons dans les phrases. Fo m. - la Fo moyenne et les limites de confiance, Hz (P=0,05); I.V. - intervalle de variation.

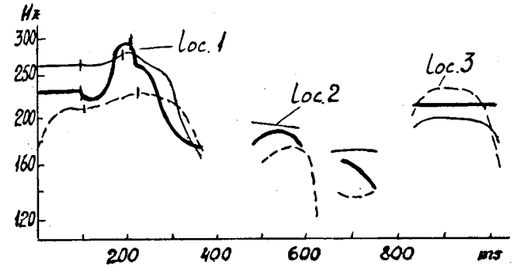


Figure 5. Contour de la phrase Cô ấy là chị của tôi /ko¹ ai⁵ la² ci⁶ kuo⁴ toi¹ / 'Cette femme est soeur à moi' lue par les trois locuteurs.

Les faits décrits permettent d'en tirer les conclusions suivantes:
1. La variabilité paradigmatique des tons a pour contrepartie une grande stabilité des relations tonales sur l'axe syntagmatique.
2. En vietnamien ainsi que dans les autres langues de ce type (voir p.ex. /7/) les effets rythmiques sont utilisés de la même manière que dans les langues non-tonales. Le rythme apparaît donc comme un procédé intonatif universel.

REFERENCES

/1/ I.S.Bystrov, Nguyễn Tài Cẩn, N.V. Stankevitch. Grammatika vietnamskogo iazyka. Leningrad 1975 (Grammaire de la langue vietnamienne).
/2/ V.B.Kasevitch. Fonologitcheskie problemy obchtchego i vostotchnogo iazykoznania. Moskva 1983 (Problèmes phonologiques de la linguistique générale et orientale).
/3/ N.D.Andreev, M.V.Gordina. Sistema tonov vietnamskogo iazyka. Vestnik LGU 1957 N 8 s.132-148 (Système des tons de la langue vietnamienne).
/4/ M.V.Gordina, I.S.Bystrov. Fonetitcheskii stroi vietnamskogo iazyka. Moskva 1984 (Phonétique de la langue vietnamienne).
/5/ Nguyễn Hàm Dương. Sistema tonov i spectry glasnykh vietnamskogo iazyka. Moskva 1963 (Système des tons et spectres des voyelles de la langue vietnamienne, thèse).
/6/ M.V.Gordina, I.S.Bystrov. Raspredelenie vietnamskikh tonov v slovare i v tekste. Utchenye zapiski LGU N 401, 1979, s.31-44 (Distribution des tons vietnamiens dans le vocabulaire et dans les textes).
/7/ R.Gsell. Sur la prosodie du thai standard: tons et accents. Paris 1979.

A PERCEPTUAL ANALYSIS OF RUSSIAN INTONATION: SOME ASPECTS

CECILIA ODÉ

Institute for Perception Research
P.O.Box 513, 5600 MB Eindhoven, the Netherlands

0. ABSTRACT

The method of stylizing pitch phenomena developed at the Institute for Perception Research (IPO) in Eindhoven, the Netherlands, has been shown to be applicable to Russian intonation. In this method, measured Fo curves are represented by the smallest number of straight-line pitch movements which will still yield perceptual equality with the original Fo curve. The aim of the research is to describe Russian intonation in terms of perceptually relevant pitch movements that combine to form complete stylized contours. On the basis of a perceptual analysis, problems on the phonetic and linguistic level can be solved.
Two of the problems encountered so far are highlighted:
1. In an excerpt of running speech a sawtooth shaped intonation pattern was found to occur frequently. This pattern seems to be associated with non-main pitch accents preceding the final pitch accent before a boundary.
2. In the same excerpt, pitch movements are described as moving between a high and a low reference line. However, in order to account for all pitch movements that reach a perceptually relevant point, more reference lines must be distinguished in the broad field between these lines.

1. INTRODUCTION

This paper deals with some aspects of the perceptual analysis of Russian intonation. Such an analysis is different from earlier approaches to the subject. To my knowledge, no perceptual analysis with a high degree of explicitness attained by means of the stylization method (see section 2) has been made of Russian intonation. This method has been successfully applied to Dutch, British English, American English and German, and has proved to be applicable to Russian as well.
On the basis of a perceptual description, more can be said about the linguistic function of Russian intonation. A linguistic description of Russian intonation, together with the explicit perceptual description, is a prerequisite for teaching intonation.
A study of intonational meaning presupposes a perceptual analysis. The reverse does not hold true.

2. METHOD

The method of stylizing pitch phenomena analyses intonation using the Linear Predictive Coding (LPC) analysis by resynthesis system (see Fig.1). This method is known as the Dutch School of Intonation; it has been developed at the Institute of Perception Research (IPO) in Eindhoven, the Netherlands, by A. Cohen, R. Collier and J. 't Hart (e.g. 1973, 1975).

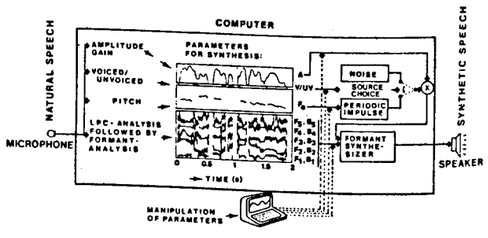


Fig.1: The analysis by resynthesis system (Nooteboom and Cohen, 1984)

The measured original fundamental frequency curves are represented by the smallest number of straight-line pitch movements which will still yield perceptual equality with the original Fo curve.
In the process from original Fo curve to the final stylized pitch contour three stages can be distinguished:
1. Measurement of the original Fo curve.
The original speech signal is digitized and analysed into thirteen parameters: voiced/unvoiced, amplitude, source frequency, five formants and their bandwidths (see Fig.1).
2. Comparison of the resynthesised speech signal with the original speech signal.
All whimsical, small fluctuations can now be removed from the pitch curve by listening and comparing the resynthesis and the natural signal. No differences should be heard between the two signals. Only then is a basis for further stylizations achieved.
3. Stylization of the pitch movements in the "clean" Fo curve.

In interaction with the computer, the Fo curve is reduced to the smallest number of straight-line pitch movements in such a way that perceptual equality between the original Fo curve and the stylization is still observed (see Fig.2). The stylized pitch movement can be made audible and compared with the original Fo curve of the same fragment of the speech signal. No differences are audible. Native subjects can verify the acceptability of the stylization.

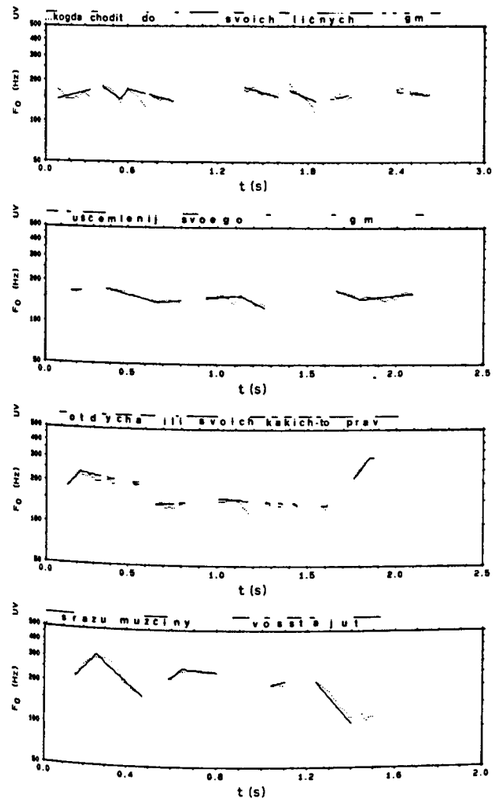


Fig.2: The original Fo curve (dotted) and close-copy stylization (solid) of the Russian spontaneous utterance (in word-to-word translation): ... as soon as it comes to their ... personal ... uh ... restrictions of their ... holiday or some of their rights men immediately revolt

A verified stylized pitch movement is called a close-copy stylization in the sense of De Pijper (1983).
The result of the process described above is a representation of Russian intonation in terms of perceptually relevant pitch movements that combine to form complete stylized contours.

3. AIM AND MOTIVES

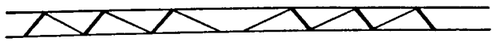
The aim of the present research is to represent larger fragments of Russian spontaneous speech in stylized pitch contours. I was encouraged by the studies of Svetozarova (1982) and Keijsper (1983) to formulate the following issues for perceptual research:
1. The number of discrete pitch movements in Russian;
2. The positions in which pitch movements occur within the contour;
3. The possible combinations of pitch movements;
4. The acceptable tolerance within pitch movement parameters;
5. The linguistic function of perceptually relevant pitch movements.
It is expected that the perceptual approach will contribute to elucidating these issues.

4. SOME ASPECTS

So far, an excerpt of Russian spontaneous speech (2 min.) has been fully analysed and described in terms of stylized pitch contours (Odé, 1986). Apart from some questions I had to face while analysing the excerpt, two problems were encountered in particular:
1. a sawtooth shaped intonation pattern was found to exist;
2. the number of reference lines and the exact distance between them in semitones could in most cases not be easily defined.

4.1 The sawtooth pattern

The intonation pattern which I call the sawtooth because of its shape, consists of a sequence of one or more pitch accents in non-final position. The prominent syllables in the sawteeth in one sequence either rise or fall (the bold line indicates the accented syllable):



This combination of pitch movements occurs frequently in the above mentioned excerpt. The sawtooth pattern precedes the final pitch accent before a boundary. The establishment of the sawtooth pattern is important. It is an explanation for the fact that many pitch accents that can be found in Russian spontaneous speech do not belong to any of the intonation constructions as described by Bryzgunova (1977). Furthermore, the sawtooth pattern may possibly contribute to the solution of an important linguistic problem: what is the reason for the fact that prosodic grouping of words in a Russian phrase does not always coincide with the syntactic structure of that phrase? Probably a prosodic group of words consists of a rising (falling) pitch accent followed by an unaccented falling (rising) part until the next accent.

Pitch accents of the sawtooth type differ from all other pitch accents that were found in the same excerpt. Modifications of sawteeth into another pattern, e.g. a pointed hat or terrace pattern (a frequent pitch accent in Dutch) were not acceptable for native speakers of Russian.
These considerations have brought me to a first careful conclusion that the sawtooth pattern
a) has not yet been described as a Russian intonation pattern;
b) consists of non-main pitch accents in non-final position;
c) groups together words prosodically which do not necessarily belong together syntactically.
Further analysed speech material seems to support these statements.
An explicit phonetic description of the sawtooth pattern can not (yet) be given. There is much varation in excursion and duration within the pattern, which makes the other features, e.g. the position in the phrase, even more important.
The problem of a phonetic description leads to the second aspect that I would like to examine in this paper: the definition of the reference lines between which pitch moves.

4.2 The reference lines
The question about the reference lines in Russian intonation is twofold:
How many reference lines are to be distinguished and, particularly, how can these lines be defined, that is: how many semitones separate the lines from each other?
Assuming that pitch movements can be described as moving between a high and a low reference line, one question arises immediately: how to handle the fact that in spontaneous Russian speech not all pitch movements have the same size? How can this phenomenon be described in terms of reference lines?
We cannot simply divide the field between the high and low line into two equal parts and assume a middle reference line (see Fig.3) if such a line does not separate perceptually relevant whole pitch movements from perceptually relevant non-whole pitch movements. Non-whole pitch movements are not always the half of whole pitch movements.

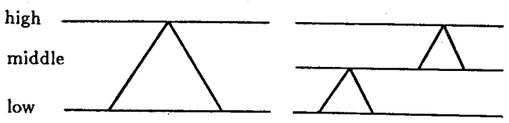


Fig.3: Whole pitch movements and pitch movements that are the exact half of whole pitch movements

If half pitch movements indeed exist and move between low and middle or middle and high reference lines, how can we define the lines on which non-half and non-whole p movements (pitch movements that are not the exact ha whole pitch movements) start or end?
In the analysis of different types of spontaneous Rus speech whole and non-whole pitch movements were fo The non-whole pitch movements were usually not the of whole pitch movements. Rather, instead of one mi reference line, it seems more appropriate to assume a high and a non-low reference line. Movements that do reach the high, non-high, low or non-low reference line at points in the field between the non-high and non-low erence line. No further subdivision of this field seems t possible (see Fig.4).

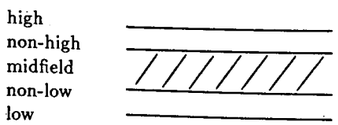


Fig.4: The subdivision between high and low reference li

The assumption of such a subdivision is suggested by variety of frequent non-whole pitch movements of the lowing type (see Fig.5):

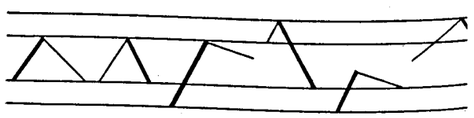


Fig.5: Non-whole pitch movements (the bold line indica the accented syllable): a few examples. The thin l demonstrates the direction of the pitch movement in t unaccented pre- or posttonic parts.

The subdivision indicated in Fig. 4 and Fig. 5 is releva for the perception of pitch movements. This means th this subdivision is also linguistically relevant. The subdi sion is not only relevant for the recognition of pitch mo ments within the accented syllable, but also for percepti relevant distinction in posttonic syllables. In the followi examples the accented syllables are perceptually identical, but the posttonic syllables ensure that the configurations differ from each other perceptually. Where the accented part occurs in isolation, e.g. in final position, this difference is neutralized (see Fig.6).

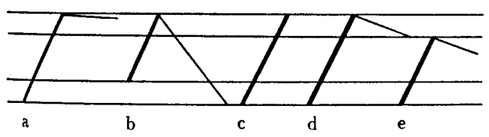


Fig.6: Pitch movements with an identical rise (the bold line indicates the accented syllable) but different posttonic syllables: a) rise + high posttonic part, b) rise + falling posttonic part, c) rise + no posttonic part in final position, d) rise + posttonic part gradually falling to the non-high level, e) rise to the non-high level + posttonic part gradually falling to the midfield.
b) and c) are always clearly different from a), d) and e). A perception experiment has been devised to verify whether a) and d) and d) and e) are indeed discretely different.

5. DISCUSSION

The solution of the two problems described in section 4 contributes to the establishment of an inventory of discrete pitch movements and discrete reference lines between which pitch moves.
If non-main pitch accents must be distinguished from main pitch accents and if the position in which the two types of pitch accents occur can be defined, the issue touches upon the discussion about how prosodic groups and prosodic boundaries can be established. If prosodic boundaries and syntactic boundaries do not coincide, i.e. if syntactic structures do not prescribe prosodic grouping or vice versa, then the issue has important linguistic consequences. It is beyond the scope of my research and this paper to go into detail, but the phenomenon should at least be mentioned here.
The subdivision between high and low reference lines indicated above can solve the problem of the great number of pitch movements that do not fit into a model that distinguishes only between whole and half pitch movements, i.e. a high, middle and low reference line. My description separates from each other pitch movements or configurations of pitch movements that perceptually differ in a very subtle way (see Fig.6, a), d) and e)), but that are probably non-identical.

6. CONCLUSION

The perceptual analysis of Russian intonation has proved to be an appropriate approach to describe pitch movements. At the same time, it stimulates the linguistic discussion of accent and intonation theory. The research also serves a practical purpose. It can be expected that on the basis of a perceptual description of Russian intonation better results will be achieved in the teaching of this often neglected subject. A replacement of the traditional impressionistic way of teaching intonation by an explicit perceptual, phonetic and linguistic description in an audiovisual course with audiovisual feedback will motivate both students and teachers to study intonation.
It is nothing new that a foreigner, even one who speaks the foreign language very well, can usually be recognized by his/her intonation and even be identified as, for instance, a native speaker of Dutch.

7. REFERENCES

Bryzgunova, E.A. *Zvuki i intonacija russkoj reči*, Moskva, 1977.
Collier, R., 't Hart, J., *Cursus Nederlandse Intonatie*, Leuven 1981.
't Hart, J., Cohen, A., "Intonation by rule: a perceptual quest", *Journal of Phonetics* 1, 1973, 309-327.
't Hart, J., Collier, R., "Integrating different levels of intonation analysis", *Journal of Phonetics* 3, 1975, 235-255.
't Hart, J., Nooteboom, S.G., Vogten, L.L.M., Willems, L.F., "Manipulation of speech sounds", *Philips Technical Review* 40, 1982, 134-145.
Keijsper, C.E., "Comparing Dutch and Russian Pitch Contours", *Russian Linguistics* 7, 1983, 101-154.
Nooteboom, S.G., Cohen, A., *Spreken en verstaan*, Assen, 1984.
Odé, C., "Towards a perceptual analysis of Russian intonation", *Studies in Slavic and General Linguistics* vol.8: *Dutch Studies in Russian Linguistics*, Amsterdam, 1986, 395-442.
de Pijper, J.R., *Modelling British English Intonation*, Dordrecht, 1983.
Svetozarova, N.D., *Intonacionnaja sistema russkogo jazyka*, Leningrad, 1982.

PAGE D'HISTOIRE DE LA PHONETIQUE ANCIENNE
LA FORME EXTERNE DE L'ALPHABET ASOMT'AVRULI EN TANT QUE MODELE GRAPHIQUE
DE LA STRUCTURE DIFFERENTIELLE DES PHONEMES DU VIEUX-GEORGIEN

SERGE N. MOURAVIEV

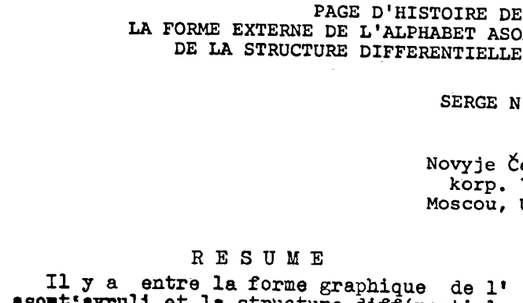
Novyje Čerëmuški 32[a]
korp. 7, kv. 37
Moscou, URSS, 113209

R E S U M E

Il y a entre la forme graphique de l'**asomt'avruli** et la structure différentielle des phonèmes correspondants du vieux-géorgien des équivalences non fortuites qui permettent de reconstituer un véritable code graphique spécialement conçu et utilisé pour bâtir les graphèmes de cet alphabet. Sachant ce code, il est possible de reconstruire la classification phonétique qu'il implique et de juger ainsi de l'intuition phonologique de son/ses auteur(s). Quant aux écarts par rapport à lui qu'on observe dans une partie des graphèmes, ils s'expliquent par les règles graphiques auxquelles toute écriture qui se veut efficace doit forcément se plier.

0. INTRODUCTION

La présente communication s'inscrit dans le cycle de nos travaux sur l'origine des alphabets paléochrétiens du Caucase [1; 2; 3] et contient la substance d'une étude monographique actuellement sous presse [4], où l'on trouvera maint détail supplémentaire et toute la littérature pertinente.

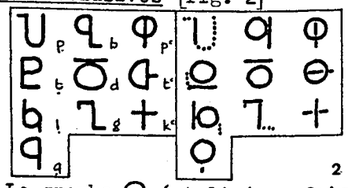


L'**asomt'avruli** est la plus vieille forme attestée de l'alphabet géorgien. Le voici tel qu'il apparaît dans les inscriptions des Ve-VIe (32 graphèmes), VIIe (h č', š ǯ, č h) et VIIIe (ɣ ž) siècles [fig. 1]. De tous les allographes reproduits ne seront utilisés ici que les paléographiquement les plus anciens, ceux qui figurent en tête de chaque graphème.

1. ANALYSE
LES EQUIVALENCES GRAPHO-PHEMIQUES

1.0 On trouvera ci-dessous un inventaire des correspondances plus ou moins régulières qu'on peut observer entre certains **graphes** (éléments de graphème) et certains **phèmes** (traits différentiels de phonèmes) de l'asomt'avruli.

1.1 **Les occlusives** [fig. 2]



a. Le graphe ○ (stylisé parfois en ⊂ ou ⊃) figure dans toutes les rangées et colonnes (7 graphèmes sur 10).

b. Le graphe | figure dans la rangée des labiales (3 sur 3) ; le graphe — dans celle des dentales (3 sur 3) ; le graphe | & — (combinaison des 2 précédents) dans celle des vélaires (2,5 fois sur 3). Pour le graphe ˌ, cf. § 1.2.1.

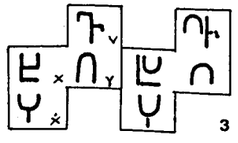
c. Le graphe rectiligne de la colonne des abruptives est placé par rapport à ○ soit à gauche, soit au-dessous ; celui des sonores, soit à droite, soit au-dessus, soit les deux (?) ; celui des aspirées, au centre soit verticalement, soit horizontalement, soit les deux (6 ou 7 sur 10).

1.2 **Les fricatives**

1.2.1 **Les stridentes** [fig. 3]

a. Le graphe ∪ ou ∩ (stylisé aussi en ⊔ ou ⊓) figure dans tous les graphèmes (4 sur 4).

b. La labiale **v** possède (comme les labiales occlusives) un graphe | ; une des deux uvulaires, **x**, possède (comme 2 des vélaires occlusives) le graphe | & — ; enfin, la pharyngale **ẋ** possède un graphe ˌ assimilable au graphe rectiligne de la pharyngale occlusive q [cf. fig 2].



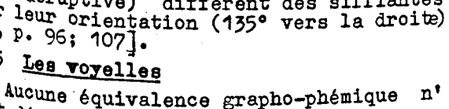
c. Le graphe rectiligne des 2 sourdes **x** et **ẋ** (comme celui des occlusives abruptives) est placé à gauche et/ou au-dessous du graphe ∪ ; le graphe rectiligne de la sonore **ɣ** (comme celui de l'occlusive correspondante **b**) est placé à droite du graphe ∩.

1.2.2 **Les moyennes** [fig. 4]

a. Le graphe ɓ ou ɠ figure dans tous les graphèmes (4 sur 4).

b. Le graphe ɓ figure dans la rangée des sifflantes (2 sur 2), ɠ dans celle des chuintantes (2 sur 2).

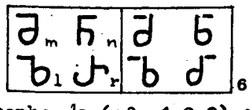
c. Le graphe — figure dans la colonne des sonores (2 sur 2) et le graphe ‾, dans celle des sourdes (1 sur 2).



1.2.3 **Les douces** [fig. 5]

Il n'y en a que deux, **w** étant considéré comme une voyelle (prononcée [ü/wi] ; cf. [3, p. 64]). Chaque graphème se compose d'un graphe incurvé ⊃ ou ⊔ et d'un graphe à crochet rectiligne ⌙ ou ⌐.

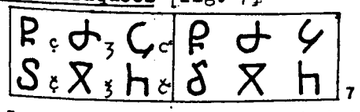
1.3 **Les liquides** [fig. 6]



a. Le graphe ɓ (cf. 1.2.2) ou ɗ, et le graphe ‾ (cf. 1.1, 1.2.2) figurent apparemment dans tous les graphèmes.

b. Dans les nasales, le second graphe et la boucle du **premier** sont orientés dans le même sens; dans les buccales, dans des sens opposés.

1.4 **Les affriquées** [fig. 7]



a. La sonore et l'aspirée chuintantes ont la même configuration que les sifflantes correspondantes, mais sont faites de lignes droites/brisées.

b. Les 2 abruptives sont faites chacune des mêmes graphes ρ et ¬ connectés différemment.

c. Les chuintantes (sauf le graphe ¬ de l'abruptive) diffèrent des sifflantes par leur orientation (135° vers la droite) [1, p. 96; 107].

1.5 **Les voyelles**

Aucune équivalence grapho-phémique n'est décelable [4, § 2.2.5].

2. SYNTHESE
LES CODES PHONO-GRAPHIQUES

2.0 Les correspondances observées ci-dessus impliquent l'existence d'un système complet de corrélations **phèmes : graphes.** En voici une reconstitution semi-hypothétique succinte d'où il découle qu'il y avait deux codes différents bien que liés entre eux : un pour les non-affriquées et un autre pour les affriquées.

2.1 **Les non-affriquées**

2.1.1 **Le code.** Chaque graphème se compose d'un élément de base (EB) incurvé et d'un élément secondaire (ES) rectiligne. L'EB désigne le degré de constriction ; l'ES, l'organe actif de phonation; l'orientation (tropisme) de l'EB,parfois, l'organe passif de phonation; la position réciproque (syntaxe) des 2 éléments, le mode de voisement.

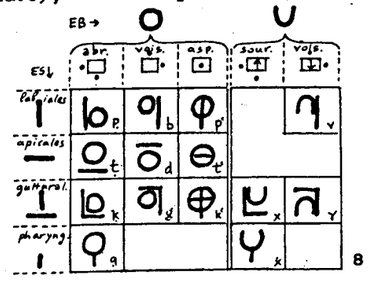
Les 4 EB sont :
○ = constriction totale (occlusivité)
∪ = constriction partielle (friction) forte
ɓ = constriction partielle (friction /liquidité) moyenne
⊍ = constriction partielle (friction) faible.

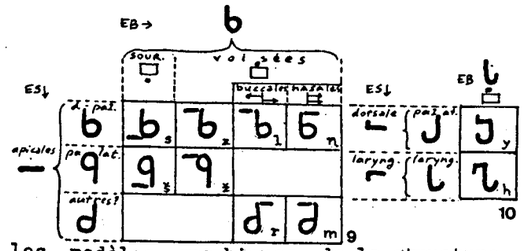
Les 6 ES sont :
| = labialité | & — = gutturalité
— = apicalité ˌ = pharyngalité
⌙ = dorsalité ⌐ = laryngalité.

Les tropismes des EB sont :
∪ = ∩ (sans signification propre)
ɓ = dentopalatalité, ɠ = palatalité, ɗ = autre localisation (?)
ʃ = palatalité, ⊍ = laryngalité.

La syntaxe obéit à deux règles :
A. ES à gauche ou/et au-dessous de EB = = abruptivité/non-voisement
ES à droite ou/et au-dessus de EB = = voisement
ES traversant EB verticalement ou/et horizontalement = aspiration
B. (valable seulement pour ɓ/ɗ)
EB et ES orientés dans le même sens = = nasalité ; en sens inverse = buccalité.

2.1.2 **Les modèles ou formes de base.** Tel que décrit et compte encore tenu de règles calligraphiques (omises ici faute de place), ce code permet de construire



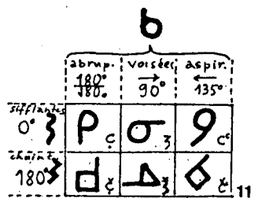


9 10

les modèles graphiques de la structure phémique des 24 consonnes non-affriquées du vieux-géorgien [fig. 8 à 10].

Deux modèles de liquides sont, toutefois, imparfaits : **m** n'est pas une apicale et **l** se prononce autrement que **s**. Cf. 2.3, 3.3.

2.2 **Les affriquées** [fig. 11]



11

A la différence du code purement combinatoire des non-affriquées (2.1.1), celui des affriquées est plutôt algorithmique. Il fait subir à un élément initial (EI) **b** emprunté aux fricatives sifflantes (= affrication) un système de 3 transformations symétriques générant d'abord les sifflantes, puis les chuintantes :

a. Trois tropismes différents, dont un réflexif, (= abruptivité, voisement, aspiration) tranforment l'EI en les 3 modèles des sifflantes ꝑ, ꝛ et ꝯ.

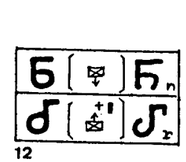
b et **c.** Leur rotation de 180° et leur rectilinéarisation transforme les 3 modèles des sifflantes en les 3 modèles des chuintantes ꝺ, ꝟ et ꝸ.

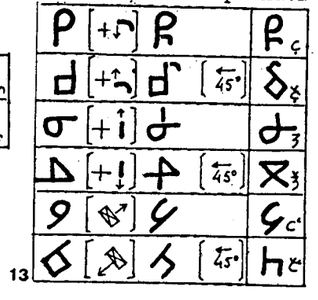
Cf. encore 2.3.

2.3 **Altérations subséquentes**

L'imperfection phono-graphique de certains liquides (2.1.2), de même que, sans doute, des particularités de leur phonation et de celle des affriquées sont la raison pour laquelle la plupart des formes de base (FB) de ces 2 groupes de graphèmes ont subi des altérations subséquentes dont la signification phonétique exacte nous échappe, mais qui n'ont apparemment pas été dictés par des considérations purement graphiques (§ 3).

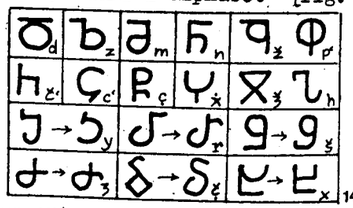
A noter que les deux liquides [fig. 12] et les six affriquées [fig. 13] sont altérés de façons analogues et que les altérations de ces dernières semblent faire double-emploi avec le code.


12


13

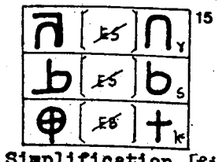
3. INTERPOLATION
LES AJUSTEMENTS GRAPHIQUES

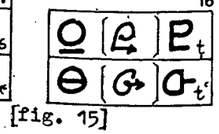
3.0 Dix-huit des trente FB ainsi construites ont été incluses telles quelles (ou presque) dans l'alphabet [fig. 14].


14

Les douze autres présentaient certains inconvénients pour l'écriture : les unes étaient trop difficiles à tracer d'un trait, d'autres se déformaient au point de devenir méconnaissables à la moindre cursivation, d'autres encore se ressemblaient trop entre elles, surtout en cursive.

D'où la nécessité : de simplifier les premières, de stabiliser les secondes et de différencier les troisièmes.


15


16

3.1 **Simplification** [fig. 15]

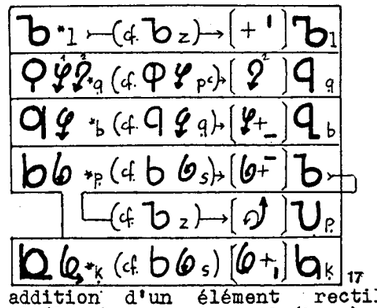
Trois FB compliquées (ɣ, **s** et **k'**) sont simplifiées par amputation d'un élément : soit de l'ES (ɣ, **s**), soit de l'EB (**k'**). Cf. encore le cas de **g** (3.4 ; fig. 18).

3.2 **Stabilisation** [fig. 16]

Deux FB instables (ꝫ et **t'**) sont rendues plus constantes par cursivation et capitalisation consécutive de leur allographe cursif. Cf. encore les cas de **q**, **b**, **p**, **k'** (3.3 ; fig. 17) et **v** (3.4 ; fig. 18).

3.3 **Différenciation** [fig. 17]

Les FB de **l**, **q**, **b**, **p**, **k**, ou leurs allographes cursifs, identiques ou semblables respectivement à **z**, **p'**, **q**, **s** (et **z**), **s**, ou à leurs allographes cursifs, en sont différenciées au moyen d'une ou deux des opérations suivantes :


17

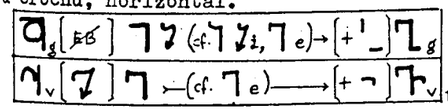
a. addition d'un élément rectiligne vertical (**l**, **k**) ou horizontal (**b**, **p**) ;

b. cursivation + capitalisation (**q**, **b**, **p**, **k**), cf. 3.2 ;

c. déformation d'un élément (**p**).

3.4 **Cas particuliers** [fig. 18]

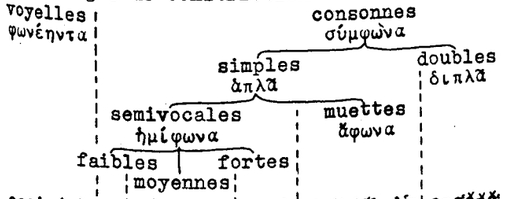
La FB de **g**, après simplification (3.1), est devenue semblable à **i** ; l'addition d'un élément vertical la rend semblable à **e**, d'où addition d'un nouvel élément, horizontal lui. La FB de **v**, après stabilisation (3.2), est aussi devenue semblable à **e**, d'où addition d'un élément, rectiligne ou crochu, horizontal.


18

4. CONCLUSION
UNE CLASSIFICATION PHONOLOGIQUE DU V^e S.

Ce système graphique sophistiqué (de même que l'ordre alphabétique [cf. 3, p. 78-9]) suppose l'existence, au V^e s. au plus tard, d'une classification détaillée des phonèmes du vieux-géorgien tenant compte de leur structure phémique. Quatre critères semblent avoir été utilisés : le degré de constriction (de nulle pour les voyelles à double pour les affriquées), les modes d'aspiration/voisement, les types de résonance et les lieux d'articulation.

a. Degré de constriction



b. Modes d'aspiration (πνεῦμα)/voisement

minces (ψιλά) :					s	š	x	x̌	p	ṭ	k	q	c	č
médianes (μέσα) :	y	h?	z	ž	v	ɣ	w	d	g				ʒ	ǯ
épaisses (δασέα) :									p̣	ṭ	ḳ		c̣	č̣

c. Types de résonance

nasales : **m** **n**, buccales : les autres

d. Lieux d'articulation

active - labiales : **m**? **v** **p** **b** **p'** ; apicales : **n** **l** **r** **s** **z** **š** **ž** **ṭ** **d** **t'**+ affriquées ?; dorsales : **y**; gutturales : **x** **ɣ** **k** **g** **k'** ; pharyngales : **ẋ** **q** ; laryngales : **h** ; passive - dentopalatales : **n** **l** **s** **z**, palatales : **y** **š** **ž** (à cette dernière division correspond la distinction plus empirique entre les affriquées sifflantes et chuintantes)

Cette analyse phonologique est remarquable par la subtilité de certaines distinctions (p. ex. entre les stridentes, les moyennes et les douces). Elle ne l'est pas moins aussi par les distinctions qu'elle ignore : entre fricatives et liquides, vélaires et uvulaires ; par la façon dont elle s'embrouille dans les liquides ; dont elle élude le difficile problème de la bifocalité des chuintantes ; enfin, par son impuissance à classer les voyelles.

Tous ces traits découlent évidemment de l'appareil théorique utilisé qui est, bien sûr, celui des grammairiens grecs. Pour ceux-ci, en effet, les semivocales ne comprenaient que les liquides et une fricative, **s** (**z** était "double", **h** un "souffle", **y** se confondait avec **i**) ; il n'y avait pas d'uvulaires ; la classification des liquides faisait problème. Celle des voyelles grecques est inapplicable au géorgien qui n'a ni voyelles longues, ni diphtongues.

Il est d'autant plus stupéfiant de constater que le ou les auteurs de l'alphabet géorgien (car cette classification leur appartient sans aucun doute) ont su non seulement appliquer au vieux-géorgien avec ses 36 phonèmes un appareil théorique conçu à partir des 24 lettres grecques, mais encore le développer de manière à rendre compte des 9 + 6 fricatives et affriquées spécifiques du vieux-géorgien et à identifier de nouveaux lieux articulatoires tels que le pharynx et le larynx.

Sans parler de l'idée révolutionnaire de représenter chaque trait distinctif par un graphe et de simuler la structure de chaque phonème en construisant la lettre qui le désigne à partir des graphes appropriés !

[1] **S.N.Mouraviev**, "Les caractères daniéliens. Les caractères mesropiens", Rev.des Etudes arméniennes XIV (1980) 55-111.

[2] **С.Н.Муравьев**, "Три этюда о кавказско-албанской письменности", Ежегодник иберийско-кавказского языкознания VIII (1981) 222-325 ; Cf. Le Muséon 93 (1980) 345-374.

[3] **S.N.Mouraviev**, "Valeurs phoniques et ordre alphabétique en vx-géorgien", Zeitschrift der Deutschen Morgenländischen Gesellschaft 134 (1984) 61-83.

[4] **С.Н.Муравьев**, "Генезис древнегрузинского письма асомтаврули" (sous presse).

LAUT UND NAME MITTELALTERLICHER BUCHSTABEN*

Horst Weinstock

Institut für Anglistik
Rheinisch-Westfälische Technische Hochschule
D 5100 Aachen, BR Deutschland

Der Vortrag verfolgt die Geschichte von Name (nomen) und Lautwert (potestas) der Buchstaben von der analytischen Silbenschrift bis zu Aelfrics Grammatik. Die konsonantischen Akrophone des hebräischen Mikroalphabets bezeichnen Lautbildungsstellen; sie ermöglichen auch polyglotte Transkription und Transliteration. Varro unterteilt die 'Mitlaute' in Konsonanten (K̄V̄) und Halbvokale (V̆K). Die Routineverschriftung des Vokalismus trennt die alphabetsystematischen Namen der Buchstaben von ihren am historisch-morphophonologischen Wandel beteiligten Lauten.

* Ein Abdruck der vollständigen und dokumentierten Fassung erscheint in Linguistik in Deutschland: Akten des 21. Linguistischen Kolloquiums in Groningen, September 1986, Tübingen, 1987, 405-421.

Quellenbefund und Forschungsstand erlauben einen raschen Einstieg und liefern ein klares Bild. Aus semantischen Zeichenschriften mit (zunächst) konkretem Bildsinn und (alsbald) abstraktem Begriffssinn entwickelte sich allmählich eine analytisch-monomorphemographische Schreibweise. Der gleichen Lauteinheit entsprach systemgebunden das gleiche Schriftzeichen. Die damals kleinsten Schrifteinheiten umfaßten Silben.

Im frühen 2. Jahrtausend v. schafften ursemitische Schreibmeister den Riesenschritt von analytischer Silbenschrift zu einem rein konsonantischen Buchstabenalphabet. Alle zweiundzwanzig Buchstaben tragen einen Namen mit lexikalischer Eigenbedeutung und Zahlenwert (Begleittext Nr. 1). Die Akrophone des Mikroalphabets vertreten die hebräischen Buchstabennamen des Makroalphabets als Laut-, Wort- und Zahlenwerte. Die Phonographe des Alphabets können auch nicht-hebräische Sprachen transkribieren und transliterieren. Das echte Alphabet schließt als Einfachsystem allographische Mehrfachbezeichnungen aus. Jeder Buchstabe bezeichnet einen einzigen Lautwert oder eine bestimmte Lautbildungsstelle; jedem Lautwert oder jeder bestimmten Lautbildungsstelle entspricht ein Einzelbuchstabe.

Ob einst durch Zufall oder aus Absicht konsonantisch, haben menschliche Willensfreiheit und sprachgeschichtlicher Sachverstand die Vokallosigkeit nachträglich zweifellos aus guten Gründen sanktioniert. Bereits Genesis 2.19-20 schildert das Sprach- und Sprechvermögen als Gottesschöpfung, den Sprachbau und die Sprachverwirklichung aber als Menschenwerk. Entsprechend der menschlichen Doppelteilhabe tragen Mensch und Menschenwerk die Züge von Körperlichem einerseits und von Geist und Seele andererseits. Das Laut- und Schriftgerüst der konsonantischen Grundlaute oder Radikale verkörpert die primäre, lexikalische, statische oder Speicherform eines Wortes, die natürliche, traditionsverpflichtete, bleibende und wiederkehrende Wortbedeutung, den potentiellen Rahmen oder die Kompetenz, das sakrosankte zeitliche Vermächtnis aus dem überzeitlich Ewigen. Dem Wesen nach archiphonemisch und ohne phonematische Gegensätze von Verschluß- oder Reibelaut, von Länge oder Kürze, hüten die allophon konsonantischen Grundlaute als Diener, Wächter, Stallwachen oder Statthalter das feste und beständige Gefüge der Sprechwerkzeuge und ihrer Artikulationsstellen. Die unveränderliche konsonantische Triliteralität jedes Wortstammes bestimmt seine Lexikalbedeutung. Den grammatischen Kontextualsinn der Worte im Satz konstruieren die morphemophon funktionalen Vokale. Der verstandesbeherrschte, für alle möglichen Sprecher konventionalisierte Geist segmentaler Vokalisierung und die gefühlsgetragene, von den einzelnen Sprechern aktualisierte Seele der suprasegmentalen Betonung oder Kantillation schöpfen die sprachliche Ausdruckskraft voll aus. Die von Vokalen tönende Stimme verwirklicht, belebt, beseelt, durchgeistigt und vollendet die menschliche Rede. Erst die (gesprochen) eindeutige Vokalisierung prägt den (geschrieben) uneindeutigen Konsonantismus. Vokale artikulieren den vergegenwärtigten Kontextualsinn und koartikulieren den überzeitlichen Pausal- oder Lexikalsinn. Der Verzicht auf vokalisierte Schreibweise sichert dem Hebräischen stets aufs neue schöpferische Sprachimpulse aus dem Zwang der Traditionsverpflichtung gegenüber der Vergangenheit und der Freiheit zum Überdenken in der Gegenwart. Textverständnis fordert einen Sprachvollzug aus der Geisteshöhe des Eindenkens und der Seelentiefe des Einfühlens.

Seit dem 10. oder 9. Jahrhundert v. drang das (nord)semitische Alphabet über die Phöniker zu den Griechen. Übereinstimmung oder

Ähnlichkeit in Buchstabengestalt, Buchstabenfolge, Buchstabennamen und (vorerst linksläufiger) Schreibrichtung schließen eine beweiskräftige Kette (Begleittext Nr. 2).

Daß trotz der Übernahme der äußeren Form die Anverwandlung der inneren ausblieb, bekunden zwei Tatsachen: Erstens neigen die griechischen Buchstabennamen zur Standardisierung des langen betonten Endvokals -ī̄ oder -ȳ̄ für Einsilber und des kurzen und unbetonten Endvokals -ă̄ für Zweisilber. Zweitens tragen die griechischen Buchstabennamen keine Eigenbedeutung mehr. Eine gewisse Sinndeutung erhellt allenfalls noch aus den Zusatzbuchstaben epsilon, omikron, ypsilon, omega mittels der Merkmalsangaben mikron 'klein, kurz', mega 'groß, lang' und psilon 'einfach, d.h. monophthongisch'.

Die Griechen ergänzten das konsonantische Alphabet des Hebräischen durch Āleph, Hē̆, Wâv, Jôd, Ayin um deren akrophone Lautwerte a e u i o. Die routinemäßig verschriftete Vokalisierung kostete im Grunde einen hohen Preis. Die lineare Festschreibung der Vokalbuchstaben zwängte die akustisch modulationsfähige Stimmtönung in starre Gewohnheitsparadigmen. Leser/Sprecher und Hörer/Schreiber vollziehen fortan den segmentalen und suprasegmentalen Konstruktionssinn über die normative Optik des Vokalalphabets. In anderen Worten: die Vokalschreibung aus Gewohnheit hatte Aussprache von Schreibung, potestas von nomen und figura, Transkription von Transliteration getrennt.

Nachplatonische griechische Sprachphilosophen achteten vornehmlich auf die Identifikation des Gemeinten. Das verlagerte den Blick von hör- oder sichtbaren äußeren Sprachformen auf die inneren des Begrifflichen. Immerhin verdankt die historische Phonologie griechischen Einsichten in sprachliche Erscheinungsbilder die dreifache Unterscheidung zwischen Buchstabenname (onoma), seinem phonemographischen Wert (character/schema) und seiner phonetisch-allophonischen Lauttönung (dynamis/ekphonesis).

Die Rudimente römisch-alphabetischer Schriftniederschläge im 7. Jahrhundert v. entstammen kümmerlichen Anfängen, in die sie in Abständen wieder zurückversinken. Immerhin festigte sich schon bald die Schreibrichtung von links nach rechts.

Die dreiundzwanzig Buchstaben des lateinischen Alphabets vereinten semitisch-griechisch-etruskische Schriftvorbilder und -gepflogenheiten. Alle etruskisch-römischen Buchstabennamen bildeten einheitlich Einsilber (Begleittext Nr. 3). Die meisten Konsonantennamen lauteten auf kurzes ĕ- an oder auf langes -ē̄ aus. Wie griechisch pī̄ oder phī̄ und lateinisch ĕf oder pē̄ lehren, begannen nachhebräische Alphabete Laute statt Lautbildungsstellen zu bezeichnen.

Die lateinischen Buchstabennamen blieben Fremdkörper ohne eigene Wortbedeutung und Zahlenwert, unveränderlich auch nach Fall und Zahl. Mehr als je zuvor in der Geschichte analytisch-phonetischer Schreibweise deckten sich Buchstabenname und Lautwert. Die Akrophone des Mikroalphabets klangen sogar mit ihren Buchstabennamen im Makroalphabet völlig gleich. Außerhalb des Gewohnheitssystems und der Bedeutungsfunktion von Eigennamen gehorchten die Vokalanteile der Buchstaben sogar dem regelmäßigen diachronischen Lautwandel, vorerst in der Lautgestalt dominanter Allophone unter dem Hauptton in offenen Silben und alsbald phonemisiert. Die Eignung analytisch-alphabetischer Schreibweise für Aussprache-schreibung und Lauttranskription schwand dahin.

In der Blütezeit römischer Grammatik zeichnete Varro (116-27 v.) für einen auf Jahrhunderte folgenschweren Rückschritt verantwortlich. Vermutlich nach griechischen Quantitätsunterschieden oder nach dem etruskischen Vorbild der silbenfähigen Dauerlaute f l m n r s und des griechischen xi zum lateinischen Spiegellaut ix sonderte Varro Halbvokale von den übrigen Konsonanten aus (Begleittext Nr. 4).

> "[...] consonantes ab e debere incipere, quae semivocales sunt, et in e debere desinere, quae mutae sunt."

Varros Riesenschritt in die falsche Richtung hielt zwei phonologisch gleichartige Buchstabengruppen für verschiedenartig. Dennoch bürgerte sich Varros Nomenklatur rasch ein und lebt bis heute fort (Begleittext Nr. 5).

bē̄ cē̄ dē̄ ĕf gē̄ ĕl ĕm ĕn pē̄ ĕr ĕs tē̄ ĕx/ĭx.

Grammatiker des 2., 3. und frühen 4. Jahrhunderts schrieben ihre Vorbilder (nicht selten wörtlich) aus. Grammatiker des 4. bis 6. Jahrhunderts erwarben sich große Verdienste, indem sie die römische Grammatik und ihr Kapitel "De lit(t)eris" überlieferten und verbreiteten. Das rund tausendjährige Lateinische Mittelalter hindurch gelangte keine Grammatik so nachhaltig zu Anerkennung, Umlauf, Ansehen, Tradition und Bewunderung wie der 'Donat'.

Um die Mitte des 4. Jahrhunderts kanonisierte der spätrömische Grammatiker Donatus, der Lehrer des Hieronymus, einerseits die häufigste Abfolge der Lehrabschnitte vox - lit(t)era - syllaba - dictio und andererseits gewisse Grundansichten über 'Buchstaben'. In der Lautlehre (vox, littera) wiederum bildete die Kunst lauten Vorlesens (lectio) eine vordringliche Aufgabe für die Beherrschung der grammatischen Tugenden von scribere, legere, intellegere, probare.

Donatus und seine frühmittelalterlichen Nachfolger grenzten die geistig verständliche und in Buchstaben verschriftbare Stimme des vernunftbegabten Menschen (vox articulata) ab gegen das Unverständliche an Schall und Geräuschen der Tiere und Dinge (vox inarticulata sive confusa). Trotz des bereits fortschrei-

tenden Problembewußtseins für das Neben- oder Gegeneinander zwischen gesprochener Minimaleinheit (elementum) und geschriebener (littera) reichte Donatus die mangelnde Begriffstrennung weiter und gebrauchte littera zur phonischen sowie zur graphischen Beschreibung. So schlich sich die begriffliche Unschärfe auch in seine verdienstvolle Wiedererweckung der Lehre von den drei Eigenschaften der Buchstabenlaute gegenüber den Schriftbuchstaben ein (Begleittext Nr. 6). Praktisch nämlich bezeichnet nomen weiterhin den gesprochenen Buchstabennamen, figura die geschriebene Buchstabengestalt und potestas seinen Lautwert samt der allophonischen Varianten.

Varros Unterscheidung der Buchstaben in Vokale (V̄: a e i o u), Konsonanten (K̄V̄: b c d g [h] k p q t und Halbvokale (V̄K̄: [f] l m n r [s x]) erfuhr durch Donatus ihre Begriffsbestimmung auf Jahrhunderte (Begleittext Nr. 7). Wie Varro billigte auch Donatus allein dem Typ Konsonant-plus-Langvokal (K̄V̄) Silbenstatus zu und systematisierte zugleich den stets gleichen prothetischen Kurzvokal der sieben Halbvokale anstatt ihres eigentlichen Konsonanten (naturalis sonus).

Macht der Gewohnheit verwehrte dem Systembruch im unphonetischen Alphabet die Heilung. Routine und unkritische Grammatiktradition störten das Wesen einer analytisch-phonetischen Schreibweise, sprengten die akrophone Entsprechung zwischen Lautwert und Buchstabennamen sowie die Möglichkeit zur Verfahrensgleichheit bei polyglotter Transkription und lateinischer Transliteration. Donatus schob das nicht aufgehobene Problem künftig unentbehrlicher Lautumschrift auf, reichte es an seine mittelalterlichen Nachfolger weiter - ungelöst, aber immer gewohnheitsvertrauter und unverdächtiger.

Erst Priscian um die Wende zum zweiten Viertel des 6. Jahrhunderts bemerkte einige verbesserungsfähige Begriffsbestimmungen. Er erblickte das Wesen der Halbvokale im auslautenden Konsonanten selbst (Begleittext Nr. 8). Über die theoretische Einsicht hinaus, hielt Priscian allerdings praktisch am eingespielten Systembruch fest.

Scharf hingegen hob Priscian das Wesen mikroalphabetisch analytisch-phonetischer Schreibweise durch Absicherung des am Buchstaben Unteilbaren (individuum) und des durch Stimme Verbindenden (compositum) hervor sowie die Trennung von geschriebenem Buchstaben (lit(t)era) und gesprochenem (elementum). Priscian brach mit Varros Deutung der figura als Lautgestalt und Teilstrecke der vox.

Wie kraftvoll Geist und Seele des Sprechenden segmental und suprasegmental in den (eben vokalischen) Sinn- und Starktongipfeln ertönen, veranschaulicht sein Leib-Seele-Vergleich zwischen Konsonanten und Vokalen (Begleittext Nr. 9). Priscian verstand die kleinsten Laut- und Schrifteinheiten als Bindeglieder des Redeflusses, die beim stimmlichen Abruf dynamisch und koartikulatorisch über sich selbst hinauswachsen. Mit Diomedes leitete er litera aus legitera für legendi iter 'lauter Leseweg' her. Das erhob potestas vom ungesättigten Lautwert zum gesättigten Füge- oder Ausdruckswert.

Isidor von Sevilla (560?-636) schärfte das Grammatikbewußtsein für analytisch-phonetische Schreibweise nur unerheblich. Doch besann sich der Spanier auf die einstige Herkunft der griechischen und lateinischen Buchstaben von den hebräischen.

Auf der Insel erfragte Tatwine (?-734) in seinem Rätsel "Versus de nominibus litterarum" zweifellos einsilbige Buchstabennamen, das heißt lateinische. Er versteht littera als figura 'Schriftgestalt' der Vokale bzw. der konsonantischen soni naturales, das heißt der potestates.

Ein weiteres Rätsel hebt an (Begleittext Nr. 10):

"Innumerae sumus, et simul omnes quaeque sonamus. / Vna loqui nequit."

Unzählbar gelten die 'gesprochenen' Buchstaben wegen der Biphonie von ⟨i⟩ /i,j/ und der Triphonie von ⟨u⟩ /u,v,w/. Zusammen ertönen sie als alphabetisches Pausalsystem und beim Kontextualgebrauch in Silbe, Wort und Satz. Ein Buchstabe allein redet nicht, weder mit Eigenbedeutung des Buchstabennamens (nomen) noch des akrophonen Lautwerts (potestas). Stets eignet dem Buchstaben der Redefluß der Stimme (vox).

Beda Venerabilis (673-735) berief sich wiederholt auf Varro und die lateinischen Kirchenväter. Zum Vergleich zog er ausschließlich das Griechische heran. Im Rahmen einer Orthographia wohlberechtigt, hafteten Bedas Blicke am Schriftbild. Daß Beda in einer Zeit fortschreitend uneinheitlicher Einzelsprachen die Pflege einheitlicher Buchstabenschreibweise(n) regeln wollte, überrascht wenig. Mehr überrascht, daß seine Auseinandersetzung mit dem geschriebenen Buchstaben (figura) zwar noch dessen gesprochene Seite erwähnte, aber sich jeder lautlichen Aussage enthielt. Das verrät 'beiläufig', wie die Unterscheidung zwischen einsilbigem Buchstabennamen (nomen) und fast deckungsgleichem Lautwert (potestas) achtlos in Vergessenheit geriet.

Mit Bonifatius (675?-754) verbindet die Überlieferung die festländische Verbreitung einer Art 'insularer' Geheimschrift. Ohne Nennung der Zeichennamen verlagerten die verschiedenen Punktzahlen und ihre Anordnung den Blick vom Verhältnis zwischen Buchstabenname (nomen) und Lautwert (potestas) vorübergehend auf Schriftbild oder Buchstabengestalt (figura) und Lautwert (potestas). Die Geheimschreibweise ersetzte Vokalbuchstaben entweder durch Punktation oder durch den jeweils alphabetisch nächsten Folgekonsonanten (Begleittext Nr. 11):

(1) a = : e = ⁝ i = · o = :: u = ⁘
(2) a = b e = f i = k o = p u = x
Bonifacius (1) B::N·F:C·⁘:S
(2) Bpnkfbckxs.

Ob in mißdeuteter Nachahmung oder aus zufälliger Erfindungsähnlichkeit, fällt an beiden (auch festländischen) Vokalschreibweisen eine Wesensnähe der Geheimschrift zu hebräischen Schreibgepflogenheiten seit dem 6. Jahrhundert n. auf. Die Punktatoren drückten Vokale durch ein bis fünf unter- oder übergesetzte Punkte aus (Begleittext Nr. 12).

Die von Palatal- nach den Velarvokalen hin angeordnete Punktation bezeichnet segmentalsyntagmatisch (also rein phonisch) mittels Vorkonsonanz; der Ersatzkonsonant bezeichnet alphabetisch (und rein graphisch) mittels Nachkonsonanz. Im Gegensatz zur graphischen Art der Geheimschrift wahrt die analytisch-phonische Punktschreibweise deren Lautübergänge (Transienz). Das erleichtert sinngemäßes Lesen, fordert aber (wie schon das Hebräische) vertieftes Textverständnis durch kritische Neubesinnung auf den Nachvollzug sach- und lautgerechter Lesung (lectio).

Alkuin (735?-804) betrachtete Sprache vor allem in seiner Grammatica und Orthographia. Er verstand den Buchstaben als kleinste Teileinheit einer ganzheitlichen lectio. Dem prudens lector oblag es, Geschriebenes lesend und sprechend geistvoll ins Leben zu rufen, auf Wohlklang (euphonia) und Trennschärfe (differentia/distinctio) zu achten (Begleittext Nr. 13).

Alkuin versetzte sich tief in die hebräische Tradition der mittlerweile mit ihren Lautwerten praktisch deckungsgleichen lateinischen Buchstabennamen. Nach Ausweis von De Rhetorica et de Virtutibus vernahm Alkuins phonetisch-etymologisch-semantische Exegese den totgeschriebenen Vokalsinn aus dem Geist und Mund des prudens lector. Wie nur wenige Gelehrte nach Cassiodor, erkannte und betonte Alkuin in Ars Grammatica und De Litteris Colendis den hohen Wert der Buchstabenpflege. Ungenaues Sprechen, Ablesen, Schreiben oder Abschreiben von Buchstaben verfälschen den Schriftsinn.

Rund dreihundert Jahre später lehrte Aelfric (955?-1020) noch einmal aus Donatus den Abschnitt "De Littera" unter lateinischer Überschrift in nunmehr heimisch-spätwestsächsischer Sprache. Wie Donatus nannte Aelfric den Buchstaben se lǽsta dǽl (pars minima) und mit Priscian untōdǽledlīce (id quod dividi non potest). Wo aber Priscian die Unteilbarkeit und das Verhältnis zwischen phonisch/phonemischen und graphisch/graphemischen Einheiten besonders hervorhob, fehlt es Aelfrics Ausdrucksweise an Unmißverständlichkeit (Begleittext Nr. 14).

Aelfric, der von Buch über Satz, Wort, Silbe zu Buchstabe segmentiert, meint mit Entzweigehen im Sinne der Anecdota Helvetica weder Diphthong noch Digraph. Er sieht den Identitäts- und Identifizierungsverlust beschädigter Buchstaben, die Wesenszerstörung, ohne die sich die Schriftgestalt (figura) eines Buchstaben nicht ablesen oder erkennen läßt. Denn im übrigen hielt Aelfric an den drei Buchstabeneigenschaften des Donatus so eng fest, daß er weiterhin die Schriftgestalt des Buchstaben nicht gegen seine Lautgestalt abgrenzte (Begleittext Nr. 15). Getreu seiner Quelle, bezeichnete Aelfric die Vokale als tönend (clypiendlīce), die Konsonanten als mitlautend (samod swēgende) und die Halbvokale als halbtönend (healfclypiende). Typologisch hallte die Varronische Trennung in Halbvokale des Typs Kurzvokal-plus-Konsonant und in Konsonanten des Typs Konsonant-plus-Langvokal wider.

An je einer Stelle seiner lateinischen "Praefatio" und seines altenglischen "De littera" zögerte Aelfric, lateinische Begriffsbestimmungen der Grammatik auf angelsächsische Spracherscheinungen zu übertragen (Begleittexte Nr. 16 und 17). Noch immer scheiterte das theoretisch-begriffliche Unterscheidungsvermögen an der praktischen Umsetzbarkeit. Weder Hrabanus Maurus noch Otfrid von Weißenburg noch auch der Angelsachse Aelfric vermochten die analytisch-alphabetische Schreibweise des Lateinischen Mittelalters auf germanische Lauttönung zu übertragen oder das phonemographemisch eigenständige Strukturprinzip zu beschreiben. Vor dem Vergleichsmaßstab lateinischer Literate zählten Transliterate aus dem Hebräischen, Griechischen, Keltischen und Germanischen als illiterat. Anders als dem ostkirchlich Griechisch-Kyrillischen haftete dem Hebräischen, Ogham-Keltischen und Runen-Germanischen der Verdacht unerwünschten Kultes an.

Der weitere Verlauf zeichnet sich bereits klar ab. Erst Roger Bacon schärft das Bewußtsein und leistet neue Denkanstöße für die sprachliche Betrachtung der analytisch-alphabetischen Trias aus Buchstabenname (nomen), Lautwert (potestas) und Schriftgestalt (figura).

PHONETIC TRANSCRIPTION AND THE VOLNEY PRIZE

ALAN KEMP

Dept. of Linguistics
University of Edinburgh,
Edinburgh, Scotland.

ABSTRACT

The institution of the Volney Prize in the early nineteenth century was intended to concentrate attention on the need for a standard system to transcribe and transliterate languages. This paper briefly discusses two of the essays submitted, and assesses how far the Volney Prize succeeded in its objectives.

1 THE BACKGROUND

This paper is concerned with the institution of the Volney Prize - an episode in the history of the development of transcription. The term 'transcription' is taken to include both (a) the recording of the phonological and/or morphological elements of a language using a specific writing system (referred to in this article as 'transcription' in a narrower sense); and (b) the recording of the graphic symbols of one writing system in terms of the corresponding graphic symbols of a second writing system (referred to henceforth as 'transliteration').

Prior to the 19th century there was no standard scheme for transcribing or transliterating languages, although a number of systems had been put forward, often with newly devised notations, using a non-roman script (e.g. Wilkins [6], De Brosses [1]). The most practical scheme then existing was that of Sir William Jones ([3]) - intended particularly for converting oriental scripts to the roman alphabet.

Constantin François Volney (1757-1820) was a prominent member of the group of savants in France known as the Idéologues. It became one of his great aims in life to make oriental literature and culture more accessible to the West, and to open up the East to the influence of the superior (as he saw it) western civilisation. With this in view he published two works in the course of his life which were intended to provide a system for transcribing (in the wider sense) the oriental languages, notably Arabic (Volney [4,5]), but he was conscious that they only went part of the way towards achieving this objective. He was aware of Sir William Jones' scheme, and believed it to be an important contribution to the advancement of transcription systems. When Volney died, in 1820, he left approximately 24,000 francs in his will "for the best work related to the philosophical study of languages" with the wish that it should "encourage all work promising to continue and bring to completion a method of transcribing Asiatic languages into European letters".

2 THE VOLNEY PRIZE

A Commission was set up to administer the Prize, consisting of members of the French Academies, and it began by asking for essays which would prepare the ground for a solution. This meant (a) setting out by what means Volney's wishes could be fulfilled; (b) determining the scope of the new system; (c) mapping out a plan of action to be followed; (d) specifying what a successful outcome might lead to.

Of the four essays submitted for this first competition of 1822 two were by librarians - no doubt committed by their profession to achieving a solution to the transcription problem. Scherer was from Munich, and Schleiermacher was later to be librarian of the Ducal library in Darmstadt. They were awarded the prize jointly, but took opposite views about one of the crucial issues - whether to aim at a transcription of pronunciation or simply a transliteration. Scherer favoured the former, Schleiermacher the latter. Over the years the members of the Volney Commission and the contestants had sharp disagreements about what Volney's real intentions had been. There is no doubt that his ultimate aim was for a universal phonetic alphabet, but the question was whether to aim for a more limited objective, attainable in the immediate future.

Some of the problems involved in a transcription of pronunciation are:

1. to limit the sounds to be transcribed - there was as yet no clearly formulated notion of the phoneme, though it is clearly implicit in the mention in some schemes of 'fundamental' or 'important' sounds;
2. to be able to convey the pronunciation accurately so that non-specialists could understand and reproduce it. This required a satisfactory phonetic terminology - only in part available at that time;
3. to choose from among competing accents of a language;
4. to allow (at least if universal use is anticipated) for the incorporation in any scheme of 'new' sounds - i.e. to have an open-ended system;
5. to provide sufficient symbols for the sound distinctions required, and symbols that were aesthetically pleasing, easy to reproduce, and yet clearly distinct from each other.

3 SCHERER

Scherer's essay is admirably clear. While recognising the above problems, he believed that the benefits to be derived from a satisfactory phonetic transcription scheme would outweigh the difficulties. He had no intention of replacing existing orthographies - his alphabet would stand alongside them, helping to make oriental languages more accessible to the learner. He foresees the possibility in due course of what he calls 'philosophical' symbols, by which he means newly devised symbols, not taken from existing alphabets. However, the need to enlist wide public support, as well as considerations of expense, meant that the alphabet, to begin with, would have to use the easily available and familiar symbols of the roman alphabet, supplemented by some Greek letters or by other devices.

He sets out what he considers to be principles of a good notation:

1. No one sound is to be represented by more than one symbol, and no one symbol is to signify more than one sound.
2. Symbols should be chosen on principles of simplicity, consistency and accessibility in printing types.
3. They should be usable in handwriting as well as in printing.
4. They must include the marking of the 'syllabic accent' - ignored in most systems of transcription.

Where possible he aimed to combine a transcription of pronunciation with an indication of the original orthography. As regards scope, while accepting that in the first place limited groups of languages would be dealt with, he sees nothing in principle against the idea of a universal alphabet. Scherer envisages (far too optimistically as it turned out) that a solution to the problem could emerge in the following year's competition, and that the necessary tools would rapidly be made available - namely (a) the full notation system, with a suitable introduction and illustrations of its use; (b) a simple grammar, a chrestomathy and a vocabulary of two of the most important languages - Arabic and Sanskrit; (c) possibly a complete transcription of a selected oriental classic work. He sees the new alphabet as having five main benefits:

1. economy in the reader's time;
2. economy in printing costs;
3. attracting new students to oriental studies;
4. aiding language acquisition in general;
5. improving communication for all those coming into contact with foreign languages - administrators, travellers, traders etc.

Scherer's approach is practical, and the system he put forward in the following year's competition adhered closely to the principles he had set out in 1822. It dealt with phonetic transcription under the heading of 'phonography', and transliteration separately as 'semiography'. He attempted to combine the two by using lower case letters for phonography and upper case for semiography. He gives as an example the name Muhammad transcribed from Arabic. The semiographic version, transliterating the Arabic consonants, would be MHMD, but the combined version would be MuHaMMaD (the inserted letters representing the Arabic diacritics for vowels and for doubling the consonant). His phonographic alphabet contains 40 letters, of which all but 12 are taken from the roman alphabet, and most of the remainder only involve slight modifications of roman letters.

Scherer's phonetics is inevitably

incorrect in places, but he makes some interesting observations. He presents the vowels in a vowel diagram, which he specifically says does not relate to the vowels of a particular language. In his 1822 essay he had introduced a third dimension. Whereas the basic vowel sounds are envisaged as on the surface of a solid elliptical body, the vowel 'mute e'(i.e. schwa), which he describes as a "vowel yet unformed" is said to be in the centre (see diagram). In his 1823 essay he talks of this vowel as either "concealing itself within the vocal sphere" or "approaching the vocal periphery". He equates it with the colour grey which he regards as a mixture of all colours. (This comparison with colours is found in a number of early descriptions.) As with most early vowel diagrams the central line does not represent vowels with a central tongue position, but front rounded vowels.

4 BRIERE

It is impossible here to give more than a taste of the essays on transcription which were submitted for the Volney Prize in the first 20 years (after that the Commission decided to drop the topic of transcription, disappointed with the results of previous competitions). They range from the most limited (both in respect of the number of languages covered, and in adhering strictly to a transliteration rather than a phonetic transcription) to the most ambitious, attempting to provide for all sounds in all languages. This end of the spectrum obviously includes those which are of more interest to the phonetician. I shall confine myself to one of these more ambitious schemes - the most ambitious in fact. It was by a certain M. de Brière - a pseudonym - his real identity is still obscure. He first put it forward in 1827, and resubmitted it in 1831. It did not win a prize on either occasion, though Brière did eventually win in 1837 with a much more limited scheme.

The 1827/1831 proposal was entitled Phonographie cyriographique or idéographique - the art of representing the movements of speech by precise letters - that is, it was a universal alphabet. After a description of the speech organs, accompanied by somewhat crude diagrams, he presents his alphabet, basing it on the productive mechanisms involved. His categories are derived from a description of articulatory movements relating to the lungs, larynx, velum, hard palate, teeth, lips, tongue tip, jaw and cheeks. To give an idea of the detail of his description, 21 possible lip positions are allowed for, and 17 different positions of the tongue tip (see Appendix for examples). Larynx raising and lowering are taken into account. Brière derives 80 subclasses of sounds, and each of these is then subject to distinctions of what he calls 'intensity' - the degree of variation within each category.

Interesting, and unexpected in a description of this period, is his recognition of variations which are frequently totally ignored in descriptions of speech. He lists these as: speaker's sex, age, temperament, physical dimensions, state of health, body posture, situation, proximity to others, timbre of voice, tones of voice, character, emotional state, airs and manners, social position, national or provincial accent, epoch in which he is living, the temperature, the time of day, simultaneous activity, and type of text (if read). His 'normal' or 'neutral' case is: A Frenchman, from Paris, aged about 30, of average height, with good constitution and health, of average social status, before a meal, not affected by any emotions, of good character, speaking in a friendly way, in a standing position, using a moderate degree of loudness, in an amply furnished room of average size, about noon on a fine day in spring in the nineteenth century.

Although variations of this kind may be disregarded for many purposes (and certainly they go far beyond what the Commission was looking for), they are evidence of an open mind and an observant ear. Brière also provided for features of connected speech, such as assimilation. He calculated that his system would allow for the description of 43,923,168 sounds! The word 'overkill' springs to mind, but he made it clear that the number of sounds which would require to be symbolised in reality would probably not exceed 220. The degree of expertise which his system would call for if applied in full is obviously far beyond what the average user of a universal alphabet would possess, but it is stimulating to find such a search after precision at a time when so many essays were little more than rehashes of previous work, involving little or no new observations.

One other particular point of interest is that Brière assigns an 'organic name' to each sound, based on its formation. It is the same idea, in effect, as Jespersen's analphabetic notation ([2]). To take one example: the Indian retroflex stop is given the name tés-lé-rou, where <t> = tongue tip articulation, <é> = aspirated, <s> = breathed out, <lé> = mid-palatal, and <rou> = high intensity. The symbols of his notation are, he says, only an imprecise compromise forced on him by the Commission's requirements, and only the organic names can specify a sound with precision.

Finally, to provide for transliteration he uses a system of subscript numerals. He exemplifies it from Modern Greek, which retains the orthography of Ancient Greek, but has a much reduced vowel system. There are six ways of representing the vowel /i/. In transliterating Brière uses /i/ with the subscript numerals 1-6, so κοιτη is represented as ki_4ti_5.

5 CONCLUSION

In all 36 essays on transcription were submitted, of which six were awarded the prize. The sad fact is that none of them was deemed by the Commission to have presented a system which they could do more than commend as worthy of further examination or wider circulation before approval could be considered. Volney's hopes for the adoption of a new system with the backing of the French Academies were never realised, though the institution of the Prize stimulated many valuable works in the wider linguistic field. Many of these are to be published as part of a major project concerned with the Volney Prize Essays in the course of the next year or so.

REFERENCES

1. DE BROSSES, Charles (1765). Traité de la formation mécanique des langues et des principes physiques de l'etymologie. Paris.
2. JESPERSEN, Otto (1889). The articulation of speech sounds represented by means of analphabetic symbols. Marburg.
3. JONES, Sir William (1788). "Dissertation on the orthography of Asiatick words in Roman letters". Asiatic Researches 1:174ff.
4. VOLNEY, Constantin François Chasseboeuf (1795). Simplification des langues orientales ou méthode nouvelle et facile d'apprendre les langues arabe, persane et turque avec des caractères européens. Paris.
5. VOLNEY, Constantin François Chasseboeuf (1819). L'alfabet européen appliqué aux langues asiatiques. Paris.
6. WILKINS, John (1668). An essay towards a real character and a philosophical language. London.

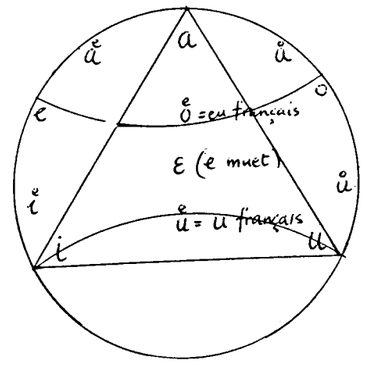
APPENDIX

EXAMPLES OF BRIERE'S CLASSES OF SOUNDS

1. pulmonité - lung movement
2. gutturalité - larynx action
3. glosso-staphylinité - tongue movement in relation to soft palate
4. nasalité - soft palate movement
5. lingualité - tongue tip movement
6. palatalité - contact with palate
7. maxillarité - jaw movement
8. dentalité - contact with teeth
9. labialité - lip position
10. oralité - mouth opening
11. genalité - inflation of cheeks

Examples of subclasses

1. gravi-gutturalité - larynx lowered
2. acuti-gutturalité - larynx raised
3. cavi-lingualité - tongue point lifted and curved
4. extensi-maxillarité - jaw pushed forward
5. lati-maxillarité - jaw moved sideways
6. distensi-labialité - lips lengthened
7. retracti-labialité - lips drawn back



Scherer's 1822 vowel diagram

E.Nushikyan

THE TYPOLOGICAL ANALYSIS OF EMOTIONAL SPEECH PROSODY

EMMA NUSHIKYAN

English Dep. Odessa State University Odessa, Ukraine,
USSR, 270021

ABSTRACT

Emotion and its linguistic expression form a system: information about emotion comes by lexical cues, syntactic structures and prosodic indicators. This paper describes some further attempt to iden tify the acoustic parameters of emotional texts in English, Russian and Ukrainian.

INTRODUCTION

In recent years a convergence of interest has developed among linguists /1/, psychologists /2/, physiologists /3/ and other specialists in the theory of emotion. Significant research has been conducted in the area of the verbal and non-verbal types of emotional expression /4/.Many recent writers on the manifestation of emotion have found it natural and useful to state that emotion, language and speech are related in some way as language being the principle mode of communication is also a means of expressing emotions and arousing them in others. The main aim of the present study is to reveal the system of linguistic devices of expressing emotions in English, Russian and Ukrainian.

APPROACH

Emotion and its linguistic expression form a system. In order to characterize and understand its function we need to consider the system properties.

Our investigation of 30000 pages of English fiction has pointed out that all levels of linguistic system may be involved in the process of emotional manifestation.

The syntactic analysis of the material under investigation proves that the structural means of expressing emotions form a special syntactic code. Language inspired by emotion undergoes a wide variety of formal changes - a breakdown of grammatical structure of the sentence, repetitions of subjects, subjects + predicates, introduction in the sentence structure of the formal elements - interjections, addresses, particles and so on.

At the lexical level we must find to what extent words alone can still yield further information in the study of emotional speech. Usually emotions are identified in the text due to the adjectives which emphasize the high degree of some quality (e.g. magnificent, awful etc.), the corresponding adverbs (e.g. terribly, exceedingly etc.), nouns (e.g. treasure, ruffian), verbs (e.g. hate, adore).

Despite the fact that in the majority of cases the emotive meaning depends on the integration of prosodic, syntactico-semantic and contextual information the role of intonation must not be ignored as there are cases in which it is the only

carrier of affective information.

SUBJECTS AND SPEECH MATERIAL

In this research ten English, six Russian and four Ukrainian speakers participated as subjects. To preserve the identity and comparability of the experimental material 115 English texts and their Russian and Ukrainian translations were chosen. These texts expressed the 16 most frequently observed positive and negative emotions. As the emotional aspect of textual prosody implies that the effective messages can be superimposed on the neutral texts the material was recorded twice - as samples of emotional and non-emotional speech. The tape was then presented to 30 Russian, 30 English and 30 Ukrainian listeners who where instructed to listen carefully to the presentation of the texts and to decide which emotion the speaker was trying to convey. A criterion of 95% correct identification was set for including the text into the further instrumental analysis.

The original speech material was instrumentally analyzed into separate components - fundamental frequency, amplitude, duration and spectral composition. Oscillograms were obtained with the help of the Visi-Pitch and M-4030-1 computer which has been programmed to yield the desired speech parameters. Spectrograms were made on Sona-Graph of the Kay Elemetrics Corporation, using the wide-band filter (300 Hz). The synthesis of emotional speech was done on formant synthesizer.

DATA ANALYSIS AND RESULTS

The typological study of emotional and neutral texts in English, Russian and Ukrainian disclosed common ways of expressing emotion. The detailed contrastive analysis of their acoustic structure has revealed that the frequency range, frequency interval of the terminal tone, of the semantic centre are the most informative parameters differentiating emotional and non-emotional speech in all the languages under study. In this way the similarity of emotion expression is manifested. The application of methods of mathematical statistics (t ratio, Student's t) proved that the difference between the above mentioned acoustic parameters of emotional and non-emotional speech was statistically significant as they belong to two different population variencies.

Our study presents experimental evidence that it is the movement of the fundamental frequency (its configuration) which bears the meaning of emotional intonation in a pecular language. For instance, the utterances expressing anger are pronounced by the majority of English speakers with a sliding scale, while Russian and Ukrainian speakers use the broken descending scale with two peaks of the fundamental frequency. The corresponding neutral utterances are pronounced with the gradually descending scale with the peak of the fundamental frequency on the first stressed syllable. On the whole in the majority of emotional texts (with the exception of the texts expressing sorrow, tenderness, offence) a more complicated character of the fundamental frequency was observed.

It appears probable that both configuration and pitch levels have to be specified for certain emotions in the languages under study. The pitch level is higher in the texts expressing anger, delight, joy, amazement and lower in those expressing sorrow, guilt, tenderness, offence as compared with the neutral texts.

The textual aspect of prosody enables us to state the degree of communicative text dynamism. Each text contained sen-

tences in which semantic and emotive information was concentrated. They were called communicatively strong, while the other sentences were called communicatively weak. It is the range of the fundamental frequency, the frequency interval of the first and the last stressed syllable which is greater in communicatively strong sentences as compared with the communicatively weak ones in all the languages under study.

The study of the hierarchical structure of the temporal composition of emotional and non-emotional texts shows that the difference can be observed at all the levels - separate sound, syllable, utterance and text.

The increase in duration of utterances expressing irony, joy, contempt comes from the increase of vowel duration. The duration properties of consonants influence are also of great importance for emotional speech. The analysis of the spectrograms of emotional speech shows a longer stop gap and a more intense burst at the release of the English p, t, k and a lenghtened interval of vocal tract contraction for ʃ, s, tʃ , dʒ . For Russian the increase in duration of fricatives s, ʃ, is of greater importance, for Ukrainian -ɦ , x , š: .

The mean duration of the syllable and the duration of the first and last stressed syllable are also characteristic differentiating parameters for emotional and non-emotional speech. The research shows that duration is the most variable acoustic parameter.

The temporal organization at the textual level reveals specific segmentation of emotional texts. Emotional state of the speaker has manifested in the change of duration of phonation parts of the texts and pauses. To measure this influence the relative informative duration parameter

θ has been introduced.

$$\theta = \frac{\sigma_{ph} + \sigma_{p}}{\sigma_{p}}$$

where σ_{ph} - phonation duration,

σ_{p} - pause duration.

The quantitative analysis of the intensity of emotional texts proves that in most cases greater total energy of the text is observed due to the increase of energy of the first and last stressed syllable. The decrease of total energy occurs in the texts expressing sorrow, tenderness, offence, worry.

Spectrographic measurements of formant frequencies of emotional and neutral speech show a constant increase of the total energy of the stressed vowel of the nucleus at the expense of F_1 and F_2 in emotional speech. We also found the frequency range enlargment as well as the greater importance of F_3 and F_4. The shift of the intensity of formant frequencies of the stressed vowels of the nucleus into higher regions was notices in the utterances expressing strong emotions (e.g. rage, amazement, threat, delight and so on) and into the lower regions in the utterances expressing sorrow, offence, tenderness. The spectrographic analysis also reveals the greather role of the high frequency noise regions of consonants in emotional speech.

Statistical analysis of the main acoustic characteristics shows that each of them can distinguish emotional and neutral texts, but none has a differentiating function. The integral use of these characteristics enables us to create the algorithm of recognition of emotional and neutral utterances. For creating this algorithm A.Wald's sequential analysis based on the likelihood ratio has been applied.

The final stage of instrumental analysis is presented in a series of experiments on synthesis of emotional speech the aim of which is to check the validity of the acoustic parameters of natural emotional speech and to form the rules for high quality formant synthesis.

The synthesis programme was as follows: first the so called "neutral" sentence was modelled, then the emotional one with the identical lexico-grammatical structure.Alltogether 30 Russian sentences expressing delight, surprise, anger, fear were obtained. The constructed model comprised the normalized contours (Fon) of an initial, mid and final accent groups (A.G.), each containing prenuclear (1), nuclear (2) and postnuclear parts (3). Fig. 1 represents the normalized fundamental frequency contours for sentences, expressing surprise (a), delight (b), fear (c) and the same neutral ones.

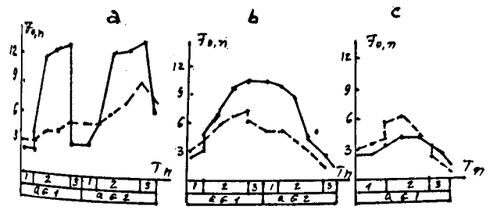


Fig. 1. Normalized fundamental frequency contours of sentences expressing (a) surprise, (b) delight, (c) fear are plotted on the contours of the same neutral ones. The solid line curve displays the emotional fundamental frequency contour, while the broken line displays the corresponding neutral one.

The high degree of perceptual acceptability of the synthesized emotional speech was proved by the auditive analysis. The tape was presented to 45 students of Russian department of Odessa University instructed to identify the type of emotion. Their answers presented in table 1 demonstrated the correct application of the synthesis rules.

Table 1. Identification of the type of emotion in synthesized speech.

Type of emotion	: Correct answers (%)
surprise	: 95
delight	: 90
fear	: 73

Our experiments on synthesis show that for expressing some emotions (e.g. surprise) fundamental frequency is the most important prosodic variable, and others need only minimal changes. For expressing other emotions some changes of formant structure must be added to a neutral sentence.

CONCLUSIONS

The systematical level approach to emotion realization in a text made it possible to single out some regularities in the interaction of components in this system and to disclose the isomorphism in its expressive means.

The results of this typological study permit us to suppose that prosodic structure of emotional texts in English, Russian and Ukrainian displays universal as well as particular properties in manifestation of emotion.

REFERENCES

/1/ Bryzgunova E.A. Emotsionalno-stilisticheskije razlichija ruskoj zvuchaschej rechi. - Moscow: MGU, 1984.
/2/ Vitt N.V. Emotsionalnaja regulatsija rechevogo povedenija // Vopr. psihologii, 1981. - N 4. - S. 60-69.
/3/ Simonov P.V. Emotsionalny mozg. - Moscow: Nauka, 1981.
/4/ Williams C.H., Stevens K.N. Emotion and speech: some acoustic correlates // J. Acoust. Soc. Am., 1972. - v. 55. - P. 1238-1250.

INTONATIONAL UNIVERSALITIES AND PERCEPTION OF EMOTIONAL INTONATIONS

IRINA YUROVA

Dept. of Foreign Languages
Institute of Aviation Instrument Making
Leningrad, USSR, 190000

ABSTRACT

The article presents the results of an investigation on the problem of universality in the intonational means for expressing emotions. Identification of seven modal and emotional connotations (regret, joy, surprise, irritation, insistence, doubt and evidence) is analysed. It is concluded that the adequate perception of emotional intonations and consequently the degree of their intonational universality depends on the type of modal and emotional connotation.

Thanks to a great number of various studies the information on the similarity of some intonational forms in different languages has been obtained. The intonational resemblance of languages is considered to be the largest in the field of expression of emotions. Actually the data obtained is not sufficient to judge whether this statement is true. The analysis of perception of intonations can serve as one of the ways of finding out intonational universalities: the identification of modal and emotional connotations in a foreign language may be regarded as a proof of universality of the means of their expression. It seems that the more alike the intonational representations of some emotional connotation in two languages (native and foreign) are, the easier this connotation will be identified in the foreign language.

The investigation based on the consept mentioned above has been carried out. It dealt with the perception of French intonations expressing surprise, joy, regret, irritation, insistence, doubt and evidence. The majority of the phrases was sampled from the recordings - suppliments to the manuals of French phonetics for foreign students /1/, /2/, /3/, their authors regarding these phrases standard. Some phrases were specially composed and recorded in the interpretation of a Frenchman to obtain the more homogeneous and complete material.

Three groups of listeners participated in the experiment. The first group (I gr.) consisted of 70 Russians who did not speak French. So they could use intonational characteristics only, Russian intonational system taken as a basis. The second group (II gr.) united 30 Russians speaking French; they understood the meaning of the phrases, had a large experience of perception of the spoken French and were well acquainted with the intonational system of French on which they could base. The third group (III gr) consisting of 14 native speakers of the French language participated in the experiment as an expert group.

The listeners of all the groups were to determine a) the presence or absence of emotion in a phrase and b) the type of emotional connotation. The phrases were presented either separately or in pair with a neutral phrase (narrative, interrogative or incomplete) having the same or similar lexical content. In order to simplify to some extent the identification of modal and emotional connotations the listeners were proposed to use the list which contained the denominations of ten connotations: seven of them were really presented while three (anger, fear, tenderness) were not included in the material to be analysed. Besides, it was specially indicated that listeners could use any other denomination of emotion and not only those given in the list. They were also allowed to mark the presence of two and more emotions underlining if they could the main one. All the answers were taken into account. The total number of the answers obtained is more than 65 000.

Statistical tests (a t-test and a sign-test) were used to compare the results obtained from different groups, in different series of experiment and to find out the contensity of the data.

The results of this analysis are presented in tables 1-3.

All the groups of listeners perceive the presence of emotional connotation (the results are consistent) in an utterance even in case when there are neither lexical nor contextual support (see table 1).

Table 1

Perception of the presence of emotional connotation in the phrases presented separately (average per cent)

Emotions \ Listeners	I gr.	II gr.	III gr.
surprise	70,0	78,4	93,6
joy	94,2	97,6	100,0
regret	93,9	99,2	98,3
irritation	77,7	56,8	88,3
insistence	83,8	81,1	98,8
doubt	64,4	53,7	92,4
evidence	85,3	91,1	96,2

The presence of emotional connotation was perceived best of all in the phrases expressing joy, regret, evidence and insistence. The results were worse for the phrases rendering surprise and irritation. The lowest results were obtained for the sentences expressing doubt, their intonational structure being similar to neutral questions.

In all cases the presense of emotional connotation was detected better by the French listeners: naturally, they were able to perceive slight distinctions between emotional and neutral phrases even without their comparison. The Russian listeners gave the similar results only when emotional phrases were presented in pairs with neutral ones (see table 2).

Table 2

Perception of the presence of emotional connotation in the phrases presented in pairs with neutral ones (average %)

Emotions \ Listeners	I gr.	II gr.
surprise	87,7	90,8
joy	99,1	100,0
regret	99,0	100,0
irritation	96,9	98,7
insistence	96,1	96,9
doubt	87,9	89,0
evidence	87,0	92,2

The Russian listeners were able to solve a more difficult problem: to determine a type of emotional connotation in a phrase presented even separately (see table 3).

In identification of utterances expressing surprise the results given by the Russian listeners were almost as good as those given by the Frenchmen: the average percentage of adequate perception by all the groups were found to be similar, all the phrases being perceived non accidentally (from the statistics point of view).

The phrases rendering regret were identified worse than the phrases expressing surprise: in all the sentences the type of emotional connotation was determined correctly, but the average percentage is already lower. The command of French by the Russian listeners (II gr.) is assumed to improve considerably the results.

Table 3

Identification of emotional connotations in phrases presented separately (average per cent)

Emotions \ Listeners	I gr.	II gr.	III gr.
surprise	54,2	60,0	68,7
joy	35,0	41,2	22,4
regret	37,9	59,8	57,6
irritation	29,2	25,3	53,1
insistence	40,1	32,6	74,2
doubt	13,0	7,1*)	40,3
evidence	5,4*)	18,9	37,2

Note. *)-the result is inconsistent.

The phrases expressing insistence were identified by the Russians much worse than by the French listeners. This proves that the determination of the type of emotion in a foreign phrase is based on the melodic component of intonation: despite intensity differences and tempo features distinguishing those phrases from neutral narrative ones the Russian listeners made many mistakes in their identification.

The identification of the kind if connotation in the phrases rendering doubt and irritation by the Russian listeners proved to be more difficult. Furthermore, the command of French (II gr.) did not improve the results.

The lowest results were obtained for connotation defined by the French authors as "evidence". Although all the groups of listeners detected the presence of emotion in these phrases (see tables 1 and 2) the Russians of the first group correctly determined the kind of connotation only in one sentence. They confused these phrases mainly with those expressing surprise and doubt. The listeners of the second group making the same mistakes however were able to identify - though without assurence - this modal connotation. It is worth mentioning that the Frenchmen themselves had difficulties in the identification of this emotion.

The identification of the phrases expressing joy is to some extent paradixical: this type of connotation was determined better by the Russians than by the French listeners. The adequate perception of these phrases by the Frenchmen might be hindered by their neutral lexical content, while the Russian listeners who based on intonational characteris - tics - only (I gr.) or mainly (II gr.) - solved the problem easier.

The perception errors may be the result of inaccurate pronunciation on the one hand, and on the other they may be accounted for the similarity of intonational presentation of phrases expressing different emotions. Thus, the listeners of the first group confused doubt and evidence, doubt and surprise.

The adequate perception of an emotion depends not only on the understanding of the phrase lexics (in our experiment the lexical content was neutral and the same phrases were produced with different emotions) but also on the knowledge of the situations of real life which determine the use of this or that intonation /4/, /5/. For the French listeners it was easy to reproduce the situation in which a given phrase was pronounced. The task was much harder for the Russians. The pre - sence of intonational context (presentation in pairs with intonationally neutral prases simplified the identification considerably. The command of French to some extent helped the Russian listeners to identify the connotations. But in general, the influence of intonational system of the mother tongue appeared to be dominant.

The acoustic characteristics of all the sentences were also analysed: fundamental frequency, sound duration and envelope amplitude were measured.

Almost all the emotional connotations studied were found to be characterised by the change of all the intonation components as compared to neutral phrases. Melody changes proved to be the most significant (as it was to be expected) for perception. The change of frequency range is observed in the majority of emotional phrases (see table 4).

Table 4

Acoustic characteristics of emotional and neutral phrases (average values)

Acoustic characteristics / Emotions	Frequency range (semitones)	Sound duration (ms)	Peak amplitude range*)
surprise	14,6	67,9	0,44
joy	14,2	84,9	0,22
regret	6,6	85,3	0,72
irritation	9,2	82,3	0,38
insistence	8,6	118,8	0,50
doubt	8,4	74,4	1,32
evidence	10,6	79,2	0,82
neutral	8,8	77,1	0,86

Note *) The peak amplitude range is the difference between the maximal and the minimal values of the relative maximal intensity of vowels in a phrase.

In case when the connotation is positive (joy) the frequency range widens. If the emotion is neither positive, nor negative it may occur both the widening (surprise, evidence) and the narrowing (doubt, insistence) of the frequency range. When a negative emotion is rendered the frequency range becomes narrower (regret), but it may remain practically unchanged or even widen a little (irritation). The latter statement contradicts in some way the other authors' data concerning the narrowing of the frequency range of a phrase in case of a negative emotion /6/, /7/.

The difference in tempo characteristics in emotional phrases is less significant. In general these utterances are reproduced slower than the neutral constructions. In some cases the tempo becomes more even (insistence), in other cases the tempo contrasts between the begining and the end of the utterance become greater (joy, regret). Besides, the lengthening of vowels and the shortening of consonants - particularly in the last syllable - is characteristic of those phrases.

The intensity structure is also important in emotional speech. To some extent it proves to depend directly on the type of emotion to be rendered (it correlates with the opinion of other authors /8/). Smoothening of the intensity structure of emotional phrases in comparison with neutral ones is found out to be their constant feature independent on the type of connotation: the relative peak amplitude range of vowels varies to a smaller extent than in neutral sentences (see table 4).

Thus, it may be assumed that the presence of any emotion is detected not only in one's native language, but also in a foreign one. Listeners detect melody , tempo and intensity deviations from neutral pronunciation, judging by the melodic structure firstly, tempo and intensity being not so important. Besides, it is the general intonational structure that is significant in fireign speech perception (the tune shape, the general tempo, the intensity level of the whole phrase). Little attention is payed to the correlation of acoustic features within the phrase.

In phrase perception in his native language the listener pays attention to intensity and tempo characteristics also, taking into consideration the slightest intonational structure. Thus, the melodic component of intonation is assumed to be the most universal characteristic of a phrase.

From the point of view of perception the hierarchy of the analysed emotional connotations may be proposed. Both the Russian and French listeners identify the phrases expressing surprise, regret and insistence with a certain assurence. So these connotations are alike to the greatest extent in those two languages.

The degree of identification of the phrases rendering irritation and joy by the Russians and the Frenchmen is quite different. It is evident that the intonational structure of these connotations is more specific.

The results of perception of utterances rendering doubt and evidence show that these connotations are differently expressed in Russian and in French.

It should be noted, however, that the character of perception of emotional connotations is connected with some other factors (psychological ones being among them) which are worth analysing.

References

1. P.Léon, M.Léon, "Introduction à la phonétique corrective ", Paris, 1968.
2. Callamand M., "L'intonation expressive", Paris, 1973.
3. G.Calbris, J.Montredon, "Approche rythmique intonative et expressive du Français langue étrangère", Paris, 1975.
4. I.G.Torsujeva , "Intonatsyia i smysl vyskazyvanija" (Intonation and meaning of utterance),Moscow, 1979.
5. G.Brown, "Listening to spoken English", Moscow, 1984.
6. G.S.Poslushajeva, "Intonatsyia vosklitsatelnyh fraz sovremennogo frantsuzskogo jazyka" (Intonation of exclamatory phrases of contemporary French), Moscow, 1971.
7. P.Léon, "Essais de phonostylistique", Studia phonetika,4, Montréal, Paris, Bruxelles, 1971.
8. N.D.Svetozarova, "Intonatsyonnaja sistema russkogo jazyka" (Intonational system of Russian), Leningrad, 1982.

PERCEPTIVE AND ACOUSTIC CHARACTERISTICS OF EMOTIONS: A Typological Research Based on the Material of Languages with Different Structures

L.V. ZLATOUSTOVA, G.Y. KEDROVA

Dept. of Philology
Moscow State University
Moscow, USSR, 119899

ABSTRACT

The present paper offers results of experimental phonetic research carried out with the purpose to solve some problems of describing different emotional states from the view of peculiarities of their phonetic expression in differently structured languages. With the help of complex techniques of auditory and acoustic analysis the main types of emotions relevant for speech communication and their phonetic parameters are determined.

INTRODUCTION

At present considerable interest is taken in the problem of phonetic expression of different emotional states, which is determined both by - the development of fundamental linguistics (mainly in the direction of communicative linguistics) and by the need to solve a number of most important applied tasks, such as diagnosing the emotional state of the speaker and listener by speech, normal and pathological emotions, identification of a person, as well as improving automatic speech synthesis and analysis. However, despite considerable success in the study of physiology and psychology of a person's emotional state, there has been produced no linguistic theory so far, which would describe and explain on all linguistic levels the similarities and variations in the phonetic expression of emotions in different languages in connection with the general type of a language structure. This task can be approached by first - revealing the typology of intonational patterns, corresponding to the types of emotional states relevant to speech communication; then - detailed description of their phonetic variants and variations; and finally by indexing emotional-phonetic "distinctive features" and their "weight" coefficients.

The paper presents the results of an experimental phonetic research undertaken with the view of solving some of the above-mentioned tasks seem to us very urgent, namely, revealing the types of emotions which are most frequent in speech communication; establishing their phonetic parameters; preliminary observations on the typology of phonetic expression of emotions in different languages. The research has been conducted on the material of one-word utterance by complex techniques including auditory and acoustic analyses and statistical analysis of the obtained data. The research has been mainly oriented towards the Russian language.

1.MATERIAL AND TECHNIQUES OF OBTAINING RECORDS

The study of Russian plays and especially of stage directions which contain lexical designation of an emotion, has revealed the most frequent emotional states, seventeen of which have been selected as having the maximum frequency coefficients. Together with the neutral emotional state they have made up the initial list of emotional states to be analysed, i.e. neutral, pleasure, joy-delight-admiration, displeasure, indignation, anger, malice-hatred, irritation, contempt, rage, irony, menace, reproach, fear-fright, entreaty, despair, distress-bitterness, surprise. With these emotions by means of modelling an appropriate situation the word "посадка" (LANDING) has been uttered by four native speakers and a professional actor (the age range - from 25 to 40) The words designating the emotional states as well as the cue word have been translated into English, French, German, Spanish, Polish, Czech and Bulgarian. The native speakers of these languages have been given the same task. To increase the degree of phonetic comparability of the naterial the words in all languages were selected so that they would be similar to the Russian word "посадка" in their rhythmical and segmental composition. The recording has been made in a studio.

2.AUDITIVE AND PERCEPTIVE EXPERIMENTS

The records have been offered for auditory analysis to both native speakers of Russian (20 - 40 persons) and native speakers of other languages (5 persons for each language). At the first stage it was necessary to identify the emotion as one from the list. At the second - to judge whether the produced utterance corresponds to the implied emotion or not. The group of Russian auditors (6 persons) were to assess the identity and difference of all possible combinations (pairs) of emotional realizations. Another group (16 persons) was to conclude as to the subjective similarity or dissimilarity of the emotions under analysis on the basis of the list of pairs of emotional states given to them.

3.ACOUSTIC ANALYSIS

Sonagrams (frequency range - up to 4 kHz) and oscillograms (intensity variations) have been made. The length of segments, changes of the voice pitch on the vowel segments and the maximum values of segments' intensity have been calculated.

4.EXPERT ANALYSIS OF SONAGRAMS

Realizations of all emotions produced by three Russian speakers and presented on sonagrams without indication of the emotion or the speaker have been offered to an expert in sonagrams with the task to class the multitude of these realizations according to certain features chosen by the expert himself. The feathres he chose and the classes they gave were registered in protocols.

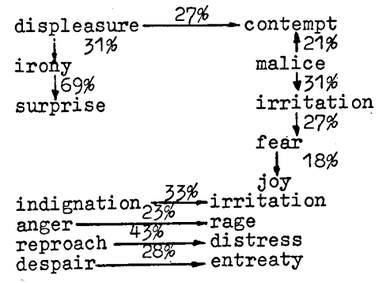
5.RESULTS

a)For the Russian language

The results of the acoustic and auditive analysis for the Russian language in the cases of all speakers show a rather high degree of concordance. Among the speakers under analysis it was possible to single out one whose emotional realizations were in auditors' opinion the easierst for identification. He also has the least disparity between the sounding and the implied emotion. His data will be used futher on as an illustrative material.
As a result, the following conclusions have been made: most of the emotions can be differentiated by ear rather easily. Anger, rage, irony, surprise, malice, irritation and entreaty are most distinctly opposed to all others; pleasure, contempt, distress, reproach and indignation are on the contrary the least distinguishable. Besides the assessment "similar stimuli", the assessment "identical stimuli" has been introduced. Alongside with the pairs made by the repetition of the same stimulus, certain realizations of irony and surprise have been recognized to be similar.

The identification of emotional states by the phonetic form of the stimulus in auditive experiments has yielded the following results: by degree of correct identification the emotions under analysis are ranged as follows: neutral (86% correct identification), surprise (88%), fear (68%), rage (75%), menace (75%), distress (69%), joy-delight-admiration (68%), entreaty (62%). Statistically relevant mistakes in the identification of emotions are given below in diagram 1 (it is noteworthy that these pairs of emotions' realizations were found similar in their phonetic form).

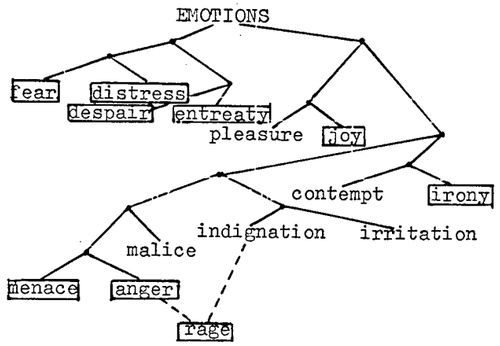
Diagram 1. Graph of confusions between emotional states in their diagnosis in speech



When estimating the acceptibility of stimuli the majority of the auditors found all stimuli natural for the given speaker except the emotion of reproach.
Thus, it is clearly seen that not all differences in phonetic form which can be perceived by ear are of equal importance for the emotions' identification. Therefore the question of relevant and irrelevant differences in the phonetic expression of emotions is closely connected with the subjective estimation of the distance between the emotional states under analysis. The subjective space of the emotions in question is as follows: surprise, reproach and neutral emotions do not form any groups and are opposed to all other emotions and to one another. The relationships among other emotions are given below in diagram 2.
The comparison of the subjective estimation of the distance between the emotions with the data of auditive experiments shows that intense emotions as well as biologically vital ones (such as reaction to a danger, signal of agression, etc.) have certain advantages in the process of identification. The data obtained as a result of perceptive analysis make it possible to range the results of the instrumental acoustic analysis as communicatively relevant parameters of sounding and peculiarities connected with the individual manner of the speaker and other cha-

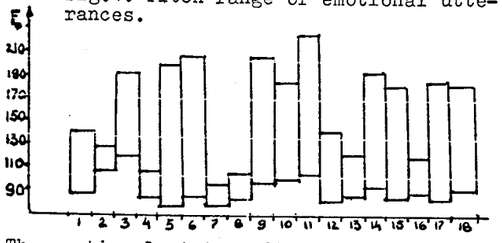
nce factors.

Diagram 2. Similarities/dissimilarities between the emotional states.



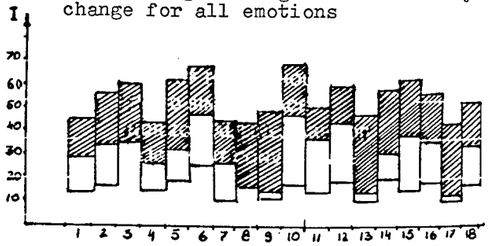
Emotions that are intonationally marked and for this reason are easily and correctly identifiable are taken in a framework.
Voice Pitch. It seems convenient to begin the presentation of the results of the instrumental acoustic analysis with the Voice Pitch. Pitch range and the average level of F$_0$ proved to be rather significant for the differentiation of emotions. Fig.1 shows the Pitch range of all emotions, 1 - being neutral utterance, 2 - pleasure, 3 - joy-delight-admiration, 4 - displeasure, 5 - indignation, 6 - anger, 7 - malice 8 - irritation, 9 - contempt, 10 - rage, 11 - irony, 12 - menace, 13 - reproach, 14 - fear-fright, 15 - entreaty, 16 - despair 17 - distress-bitterness, 18 - surprise.

Fig.1. Pitch range of emotional utterances.



The emotional states with marked Pitch range are: pleasure, displeasure, malice, irritation, despair (minimum range of F$_0$), surprise and irony (maximum range of F$_0$). On the whole, the Pitch range of all emotions except those with marked minimum range, is larger than neutral. According to average Fo value, displeasure, malice and irritation are lower than neutral, while joy, contempt and irony are higher than neutral.
According to the form of Fo, two types of the contour are important for the expression of different emotions: level contour (level-falling) and rising (rise-fall)one. It is noteworthy that the form of Fo contour influences also the tempo deformations of the rhythmical word structure. The tone contour is an important parameter for the emotions of menace (level), distress (level-falling), contempt (level-falling), malice (level-rising), entreaty (rise-fall). This parameter is especially significant for the realization of surprise and irony in one of its phonetic variants: it is rising and rising-falling in these cases.
Intensity. The overall range of intensity changes is not very informative: on the whole for all emotions it is larger than for a neutral utterance (except displeasure and irritation). See Fig.2.

Fig.2. Range of segments' intensity change for all emotions



More informative is the range of intensity change on vocal segments (the shaded area). It is minimal for displeasure, anger, rage, menace, irony and malice. The average intensity of vowels and the maximum values of intensity correlate with the differences in overall intensity of the stimulus. It becomes possible to differentiate between intense emotions, i.e. anger, rage, joy, menace and entreaty, and weak emotions, i.e. distress and contempt. For intense emotions the average peak values of intensity on one segment are 31.6 - 34.1 dB, for weak ones - 18.3-20.8 dB, for neutral utterance - 26.6 dB.
According to the relative intensity of syllables in the total intensity of stimulus we can make up a group of emotional realizations with strong emphasis on the end. These are: anger, fear, menace, irony and surprise (though in a lesser degree). Similarly we can form a group of emotions with weakened post-tonic part. They are: contempt, distress, irritation, and to a smaller extent - joy and malice. Besides, irritation, contempt, indignation, distress and fear are grouped as having an emphasis on the stressed syllable. The relative intensity of syllables within the total intensity of the word is as follows:

emotions with strong emphasis on the end

anger	.21	.44	.35
fear	.22	.46	.32
menace	.24	.42	.34
irony	.26	.38	.35
surprise	.26	.41	.32
indignation	.23	.48	.28

emotions with weakened post-tonic part

contempt	.33	.48	.19
distress	.32	.47	.21
irritation	.28	.51	.21

Duration. The fluctuations of the total word duration from speaker to speaker are so substantial that it is impossible to rely on the absolute values when judging the changes of tempo. It can be pointed out in general that the maximum duration is typical of rage and menace, the minimum - of irritation and displeasure. According to the duration of unstressed syllables in relation to the stressed one, the emotions under analysis can be ranged as follows:

for post-tonic part		for pretonic part	
distress	1.02	rage	.38
neutral	.98	irritation	.36
menace	.89	distress	.33
fear	.88	menace	.32
anger	.88	pleasure	.28
pleasure	.85	displeasure	.28
irritation	.82	irony	.28
rage	.81	fear	.26
malice	.80	malice	.26
displeasure	.78	joy	.25
entreaty	.78	indignation	.25
joy	.78	entreaty	.24
surprise	.77	anger	.23
irony	.75	surprise	.22
contempt	.74	despair	.20
indignation	.73	contempt	.18
despair	.58	reproach	.14
reproach	.51	neutral	.36

As it can be seen, distress and menace have the maximum duration of unstressed syllables, which leads to levelling-out of all the syllables. On the contrary, in despair and reproach we find considerable reduction of post-tonic and in reproach and contempt - of pretonic syllables. In the realizations of irritation and rage pretonic syllable is marked by unusually long duration. On the whole the ratio of longer and shorter syllables in emotional speech is different from that in neutral speech. The former has greater duration emphasis on the stressed syllable.
The length of certain phonetic segments in respect to the length of the whole word makes it possible to reveal cosiderable increase of sibilants (in our case S) in the emotional states of malice, rage, displeasure as well as the decrease of voiceless plosives in post-tonic part in rage and anger.
The share of vowel segments in the total duration of the word shows the degree of the word's vocalization. According to this parameter we distinguish between "voiceless" emotional states, namely dispair, displeasure and contempt - and "voiced" emotions, i.e. anger and rage.
Spectrum. The results of the expert analysis of sonagrams have not enabled us to single out definite spectrum peculiarities connected with this or that emotional state. According to such features as the position and relative intensity of upper formants (F3 and F4), frequency area and localization of noise in plosive consonants and spirants, only the speaker's individuality can be determined with sufficient validity. Thus, it has to be stated that frequency range up to 4kHz might be not large enough to reveal the acoustic differences in the spectrum (to be presented in sonagrams) which could be correlated with those in timbre.

b)Results of the analysis for other European languages

The results of the perceptive and instrumental acoustic analyses of some other European languages have been used in the first place to confirm or refute the hypotheses concerning the relevance of these or those phonetic features of different emotions as well as to define the relative weight of different parameters. As it appeared, the phonetic system of a language and caused by it "phonetic background" of a listener influence greatly not only the correctness of identification but the very possibility of making some definite conclusion about the emotional state of the speaker. In general it is more difficult to judge emotions in a foreign language than in a native one.
For Russian auditors identifying the emotional state, the most weighty parameter is the form of tone contour if only it is not falling. Despite the peculiarities of segments' duration or their intensity, such utterances are perceived as surprise. Deformation of duration structure, like for example in languages with phonologocally long vowels, also leads for Russian listeners to recurrent mistakes in emotions' perception, especially of those characterized by serious deformations of rhythmical structure typical of a neutral utterance, i.e. despair, entreaty and irony. Differences in greater or lesser tenseness of the articulatory basis of a language tell on the perception of the emotions based on contrasts in intensivity (both between words and segments of words). The best identifiable in all languages are such emotions as surprise, distress, anger rage, despair-entreaty, irony and also neutral utterance.

THE THEORY OF COMPLEMENTARY DISTRIBUTION AS A TOOL OF DISCOVERY: AN ANALYSIS OF THE PHONEMIC STATUS OF THE SIBILANTS IN THE VENEZIA GIULIA VARIANT OF STANDARD ITALIAN

CARLO FONDA

Concordia University
Montreal, Quebec, Canada
H4B 1R6

Abstract

In Standard Italian the phonemic status of [s] and [z] in the intervocalic position has remained to this day a contentious issue. Given the impasse of finding a satisfactory explanation of the sporadic nature of voicing in the Italian sibilants within the taxonomic framework of the phonemic system of Standard Italian, I repeated the attempt in the case of the Venezia Giulia dialect, and for the following reasons first, this is the dialect I am most familiar with; second, the experiment might prove an excellent opportunity to test A. Liberman's claim that the theory of complementary distribution is useless as a tool of discovery.
The following are my findings:

1- In the Venezia Giulia dialect sound system /z/ is a distinct and separate phoneme not to be confused with the voiced allophone of /s/.
2- Contrary to what A. Liberman maintains the theory of complementary distribution can be a useful tool of discovery.

In Standard Italian (S.I.) the phonemic status of [s] and [z] in intervocalic position has remained to this day a contentious issue. Witness analysts like L. Romeo [1] and J. Arce [2] who maintain that [s] and [z] are separate phonemes while others like R.J. Di Pietro [3], R.A. Hall Jr. [4], Christopher Court [5], G. Porru [6], place both sounds in the same class albeit [z] is an allophone of [s]. Among the recent proposals that were made to arrive at a satisfactory explanation of the sporadic nature of voicing in the Italian sibilants the one made by R.J. Di Pietro is noteworthy particularly because of the novelty of his approach. According to Di Pietro, the contrast between [s] and [z] can be solved in terms of a small number of phonological rules operating as matrices of co-occurring distinctive sound features. To this end he submits two alternative versions of a generative grammar model. In his first version the author formulates a general rule in which he posits intervocalic [s] as being always voiced. This on the grounds that morphemes with [z] are more numerous than those with [s]. In the alternate version he posits the Tuscan dialect as a pattern because "it is the most diversified of the general types". Accordingly, each morpheme containing an intervocalic /s/ is given the voiced feature. There are some serious difficulties with Di Pietro's proposed models, the choice of the Tuscan pronunciation being one of them. This regional dialect where words as cortese "curteous", francese "French", paese "country", quaresima "Lent" are pronounced with voiceless /s/ in Siena and with voiced /s/ in Florence does not seem quite appropriate for the formulation of phonological rules able "to furnish us with a simple way to incorporate regional variations of the standard language within the total grammar". More questionable is Di Pietro's choice of [z] as the underlying form of /s/, when it occurs in intervocalic position, on the grounds that "morphemes with [z] are more numerous than those with [s]". A choice of [z] as the more natural or as the unmarked member of the pair is in violation of the implicational law whereby a language does not have [z] in its sound inventory unless it also has [s] while the converse is not true. The information about the pronunciation of /s/ as either [s] or [z] in a morpheme that is not predictable on the basis of general rules is contained in the underlying representation, and the phoneme for the two sibilants is clearly [s] and not [z]. Witness Latin, Spanish and many Southern Italian dialects that all have [s] but not [z].
Given the impasse of finding a satisfactory explanation of the sporadic nature of voicing in the Italian sibilants within the taxonomic framework of S.I., I decided to repeat the attempt in the case of a regional dialect. For the experiment I selected the Venezia Giulia (V.G.) variant, in particular that of Trieste and Istria, and for the following reasons first, the V.G. phonetic system is quite close to that of S.I.; second, this is the dialect I am most familiar with; third, the experiment would give me an opportunity to test A. Liberman's [7] claim that the theory of complementary distribution (C.D.) is useless as a tool of discovery; fourth, should the experiment prove successful its findings might cast some light on the problems involved in S.I.

The V.G. phonetic system has a long and complex history which to this day has not yet been clearly understood. This situation is partly due to the fact that the V.G. vernacular literature is very modest and, furthermore it contains variant spellings for the same word; partly because, not infrequently, an original Latin phoneme appears to have shifted into two and even more different phonemes. Furthermore, the fact that phoneme /z/ does not exist in S.I. and that, in Italian spelling, its symbol is used to represent the two apicodental allophones [ts] and [dz] add to the complexity of the problem.
In the V.G. dialect [s] and [z] are present in C.D. as follows: The sibilant is voiced when followed by a voiced consonant, e.g. [zberla], schiaffo, "slap", and it is voiceless when followed by a voiceless consonant, e.g. [spaiar], spogliare, "to undress", and in word final position, e.g. [pʊs:], piscio, "piss". Before vowel, in word initial position and intervocalically, the two sibilants are in free alternation e.g. [se], se, "if"; [ze], è, "is"; [rʊzʊ], riso, "rice"; [lɛ·so], lesso, "boiled". How are we going to classify these contrasts? As partially allophonic? or as partially phonemic? or neither? To solve this problem let's analyse each environment in which the two sibilants are in contrastive distribution.

Word Initial Position

Contrary to S.I. prevocalic [z] does occur in word initial position in a significant number of words, e.g.:

	V.G.	S.I.	
1)	[za]	già	"already"
2)	[zaia]	moltitudine	"a large number of"
3)	[ze]	sei	"you (sing.) are"
4)	[zekʊn]	zecchino	"sequin"
5)	[zenero]	genero	"son-in-law"
6)	[zʊma]	grande freddo	"very cold"
7)	[zʊmeta]	freddolino	"cool air"
8)	[zʊŋgo]	zinco	"zinc"
9)	[zʊnɔčo]	ginocchio	"knee"
10)	[zʊzania]	zizzania	"darnel-grass", "dissension"
11)	[zʊzola]	freddo	"cold"
12)	[zo]	giú	"down"
13)	[zɔgar]	giocare	"to play"
14)	[zɔgatɔlo]	giocattolo	"toy"
15)	[zɔgo]	gioco	"play"
16)	[zɔnta]	giunta	"a part added"
17)	[zɔntar]	aggiuntare	"to unite by addition"
18)	[zɔventu]	gioventu	"youth"

We submit that, in V.G., prevocalic [z] in word initial is not an allophone of /s/ but a distinct, separate phoneme. Witness the following minimal pairs:

	V.G.		S.I.	
[zʊma]	:	[sʊma]	cima	"rope used on shipboard", "summit"
[zʊmeta]	:	[sʊmeta]	diminutive of [sʊma]	
[za]	:	[sa]	sa	"he/she knows"
[ze]	:	[se]	siete	"you (pl.) are"
[zo]	:	[so]	so	"I know"
[zʊzola]	:	[sʊzola]	giuggiola	"jujube"

Also historical grammar gives some evidence that the V.G. prevocalic [z] is a phoneme. Lexemes 6 and 7 are borrowings from Slovenian or Croatian. In both languages /z/, whose orthographic symbol is also ž, is not an allophone of /s/ but a distinct, separate phoneme. Lexeme 3 has been traditionally spelt xe in vernacular writings since as far back as the XII Century and possibly earlier. Its history is akin to French ce, both forms being derived from Latin ecce, that in Low Latin came to be used with the verb esse, "to be", to give prominence to a noun, an adverb and the like. As to the remaining lexemes the prevocalic [z] is clearly not an allophone of /s/ but the result either of a shift from the Latin or Italian voiced affricate /j/ e.g.:

Latin	S.I.	V.G.	
[iam]	[ja]	[za]	"already"
[iener]	[jenero]	[zenero]	"son-in-law"
[ienu]	[junɔk:io]	[zʊnɔčo]	"knee"
[iugam]	[ju]	[zo]	"down"
[iɔkas]	[jɔko]	[zɔgo]	"play"
[iuventas]	[jɔventu]	[zɔventu]	"youth"

or of a shift from Standard Italian /dz/, e.g. [dzʊŋko] > [zʊŋgo], [dzʊdzania] > [zʊzania], [dzek:ʊno] > [zekʊn].

Intervocalic Position

Save for the two notable exceptions that we shall examine in the next paragraph, we can say that, as a rule, in the V.G. dialect, the intervocalic /s/ is voiced e.g. [kaza], casa, "house"; [spoza], sposa, "bride". The sibilant is also voiced in words in which it is the result of a shift from Latin or S.I. /č/, /j/ and few other consonants, e.g. /k/ and /dz/:

Latin	S.I.	V.G.	
[fačʊmas]	[fač:amo]	[fazemo]	"we make"
[rationem]	[rajone]	[razon]	"reason"
[mediam]	[medz:o]	[mɛ·zo]	"middle"
[buk:a]	[buka]	[buzo]	"hole"

Before trying to establish the phonemic status of the intervocalic voiced sibilant it will be necessary to take into consideration the two exceptions we mentioned earlier. They are as follows:
1. Intervocalic [s] is voiceless in all words that in S.I. have a geminate in the correspondent position. The V.G. dialect has no long consonants consequently all S.I. geminates shift into short consonants in V.G., and the vowel preceding the shortened consonant shifts from short and open into tonic long or semi-long, e.g.:

S.I.	V.G.	
[pɔs:o]	[pɔ·so]	"I can"
[fɛlɪčɪs:ɪmo]	[fɛlɪčɪ·sɪmo]	"very happy"

2. The intervocalic [s] is voiceless in all words that in S.I. have phoneme /š/ in the correspondent position, e.g.:

S.I.	V.G.	
[lašare]	[lasar]	"to leave"
[bʊša]	[bʊsa]	"snake"
[lʊšo]	[lʊso]	"smooth"

We submit that in V.G. [s] and [z] in intervocalic position are two separate phonemes. Witness the following minimal pairs chosen at random, e.g.:

[bɑ·so]	:	[bɑ·zo]	[čɛ·sa]	:	[čɛ·za]
"low"		"kiss"	"cease"		"church"
[pɛ·se]	:	[pɛ·zɛ]	[spɛ·so]	:	[spɛ·zo]
"fish"		"scales"	"often"		"spent"

Historical grammar offers some evidence that, at one time, the two sibilants were treated as separate phonemes. From ancient texts [8] dating back to the XII and XIII century written in the old Venetian dialect (O.V.) it appears that the sibilants s and z were often graphically represented by two distinctive different symbols. For example, a long s is often represented by symbol s either short or long, e.g.:

S.I.	O.V.	
essa	esa or essa	"she"
possessione	posesione	"possession"
essere	esser	"to be"
rosso	roso	"red"

Symbol s is also used in O.V. to represent intervocalic [s] in all instances in which the voiceless phoneme is the result of a shift from S.I. /š/, e.g.:

S.I.		O.V.		
lascio	[lašio]	laso	[laso]	"I leave"
fascina	[fašɪna]	fasina	[fasɪna]	"faggot"

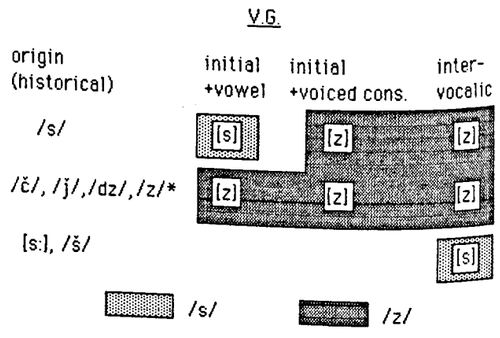
Conversely, the intervocalic voiced sibilant [z] is graphically represented by symbol x, e.g.:

S.I.		O.V.		
speso	[spɛzo]	spexo	[spɛzo]	"spent"
casetta	[kazɛt:a]	chaxella	[kazɛla]	"small house"

Symbol x is also used to represent the intervocalic [z] in words in which the voiced sibilant is the result of a shift from S.I. /č/ or /j/, e.g.:

S.I.		O.V.		
facciamo	[fač:amo]	faxemo	[fazɛmo]	"we make"
piace	[piače]	plaxe	[plazɛ]	"it pleases"
ragione	[rajɔnɛ]	raxon	[razɔn]	"reason"
cugina	[kujɪna]	choxina	[kɔzɪna]	"cousin"

To conclude, the implication of this presentation is that:
1. in the V.G. dialect the sibilants /s/ and /z/ are two distinct separate phonemes. Phoneme /s/ realizes itself as [s] or [z]. Both allophones are present in C.D. with [z] occurring before voiced consonants. Phoneme /z/ has no voiceless allophone. It occurs in word initial position before vowels and, in intervocalic position.
The following table summarizes the situation of the two sibilants as described above:



* From the Sloven z or O.V. x

2. Historical grammar suggests that the V.G. phoneme /z/ can be traced back to Latin x. This symbol, that in Classical Greek was used to transcribe the aspirate stop [kʰ] was borrowed by Latin scribae from Western Greek (Southern Italy and Sicily). Since x represented a sound not found in Latin, it gave trouble to the Romans in borrowed words. The nearest sound that the Romans had was the unaspirated [k] by which they accordingly at first represented it. Later on they used x to transcribe the free alternation s/z. This peculiar use of the symbol x is still evident in French in such alternations as soixante and sixième where intervocalic x is voiceless in the former and voiced in the latter. Also the V.G. dialect offers few examples of the free alternation s/z in intervocalic position that can be accounted for by positing, not unlike French, Latin x as the alternating sibilant, e.g.:

S.I.		V.G.		
faccio	[fač:o]	faxo	[faso]	"I do"
facciamo	[fač:amo]	faxemo	[fazɛmo]	"we do"

This situation also suggests that the confusion, in S.I., with regard to the two sibilant allophones [s] and [z] originated when Italian scribes substituted the Latin s for the Latin x. If our findings are correct then A. Liberman is wrong there where he maintains that complementary distribution is useless as a tool of discovery for "allophones can never be obtained before phonemes, and all attempts to reverse the sequence to obtain allophones in order to assemble them into phonemes is self-deception". It is by assembling the allophones [s] and [z] that we were able to prove the presence of the voiced sibilant phoneme /z/ in the Venezia Giulia dialect.

Bibliography

[1] L. Romeo, «Sibilants in Standard Italian: Facts and Fiction in Phonemic Analysis», IRAL, IV, 1, March, 1966, pp. 1-5.

[2] J. Arce, «Il numero dei fonemi in italiano in confronto con lo spagnolo», Lingua Nostra, 23, 1962, pp. 48-52.

[3] R. J. Di Pietro, «Phonemics, Generative Grammar and the Italian Sibilants», Studia Linguistica, V, 21, 1961, pp. 96-106.

[4] R. A. Hall, jr., «Italian [z] and the Converse of the Archiphoneme», Lingua, 9, 1960, pp. 194-197.

[5] C. Court, «On /s/ and /z/ in Standard Italian», Lingua, 18, 1967, pp.290-295.

[6] G. Porru, «Anmerkungen über die Phonologie des Italianischen», Travaux du Cercle Linguistique de Prague, VIII, 1939, pp. 187-208.

[7] A. Liberman, «On the Uses of Complementary Distribution», Proceedings of the Tenth International Congress of Phonetic Sciences, Edited by A. Cohen and M.P.R. Van den Broecke, Dordrecht, Holland: Foris Publications, 1984, pp. 647-649.

[8] Testi veneziani del Duecento e dei primi del Trecento, Edited by Alfredo Stussi, Pisa: Nistri-Lischi Editori, 1965.

For the inclusion of the table I am indebted to professor Gilbert Taggart of Concordia University.

THE SPECIFIC CHARACTER OF THE INITIAL SOUND-TYPE ABBREVIATIONS PHONOLOGICAL SYSTEM

DMITRENKO S.N.

Institute of Russian Language Moscow 121019 USSR

On the grounds of the basic system (BPS) of the Russian literary language, as outlined in (1), we endeavoured to find out whether system relations exist within the phonemes that are singled out in initial sound-type abbreviations (ISTA) as a particular lexical class of words, and if such relations exist, then, in what way do they differ from the relations observed in the BPS. The system of phonemes within the initial sound-type abbreviations are analyzed here not in isolation from, but in comparison with the BPS.

In the ISTA as the material shows, there is a peculiar phonological system of its own. M.Ya.Glovinskaya (2), substantiating the need to single out a subsystem of loan-words, defines three criteria which, in her opinion, may relate loan-words to a special phonological subsystem of the Modern Russian literary language:

1. The pronunciation of these words is such that it is impossible for the majority of words of the Russian literary language. Here "variation of the same phonemes according to positions is not the same in different words".

2. The pronunciation of the majority of loan-words is variable: "the same word is pronounced by different people in different ways and also... one and the same person pronounces one and the same word in a different manner. Thus the behaviour of the phonemes in these words is defined not by the determinate regularity as in the basic system, but statistically".

3. "In the basic system regularities in the phonemes' behaviour essentially coincide with the rules of pronunciation... The phonetic regularities of the subsystem define only the phonemes' behaviour and do not coincide with the orthoepic rules" (2). It would seem that these criteria which belong mainly to the pronouncing system of the Russian literary language are not sufficient for substantiating the singling out of similar groups of words into special phonological subsystem. First of all, it is necessary to define the characteristics of the basic phonological system of the majority of the words of the Russian literary language. The BPS is characterized by its own pecular contraposition of the same sound units to other ones in definite positions: accented and unaccented vowels-before hard and soft consonants, consonants before vowels and also before other consonants at the end of the word. The different possibilities for contraposing some sound units to others in definite positions allows to single out in the phonological system of the Russian literary language strong and weak phonemes. The BPS has a quite definite number of weak and strong phonemes which are distributed strictly according to their positions. The ISTA phonological system is not only characterized by a contraposition of some sound units to others in definite positions which differs from the BPS and, accordingly, by different possibility to quality differently some or other sound units in the same position.

The specific character of the ISTA phonological system consists not only in the different distribution of the phonemes (as a rule, the same phonemes which are represented in the BPS) according to their positions, but also in the possibility (or otherwise) of contraposing some or other phonemes to each other in a definite position. The latter is connected both with the presence (or absence) in the system of the phonemes of abbreviations of a special type, and with the real representation (or lack of representation) of some or other phonemes in a definite position preset by the material.

In the phonological system of a sound type's initial abbreviations the following regularaties may be noted which exist within consonant phonemes taken from the abbreviations of a given type and determined by system relationships:

1. The absence at the end of the word of a contraposition of consonants according to hardness-softness. In the ISTA in the end position there can be only hard consonants (this concerns both paired, according to voiceless-voiced features of noise sounds, and unpaired according to the same feature, sonoures sounds). For example: КЭАМ (концентрат эмульсии антраценового масла), ЛОИКФУН[X] (3) (Ленинградское общество исследователей культуры финно-угорских народностей), ОВИР (общая величина измерения расстояния до цели), ИМЭЛ[X] (Институт Маркса-Энгельса-Ленина при ЦК КПСС), РИП (радиоэлектронный измерительный прибор), РУВ (ручной установщик взрывателей etc.

Hard sonorous sounds in the ISTA are phonologically qualified as weak by the hardness-softness of the phoneme and hard noise sounds as weak by the hardness-softness and voiceless-voiced features (in the BPS weak sounds by two features in consonant phonemes cannot be at the end position of the word; here denote weak by voiceless-voiced consonant phonemes which are opposed to each other by hardness-softness: Ве/c_2/-ве/c_2'/, ра/T_2/, orthogr. рад-ра/T_2'/ etc.; the sonorous sounds in the BPS at the end of the word are opposed to each other by hardness-softness, thus both hard and soft consonants are possible: Во/Н/-Во/Н'/,МО/Л/-МО/Л'/,etc.;

2. The contraposition of consonants according to hardness-softness in a position before /e/. This position in the ISTA is strong: /с'е/ РП[X] (Социалистическая еврейская рабочая партия, дорев.) - /се/В (Совет Экономической Взаимопомощи). In old Russian words in the positions before /е/ there is no contraposition of consonants by hardness-softness. Only the utilization of barrowed words allows to broaden the position of consonants before /е/, including in it hard consonants. For example /пе/р, /ме/р;

3. The absence of consonant contraposition according to hardness-softness before /и/. In the ISTA phological system there is no combination of the"consonant (hard)+/и/" type (phonetically /ты/, /ны/ etc. (4); Here the position of consonants before /и/ is not strong, just as in the BPS, but weak (like the position of consonants before /е/ at the junction of the inflexion in the BPS, where is no contrapositioning of consonants according to hardness-softness; here, from the view point of phonetics, there is a soft con-

sonant);

4. The absence of the contraposition of consonants by hardness-softness in positions before consonants. These consonants are phonologically qualified either as weak according to the hardness-softness criterion, or as weak by two criteria (hardness-softness and voiceless-voiced): before /м / in the BPS both hard and soft consonants are represented: о/тм/ахнуться -/т'м/а, по/дм/анить - ве/дм/а, /см/отреть - ве/с'м/а, ра/зм/ахво/з'м/у, ко/рм/ - су/р'м/а, хо/лм/ - па/л'м/а, and in the ISTA in this position there are weak phonemes according to the hardness-softness criterion (phonetically hard consonants): ТМУП (Трест медицинских наглядных пособий), СМУ (сложные метеорологические условия), ЗМА (завод медицинской аппаратуры), ЛМОИ[X] (Ленинградский молочно-огородный институт), etc.; before /т/, /с'/, /к/ in the ISTA phonological system there is a weak consonant phoneme /$м_I$/ according to the hardness-softness criterion: МТИЛП (Московский технологический институт лесной промышленности), МСИ[X] (Московский санитарный институт), МКАД (Московская кольцевая автомобильная дорога); f.ex. in the BPS strong and weak consonant phonemes by the hardness-softness criterion are represented: ло/мт'/и - познако/м'т'/е, (о)ха/мс'/е - познако/м'с'/я, кро/мк/а - познако/м'к/а; before /т/,/ж/,/к/,/г/,/г'/,/б/,/$б_I$/,/$ш_2$/ in the phonological system represented is the weak consonant phoneme /$н_1$/ by hardness-softness: НТОВТ (Научно-техническое общество водного транспорта), МАНК (Международная ассоциация научного кино), НГИМИ (Новосибирский государственный институт мер и измерительных приборов), НБАД (Ночная бомбардировочная авиационная дивизия), ИНБИ (Ордена Ленина институт биохимии им. А.Н.Баха (АН СССР), РЯНШ (Русский язык в национальной школе /magazine/; in contrast to the BPS, where strong consonant phonemes by hardness-softness are represented: ко/нт/ора - ко/н'т/о, и/нж/ирко/н'ж/е, ма/нк/а-ко/н'к/а, ко/нг'/ениальный-де/н'г'/и, До/нб/асс-го/н'б/а, и/нб'/ирь- /о/ го/н'$б_I$/е, пу/$нш_2$/-ме/н'ш/е, before/т/, /к/, /$к_I$/, /м/, /$м_I$/ in the ISTA phonological system the weak consonant phoneme /$р_1$/ by hardness-softness is represented: АРТА[X](Артиллерийская радиотехническая академия), ВИРГ (Всесоюзный научно-исследовательский институт разведочной геофизики), РКИИГА (Рижский Краснознаменный институт инженеров гражданской авиации им. Ленинского комсомола), НИРММИ[X] (Всесоюзный научно-исследовательский институт растительных масел и маргарина); e.g. the BPS, before these consonants /р/ and /р'/ are opposed to each other:/рт/а-ко/р'т/о, ка/рг/а-се/р'г/а, ко/рк/а-Бо/р'к/а, ко/рк'/и-Бо/р'к/и, ко/рм/а-су/р'м/а, ко/рм'/ить-су/р'м/ить, before /п/, /$п_I$/, /б/, /в/, /$в_I$/, /$т_I$/, /д/, /$д_I$/, /с/, /з/, /$з_I$/, /к/, /г/, /м/, /$м_I$/, /н/, /$н_I$/ in the ISTA phonological system there is the weak consonant phoneme according to hardness-softness /$л_1$/ (from the view point of phonetics it is a hard consonant): СНИЛПЭ (студенческая научно-исследовательская лаборатория полупроводниковой электроники), УЛПИ (Ульяновский политехнический институт), ЛБАН (Львовская библиотека Академии наук УССР), ОЛЛВО (Ордена Ленина Ленинградский военный округ), ЛВИМУ (Ленинградское высшее инженерное морское училище им. адмирала С.О.Макарова), ЛТИХП (Ленинградский технологический институт холодильной промышленности), ЛДОК (лесопильно-деревообрабатывающий комбинат), ЛДИС (лазерный доплеровский измеритель скорости), ЛНИЛСЭ (Ленинградская научно-исследовательская лаборатория судебных экспертиз), УЛЗУ (универсальное логическое запоминающее устройство), ЛЗИИ[X] (Ленинградский заочный индустриальный институт), ЛКИП (лаборатория контрольно-измерительных приборов), ЛГОК[X](Лебединский горно-

обогатительный комбинат), ЛМОИ[X] (Ленинградский молочно-огородный институт), УКНИАЛМИ[X] (Украинский научно-исследовательский агролесомелиорационный институт - it is an abbreviation of the mixed type), ЦНИЛНОТ (Центральная республиканская научно-исследовательская лаборатория научной организации труда), ЛНИЯ[X] (Ленинградский научно-исследовательский институт языкознания), e.g. the BPS, where the contraposition /л/-/л'/ is possible: по/лш/арты-пу/л'п/а, ко/лб/а-па/л'б/а, мо/лв/а-ма/л'в/а, мо/$лв_I$/е-ма/л'в'/е, по/лт'/ина-/о/ ко/л'т/е, по/лд/ома-/л'д/ы, по/лд'/еревни-/л'д'/ина etc.: before the consonants /$т_1$/, /к/, /$к_1$/, /г/, /г'/, /б/, /$б_1$/, /с'/in the ISTA phonological system there are weak consonant phonemes by two criteria: /$п_3$/ before /$т_I$/, /к/, /$к_I$/, /$ф_3$/ before /г/, /г'/, /б/, /$б_I$/, /ф/: ОПТИ (отдел производственно-технической информации), ГУПП КА[X] (Главное управление политической пропаганды Красной Армии), ПКИАМ (проектно-конструкторский институт автоматизации и механизации); in the BPS before these consonants there are weak consonant phonemes by the voiceless-voiced criterion which are opposed according to hardness-softness: о/$п_2$т'/ереть-сн/п'$_2$т'/е, ла/$п_2$-к/а-сн/$п'_2$-к/а; ВТИЗ[X] (Всесоюзный трест строительно-технических изысканий - an abbreviation of mixed type), ВСЕГЕИ (Всесоюзный селекционно-генетический институт - an abbreviation of the mixed type), ВКАС (Военная Краснознаменная академия связи) represented in the BPS before these consonants are weak phonemes according to the voceless-voiced criterion which are opposed to each other as to hardness-softness: /$Ф_2$т'/ереть-сла/$Ф'_2$т/е, во/$Ф_2$с'/есла/$Ф'_2$с'/я, ло/$Ф_2$к/о-сла/$Ф'_2$-к/а; ДГЭС (Днепропетровская гидростанция), ВОДГЕО (Всесоюзный научно-исследовательский институт водоснабжения, канализации, гидротехнических сооружений и инженерной гидрогеологии - an abbreviation of the syllabic type), ДБАЭ (дальнебомбардировочная авиационная эскадрилья), ТБИЗ (Тбилисский инструментальный завод - an abbreviation of the mixed type), ТФАН[X] (Туркменский филиал АН СССР); in the BPS before these consonants, there are weak consonant phonemes as to the voiceless-voiced criterion, which are opposed to each other as to hardness-softness: о/$т_2$г/адать-ну/$т'_2$-г/а, о/$т_2$г'/ибать-ну/$т'_2г_I$/е, о/$т_2$б/авить-сва/$т'_2$б/а, о/$т_2$б'/ить-сва/$т'_2$б'/е, о/$т_2$ф/ормовать-/$т'_2$ф/у.

All the above, it would seem to us, convincingly points to the esistence of a phonological system of initial sound-type abbreviations, differing from the BPS.

(1). Русская грамматика: В 2-х т. М.: Наука. - 1980. - Т.I. Раздел "Фонология".

(2). Гловинская М.Я. Об одной фонологической подсистеме в современном русском языке // Развитие фонетики современного русского языка. - М.: Наука, - 1971. - С.55.

(3). Here and so forth[X] denote
"a) at present this organisation (institution etc.) does not exist,
b) this name is substituted for another,
c) at present this abbreviation is out of us"
Алексеев Д.И. Произношение сложно-сокращенных слов и буквенных аббревиатур // Вопросы культуры речи. - Вып.4. - 1963. - С.22-23.

(4) In the ISTA ГОМЗЫ (Государственные объединенные машиностроительные заводы) there is a composition /зн/ which is composed with the unitial sound and the inflexion of the word заводы.

ФОНЕТИЧЕСКАЯ ПРОГРАММА СЛОВА КАК ОСНОВНАЯ ПРОИЗНОСИТЕЛЬНАЯ ЕДИНИЦА

Л.Э. Калнынь

Институт славяноведения и балканистики АН СССР
Москва

Sound speech is naturally segmented according to phonetic word program, which is present in the conscience of speakers in the process of speech production and speech perception. Phonetic word program as a wholesale unit is organized according to interrelations, which should not be considered only as a combination of discrete elements of the sound/phoneme type.

Being set as a process extended in time, the phonetic word program is determined by the basis of articulation, by distaxia of linear bonds (interrelations), by communicative non-equivalence of certain word parts.

The problem is exsamined on the material of Russian phonetics.

I. Основой фонологического моделирования является представление о звуковой речи как о последовательности дискретных единиц типа звук/фонема. Путем операций с дискретными единицами устанавливаются правила их сочетания и противопоставления, делаются заключения об инвентаре фонем и их функциональных характеристиках. Логика непротиворечивой языковой модели предполагает, что она максимально способствует пониманию сущности интерпретируемого ею объекта. Однако знание об устройстве фонемной системы языка отнюдь не дает гарантии: I/ понимания звучащей речи, 2/ умения произнести иноязычное/инодиалектное слово, 3/ выявления причин устойчивости/изменчивости отдельных фрагментов языкового идиома.

Это заставляет предполагать, что в звуковой речи актуализируются правила и связи, которые не могут быть сведены к комбинации/нанизыванию дискретных единиц типа звук/фонема.

Умение делить слова на звуки у носителей языка обычно вырабатывается на основе ассоциаций с буквенной передачей слов при фонетическом типе письма /хотя и при этом может выделяться название буквы, а не собственно звук языка/; с отсутствием названного умения связаны и известные факты неправильной идентификации вырезанных из слова звуков. Более естественно деление слова на слоги - слогоделение может произвести каждый говорящий, руководствуясь при этом интуитивным ощущением правильности слогораздела [I]. Но, как представляется, натуральной и основной единицей членения звукового потока является фонетическое слово. Именно оно как фонетическое целое присутствует в сознании говорящих при порождении и восприятии речи.

Организация слова как целостной единицы регулируется специальным набором правил, которые могут быть интерпретированы в виде задаваемой в сознании говорящих произносительной программы. Рассмотрим эту проблему на материале русской литературной и диалектной фонетики.

2. Фонетическая программа слова определяется:
I/ артикуляционной базой идиома;
2/ правилами линейной организации звуковой последовательности;
3/ коммуникативной неравнозначностью разных частей слова.

Артикуляционная база, создавая общую звукообразующую установку, общую нюансировку речи, характеризуется не только комбинацией органов речи, но и артикуляционными переключениями, что осуществляется в пространстве более длинном, чем звук. В принципе можно считать, что полем этих переключений является именно слово, т.е. звуковая последовательность, ограниченная с обеих сторон потенциальной паузой.

Правила артикуляционной базы стабилизируют фонетику слова, компоненты которой связаны и не поддаются избирательной замене - артикуляционно перестроена может быть лишь вся звуковая цепь. Даже особенности, внешне как бы относящиеся к изолированному сегменту, в действительности определяют окружающую артикуляционную среду - так γará и gará различаются не только способом образования первого согласного, но и качеством первого гласного - после взрывного заднененебного он более высокого и напряженного образования, чем после фрикативного.

Звуковая последовательность, образующая фонетическое слово, организуется преимущественно по правилам дистактных связей, т. е. вне непосредственного артикуляционного контакта сегментов. Фонетическая программа слова задается как процесс, протяженный во времени. Связь между артикуляциями планируется до произнесения слова, а реализация плана начинается с инициальной артикуляции. Согласно этому первый звук приспосабливается к следующему еще до произнесения последнего.

Дистактность связи даже между рядом стоящими звуками подтверждается сохранением результатов регрессивного позиционного изменения вопреки исчезновению позиции, вызвавшей его. Ср. прерванное произношение, когда первый звук демонстрирует качество, полученное от только подразумеваемого, но не произнесенного звука (s'-/p'isát', op-/t'anút', диал. om-/mán, og-/mán); диал. упрощение конечных консонантных сочетаний с сохранением позиционной мягкости оставшегося согласного (kós't'i, kos', ẑýz'n'i, ẑys').

Дистактны изменения согласных в условиях внешнего сандхи - ср. звонкость согласного в предвидении инициального гласного или сонанта в следующем слове (т.е. d#V, Son). Именно дистактность связей в фонетическом слове отражает диалектное явление сохранения звонкости согласного перед глухим после нулевой редукции разделявшего их гласного [2] - гласный продолжает присутствовать в сознании говорящих как элемент, определяющий качество предшествующего согласного. На этом же основан известный факт синтагматической прозрачности v при уподоблении шумных по участию голоса (g vdov'e, но k vam) ; то же относится к l в диал. γl'd'ín'i , но k l'úd'im.

Собственно контактны лишь прогрессивные уподобления в звуковой цепи - в рус. диал. vás'k'a , но skáska , замена k > k' может произойти только при условии, что s' уже произнесен; то же t > č после ǯ в jéǯče при dájt'e; t>ć после ś в śćán'et при t'án'et.

Контактна ассимиляция гласных предшествующему согласному (ta, t''a) и дистактна - последующему (t'et' и t'ot , t'at и t'et').Характерно, что результаты дистактных связей в большей степени замечаются аудитивно, и они же проецируются на фонемный уровень.

Дистактность связей в рамках фонетической программы слова очевидна при выборе гласных на основе принципа межслоговой вокальной гармонии, что свойственно русской фонетике в большей мере, чем это принято считать. Регламентированность последовательного выбора гласных в слове лежит в основе тех типов безударного вокализма, которые не различают гласные неверхнего подъема. Ср. лит. аканье [CəCʌ/C'i], [C'əCʌ/C'i]: ə в трехсложном слове сигнализирует возможность следующих y, u, ʌ, но не a, o, ə после твердых согласных, i, u, но не e, a, o, ʌ - после мягких; гласный же ʌ предполагает следующий слог только под ударением; диссимилятивное аканье - [CəCá], где предударный гласный ориентирует только на a под ударением.

Подобие гласных разных слогов лежит в основе тех типов русского диалектного вокализма после мягких согласных, которые традиционно принято объяснять уподоблением гласного согласному в рамках сочетания VC. Это - ёканье и умеренное яканье. Из имеющихся интерпретаций этих вокальных типов наиболее близкой к существу явления кажется та, которую дал П.С.Кузнецов [3]. Выбор предударного непереднего (o, a) или переднего (e, i) гласного производен не от твердости-мягкости следующего согласного, а от вокального элемента следующего слога. Перед сегментом, имеющим завершение низкого тона, произносятся o, a ; при высоком тоне конца сегмента, что образуется и гласными ˙o, ˙a, ˙u, произносятся e, i. Эти отношения формируют модель слогового сингармонизма типа [o/a(C,CC)á, ó, ý, ú] [e/i(C', CC')í, é, ˙ó, ˙á, ˙ú].

Слоговой сингармонизм проявляется и в динамике диалектного вокализма - при освоении литературной нормы замена предударного о на а раньше всего происходит перед слогом с ударным а; предударный а после мягкого согласного в первую очередь устраняется перед ударным е, в последнюю - перед а [4].

Взаимозависимость качества гласных в слове накладвает известные ограничения на количество слогов в слове - принцип вокальной гармонии в сочетании с большой длиной слова создает произносительные неудобства. Наиболее распространенная длина русского слова до четырех-пяти слогов. Более длинные слова образуются с помощью ограниченного количества суффиксов или являются заимствованными по своему происхождению. Характерно, что именно в акающих говорах, где вокальная гармония очевидна, имеется тенденция к сокращению слогов в слове путем редукции до нуля безударных гласных.

Дистактные связи принципиально суперсегментны - даже, если полем их реализации являются рядом стоящие звуки. Но при этом в русской фонетике более важно качество второго сегмента и его воздействия на первый. Поэтому, как было показано, результаты регрессивного изменения могут быть изолированы от следующего сегмента /т.е. он может действовать как еще не реализованный элемент плана/, но устранение самого результата регрессивного изменения автоматически меняет следующий сегмент /при звонком сандхи произношение глухого согласного в конце слова влечет за собою глухость согласного в начале следующего слова/.

Анализ звуково* речи через призму фонетической программы слова приводит к выводу, что суперсегментные связи реализуются в русской фонетике шире, чем это принято

считать. Возможны случаи, когда явления, традиционно относимые к числу сегментных, обнаруживают суперсегментное содержание, если их рассматривать как компонент фонетической программы слова. Пример: при замене рус.диал. фрикативного взрывным g результат оглушения в x перед паузой сохраняется. В сегментном отношении устранение этой диалектной черты не связано с овладением новой артикуляцией или новой последовательностью сегментов. Оно связано с овладением новой фонетической программой согласно которой звонкий задненебный в предвидении паузы должен заменяться согласным k, а не x. Суперсегментная связь распространяется на серию ɣ - x - # или g - k - # .

Преобладание суперсегментных связей в фонетическом слове способствует его фонетической стабильности и выделению как целостной единицы в звуковом потоке. Можнр также отметить, что при конкуренции сегментной и суперсегментной мотивации явления сам факт его устойчивости может расцениваться как аргумент в пользу его суперсегментной обусловленности. Подчеркиваем, что имеем в виду факты, суперсегментное содержание которых не лежит на поверхности, в отличие, например, от таких явлений, как акцентный или интонационный контур слова.

Компоненты фонетической программы слова тесно взаимосвязаны. Иногда нельзя решить, обусловлена фонетика слова артикуляционной базой идиома или правилами линейной организации звуковой последовательности. Пример: рус.диал. **в>й/w** перед согласным и на конце слова, только твердость губных перед паузой могут быть результатом:
I/отсутствия навыков в переключении: губно-зубного сближения в консонантную артикуляцию и в паузу - это возможно лишь из билабиального сближения; палатализованности в губном ряду в паузу - это возможно только из артикуляции твердого согласного;
2/ произносительной программы, согласно которой **в** заменяется **й/w** мягкие губные твердыми в предвидении следующего согласного или паузы; отсутствие указанной звуковой мены для говорящего автоматически сигнализирует и иное качество следующего сегмента.

Видимо, актуальность обеих мотивировок делает эту черту русской диалектной фонетики очень устойчивой.

В естественной речи слово как фонетический процесс не может иметь на всем своем протяжении одинаковый уровень коммуникативной нагрузки - в одних частях слова он выше, в других - ниже. Это находит отражение и в артикуляционной специфике разных частей слова. Чем выше коммуникативная значимость сегмента, тем более старательно он произносится, а это, в частности, способствует реализации сложных, отчасти и неудобных артикуляций. В этой связи можно отметить, что в рус. слове коммуникативная нагрузка снижается от начала к концу слова. Этому соответствует развитие артикуляционного движения от более сильного /сложного/ к более слабому /простому/. Так, консонантное оформление начала слова сложнее, чем оформление конца - инициальных консонантных сочетаний разного типа больше, чем конечных.

Как известно, в русской фонетике к числу неудобных относятся инициальные сочетания сонорных с шумными (Son+t) и конечные сочетания шумных с сонорными (t+Son). Эти сочетания противоречат доминирующей тенденции к повышению звучности консонантной последовательности в начале слова и ее спаду в конце [5] . Тем не менее не только в литературном языке, но и в диалектах, чья фонетика свободна от давления графических ассоциаций, присутствуют инициальные сочетания типа (Son+t) В диалектах это отступление от принципа восходящей сонорности тем реже, чем длиннее консонантное сочетание. Так, нет трехчленных сочетаний, начинающихся на латеральный или вибрант, единичны, имеющие в своем начале назальный согласный; для четырехчленных сочетаний обязательно иметь в своем конце сонант. В то же время в тех же диалектах консонантное оформление конца слова не имеет отступлений от принципа понижения сонорности. Сочетания типа (t+Son) устраняются вставкой интерконсонантного гласного, утратой сонанта или приданием ему слоговости. В последнем случае сонант переводится в категорию сегментов, на которые не распространяются критерии построения консонантных сочетаний. Понижение сонорности в конце слова достигается и оглушением сонантов[6] .

Эти факты показывают, что начало слова, открывая реализацию фонетической программы, находится в сфере большего внимания со стороны говорящих, чем конец слова, где действует более определенно критерий произносительного удобства. В целом это отражает разную коммуникативную нагрузку начала и конца слова.

В таком же соотношении находятся предударная и заударная части слова в рамках его фонетической программы. Предударная часть слова коммуникативно более нагружена и поэтому находится в сфере большего внимания со стороны говорящих, чем заударная. Это может проявиться в более сильной редукции заударных гласных в сравнении с предударными в разговорной речи и в диалектах [7] . Приоритет предударной части слова просматривается и в характере тех изменений, которые происходят при утрате интервокального **j** - стяжение охватывает заударные гласные или ударный + заударный но не предударные [8] .

При перестройке безударного вокализма с целью устранения из него диалектных особенностей раньше и с большим успехом это происходит в предударной части слова. Например, такая рус. резко диалектная черта, как безударный о после мягких согласных, устраняясь или ослабляясь в предударной части слова, стойко держится в заударных слогах ((T.e.v'esnó || v'osnó, но только v'ét'or, úm'or). Этим же объясняется сохранение более архаической фонетики в послекорневых морфемах, находящихся преимущественно в заударном положении [9] .

3. Рассматривая слово как реализацию фонетической программы, мы подчеркиваем, что компонентами этой программы являются аудитивно воспринимаемые элементы, которые принято квалифицировать как сегментные единицы без учета их суперсегментной обусловленности. Собственно, обоснование фонетической программы слова необходимо именно на аудитивном уровне, так как инструментальным методом причинно-следственные связи в звуковой цепи, образующей слово, устанавливаются с достаточной степенью определенности [10] .

Интерпретация звуковой речи как серии фонетических программ слова не является лишь приемом научного анализа, но отражает объективную реальность. Это подтверждается, в частности, возможностью создания на основе этих программ фонетически правильных, хотя и лишенных смысла текстов. В то же время синтез речи как нанизывание актуальных для данного языка фонематических единиц без учета дистактных связей, характерных для фонетических программ слов, не только создает впечатление фонетически неестественной речи, но и может затруднить ее понимание.

Будучи основной произносительной единицей, фонетическая программа слова играет стабилизирующую роль в фонетике идиома. Как устройство более сложное, чем дискретная единица, программа устойчива по отношению к разнообразным воздействиям извне /например, в условиях языковых контактов/. При этом сохраняться может состояние структурно менее "удобное" /более сложное/ [11] .

Фонетическая программа слова функционирует как основной коммуникативный элемент - при порождении речи и при ее восприятии звуковая цепь членится на отрезки, в целом регламентированные фонетической программой. Это, в частности означает, что эффективная постановка произношения, его коррекция осуществляется только через слово, но не звук.

На фоне сказанного можно сделать вывод, что расчленение слова на дискретные единицы типа звук/фонема не имеет объективных оснований в фонетическом процессе, каковым является звучащая речь. Выделение этих единиц следует расценивать как операционный прием научного анализа, дающий возможность создать статическую модель процессуального по своей сущности объекта. Именно поэтому фонематические характеристики, с одной стороны, могут строиться на разных логических основаниях, а с другой, никогда не отражают всех, даже аудитивно воспринимаемых особенностей звучащей речи; не всегда могут быть однозначно определены границы дискретной единицы. Принципиальное различие между реальным строением звуковой речи и ее фонематическим моделированием надо в большей степени принимать во внимание при характеристики звукового строя языка.

Примечания.

1. Kalnin L. To the question of functioning of the Syllable. Paper presented to the Tenth International Congress of Phonetic Sciences (Utrecht 1983). М., 1983.
2. Пауфошима Р.Ф. О структуре слога в некоторых русских говорах. В кн. - Экспериментально-фонетические исследования в области русской диалектологии. М., 1977, с. 217; то же в кн. Русская разговорная речь М., 1973, с.60.
3. Кузнецов П.С. О гласных первого предударного слога в некоторых владимирских говорах. В кн. - Бюллетень диалектологического сектора Института русского языка АН СССР, в.4. М., 1948.
4. Войтович Н.Т. К вопросу о путях развития аканья в восточнославянских языках. В кн. - ОЛА. Материалы и исследования. М., 1974, с.33; Кириллова Т.В. Динамика диадектного вокализма в советский период. АДД. Калинин 1980.
5. Аванесов Р.И. Фонетика современного русского литературного языка. М., 1956, с.41.
6. Калнынь Л.Э., Масленникова Л.И. Опыт изучения слога в славянских диалектах. М., 1985.
7. Русская разговорная речь, с.46.
8. Русская диалектология. Под ред. Р.И. Аванесова и В.Г.Орловой. М., 1964, с.91.
9. Калнынь Л.Э. Об организации фонологических систем смешанного типа. ОЛА. Материалы и исследования. М., 1975, с.108.
10. Ср., например, Сорокин В.Н. Теория речеобразования. М., 1985.
11. Chloupek J. Aktualní otázky dialektologie. Jazykovědne Sympozium 1971. Brno, 1973.

THE STATISTICAL ANALYSIS OF INITIAL AND FINAL CONSONANT CLUSTERS IN ARMENIAN

ROBERT MARKOSOVICH TOKHMAKHYAN

Institute of Language, Abovian st. 15, Yerevan, Armenia
USSR, 375001

ABSTRACT

Though it is generally accepted that there are no initial clusters in Armenian the examination of the Modern Armenian word-stock has revealed a considerable number of initial and final clusters, met in 2676 and 135600words accordingly. The types of initial clusters are more limited as compared to final ones. The most frequent types are those with sonants r, y, and fricatives s, š.

It is accepted that there are no initial consonant clusters in Armenian and the use of the vowel "ə" in them prevents the formation of such clusters. This concept being true in general, is not to be applied to all the cases. According to our computational analysis of Modern Explanatory Dictionary of Armenian by E.Aghayan, 2676 out of 135600 words deviate from this rule. Some of those words belong to the native word stock and some of them are late borrowings. A number of those words reveal two pronunciation variants. One of them with the vowel "ə", the other without it. Academician Jahukian regards them as "instable models". Yet a considerable number of the above-mentioned 2676 words have a unified pronunciation that is without the vowel "ə". Therefore we can state that for some cases initial consonant clusters are to be found in Armenian too. Moreover if in those cases the pronunciation of words with the vowel "ə" occurs they should be considered substandard.

An initial double consonant clusters the second element as a rule is the sonant "y" or "r". Out of 30 consonants, 17 are used in combination with "y" (byur, gyuɣ, dyurin ťyur, lyard, xyus, cyurel, kyank', hyur, jyun, čyuɣ, myus, nyard, šyuɣ, syun, p'yunik, fyur'er). In the dictionary we can find 984 words with the initial consonant + "y" combination, the majority of which being combinations by, dy, jy, hy. Combinations with other consonants and "y" occur each in less than a dozen words. In combinations consonant "r" there may be 8 different consonants (brigadir, dramaturg xramat, krem, prizma, traktor, k'ristonya, franseren). The total number of such combinations is 444, a considerable number of which are late borrowings. Among them combinations "tr" and "pr" occur in 265 words. The total number of combinations in which the second component is expressed by "y" and "r" is 1428, that is more than half of the above-mentioned 2676 words.

Words with other combinations are not numerous. There we must mention consonant + the sonant "l" combinations (blank, klasic'izm, šleyf, planer, slavon, gladiator), 191 words all. As the first component of double consonant clusters as a rule fricatives "s", "z" and "š" occur more frequently. For example "s" + consonant combinations occur in 846 words (staр'vel, slavon, sxema, smoking, skuteɣ, syuže, spitak, svastika, stamoks, sc'enar, spɒp'el, Sfinks). Among them "sp" and "st" types occur in 646 words.

In "z" + C combinations the second component may consist only of [b], [g] and [ɱ] occlusives (zbaɣvel, zgušanal, zmaylel). There are 192 cases out of which 145 make "z" + "g" combinations.

The number of "š" + C combinations is 70 (špar, švayt, štab), and there is only one case of "š" + C (šč^{c}i).

Of the words with initial consonant clusters to be found in the dictionary, 30 have as the first component the consonant "k'" [k] (k'vant, k'sanhing). There are also a few borrowings having unusual initial clusters (tvist, psixologizm, pfenig).

In Armenian the number of consonant clusters consisting of three component is relatively few. The third component of such clusters is often "y". However, their number is not great (only 28). More numerous are the words with initial "stv" and "str" clusters. All in all 120 words.

Thus we may state that 2 per cent of the total number of words in the dictionary contains initial consonant clusters which exclude the existence of vowel "ə" in Modern Armenian literary pronunciation.

Final consonant clusters occur in considerably more words. In double consonant clusters all the consonants may be used as the second component with the exception of "y" and "ɾ", whereas the first component may be expressed with any consonant but [d], [ǯ], and [j].

The analysis reveal 13672 words with final consonant clusters. In more than half of those cases the first component is expressed either by sonants [r] and [y].

[r] is followed with any consonant (borb, bard, marg, hard, parz, kirt', haverž, marx, gorc, hark, šnorh, perč, t'arm, noktyurn, gorc, p'arč, kerp, lurǰ, vars, nerv, gort, harc', arp', jirk^{c}, torf) with the exception of 5 consonant l, ɣ, y, r, ř. The total number of such clusters is 4285 where cases with rd, rt, rǰ, rk, rz, rc, rk', rt', rp, make the overhelming majority.

[y] is followed with any consonant (ayb, ayg, ayd, huyz, xayt', huyz, ayl, šeyx, ayc, makuyk, anpačuyč, aboym, jayn, žptaňyš, žayr, ays, ayt, hayr, c'uyc', kahuyk', dreyf), with the exception of consonants h, j, ɣ, y, v, p'. The total number of such clusters being 3056.

"n" + consonant combinations make almost half of the number of such clusters, the second component of which may be expressed by any forelingual, backlingual occlusives, africates and the fricatives s, z, š (žang, hand, bronz, kant', p'arapanc, sunk, ganj, čanč, řevanš, šunč', jinǰ, avans, kent, c'anc', hank'). The total number of such words is 2694.

"n" is not followed by the labials "b", "p", "p'". In "m" + consonant combinations, the number of which being 312 (xumb, saṁt', himn, zamš, amp, toms, ek'spromt, dramp', himk', triumf) the two third of which represent "m+b" clusters.

"ɣ" + consonant combinations make 555 cases (kuɣb, heɣg, naɣd, t'uɣt', p'iɣc, p'eɣk, iɣj, kaɣč, koɣm, zeɣč', koɣp, gaɣǰ, uɣt, k'aɣč, čeɣk^{c}).

In double consonant clusters the first component is often expressed by the fricatives s, z, š. s + consonant make 812 words [xaris, hask, parisp, hast, xosk']. [s + t] out of which is the most frequent (673 cases).

"z" + consonant make 383 words (skizb, azg, azd, kazm, vazk') [z + m] of which is the most frequent (335), "š" + consonant make 384 words (vašx, mašk, drošm, hašt, hrašk') [š+t] out of which is the most frequent (264).

The rest of the above-mentioned 13672 words that is to say 1180 words may be classified according to the second component. Here we usually observe voiceless k', k, t. Consonant + k' combinations make 704 words, among them more frequently occur ck' [xosvack'] - 312 words, c'k' [gnac'k'] - 96 words, řk' [kařk'] - 86 words, tk' [hetk'] - 77 words. The characteristic feature of final double consonant clusters in Armenian, as a rule, is their voiceless pronunciation. Out of the 13672 words only 2714 have distinctly voiced pronunciation. The rest (8258 words) have voiceless pronunciation. 2700 words may have double pronunciation (voiced or voiceless).

In Armenian there are only 20 words with 3 consonant clusters. 18 of them (anc'k', kurck', hamaynk', c'aytk', rarbk', partk') as the third component have [k'], -nc'k' cluster makes 12 words, two borrowed words [punkt, boršč'] have different components. As you can see the third component is always a voiceless consonant.

Thus we may conclude that in literary Eastern Armenian there exist mainly double consonant clusters either in initial or final position and the number of 3 consonant clusters are comparatively few. We have included borrowed words in our survey and we may state that we observe a tendency of increase of the number of such words. The without [ə] utterance of consonant clusters in such words prevails especially in the pronunciation of the young generation.

CONSONANT GEMINATION IN ENGLISH LOANWORDS OF JAPANESE

JOHN H. KOO

Dept. of Ling. & For. Languages
University of Alaska
Fairbanks, Alaska 99775 U.S.A.

YAYOI HOMMA

Dept. of English
Baika Womens College
Ibaraki-shi, Osaka 567 JAPAN

ABSTRACT

The study is primarily concerned with the consonant gemination (CG) in English loanwords of Japanese. It examines the phonetic conditions under which CG takes place. The experiment reveals that the younger generation is more sensitive to the morpheme condition that prohibits geminating a voiced stop, and that both the younger generation and the older generation today tend to avoid gemination.

Historically, Japanese has been very receptive to loanwords. It would be difficult to find any other language in the world which has been as hospitable to loanwords as has Japanese (Miller, 1967: 236). We find as many as 25,000 loanwords in Arakawa's dictionary (1961) and 34,000 in Iida and Yamamoto's dictionary (1983)[1]. In 1964 the Japanese National Language Research Institute announced that close to one half of the contemporary Japanese words were of Chinese origin, and the words borrowed from other languages, mostly European, amounted to 10%, the greater portion of which came from English. The native Japanese words constituted only 37%. This situation often makes language purists frown.

Most loanwords were first introduced in the Japanese language and reproduced by the borrowing speakers who probably did not have a native control of the source language (SL). The same English words are thus borrowed very frequently with different pronunciations. According to Higa (1979), such words as 'cotton' and '(sewing) machine' have been borrowed by the Japanese dressmakers' circle as katan and mashin[2]. The following are some examples that have more than one variant pronunciation: chekko/cheko 'Czech', firippin/firipin 'Philippines', chiraa/teraa 'tiller', hatsu/haato 'heart'. Moreover, some loanwords such as runessansu 'Renaissance' and sutaffu 'staff' seem to be via spelling, i.e., due to spelling.

The word borrowing is not totally random or unsystematic. The environments in which consonant gemination (CG), for example, occurs are not altogether unpredictable, because the borrowed words (BW) are reshaped to conform to the phonological constraints of the borrowing language (BL)[3]. More precisely, when a source word (SW) contains a segment in the environment which is alien to BL sound pattern, it is modified in conformity with the sound pattern. The reshaping, or the change, which arises due to the two different phonological systems, frequently results from the addition or substitution of sounds, e.g., setto 'set', roosu 'roast'. In Japanese setto, which is from English 'set', we observe addition of the obstruent t (i.e., consonant gemination) and the vowel o (i.e., epenthetic vowel insertion). In some cases, BW's become reshaped to such an extent that they are unrecognizable to the source language (SL) speakers. The following are only a few such examples: katto 'cut', naasu 'nurse', tsuaa 'tour'.

In the present study, we are concerned only with consonant gemination (CG) in English loanwords in Japanese. The study attempts to provide some of the phonological factors that are attributable to the gemination in the loanwords.

There have been various efforts to explain CG of English loanwords in Japanese. Among the most serious studies are Kunihiro (1963); Ohso (1971); and Lovins (1973). They have, however, sufficiently demonstrated the need for more rigorous research in this area. We are thus endeavoring here to supplement some new additional information and reinterpret some of their findings. The following are some of the factors that are responsible for CG:

(1) The Vowel Length Preceding the Word Final Consonant. The final stop or affricate of SW's is systematically geminated if the preceding segment is a short vowel. As the result of this, the syllable so added to increases in its prominance. Consider the data below[4]:

poketto 'pocket'[5], kasetto 'cassette', sunappu 'snap', ṡiroppu 'syrup', panfuretto 'pamphlet', soppu 'shop', baggu 'bag', guddo 'good', katorikku 'Catholic', eggu 'egg', biggu 'big', maccꞑi 'match', suiccꞑi 'switch', kareʒʒi 'college', ankareʒʒi 'Anchorage'

One reason for CG may be due to the durational difference in vowels between the two languages. As also noted by Miller (1967: 255), the borrowed English vowels are heard by the Japanese as markedly short as compared to their corresponding vowels in Japanese [6]. CG is thus used to compensate the difference, so that the prominence of the source vowel may be highlighted and the balance between the two syllables may be sustained. Notice that the sequence -VCCu# in BW's, in which the low pitched u is devoiced following a voiceless consonant, becomes much closer in pronunciation to its corresponding sequence in SW's.

The obstruent that follows a long vowel or diphthong is thus not subject to CG: kooto 'coat', meekaa[7] 'maker', keeki 'cake', paato 'part', seebu 'save', sukoopu 'scope', puropoosu 'propose'. This is due to the unpermitted sequence VVCC in Japanese. Only the consonant clusters allowed in the phonological system of Japanese are the clusters of two identical voiceless obstruents pp, tt, kk, ss, and the mora nasal cluster (e.g., kondo [kǫndo] 'this time', konban [kǫmban] 'this evening', kongetsu [kǫngetsu] 'this month', konsei [kǫnsei] 'mixture')[8]. In this case the first segment of the geminate acquires one mora's duration.

(2) Morpheme Boundary. In the previous studies, the morpheme boundary in a multi-morpheme word has not been rigorously discussed. Rather the phonetic syllable boundary [9] was used as being partly responsible for CG (Ohso, 1971; Lovins, 1973). In our study, it will be shown that the morpheme boundary bears more relevance to CG. We claim that the source word (SW) 'butter', for example, which is pronounced as bataa with the medial consonant ungeminated, is attributable to the absence of the morpheme boundary, whereas SW 'batter', which is pronounced as battaa with the medial consonant geminated, is due to the presence of the boundary. Consider the following data:

basukettobooru (basket-ball), nekkuresu (neck-lace), hettoraito/heddoraito (head-light), fureʒʒaa (fresh-er), suiccꞑingu (switch-ing), fiʒʒingu (fish-ing), ataccꞑimento (attach-ment), kurokkuwaizu (clock-wise), bukkiṡu (book-ish), ṡoppingu (shop-ing), adomittansu (admit-ance), bukkusu (book-s), ċippusu (chip-s), koohii dorippaa (coffee drip-er)

We note above that the morpheme final obstruent is systematically geminated as is the word final obstruent, whether the following morpheme is a substantive or a derivational affix. How about BW nekutai 'necktie' then? It is not exceptional to the rule. Consider the loanwords in which the same morpheme 'neck' consistently undergoes CG at the boundary: nekkupiisu 'neck-piece', nekkurain 'neck-line', nekkubando 'neck-band'. We interpret the degemination as due to the fact that the word was perceived as one morpheme word like such words as kakutasu 'cactus', nekutarin 'nectarine', and napukin 'napkin'.

(3) Stops before s in Word Final Position. We observe that k of the cluster ks#, which is represented by x in some SW's, and p of the cluster ps# are regularly subject to CG as is seen in the examples:

rirakkusu 'relax', bokkusu 'box', sekkusu 'sex', mikkusu 'mix', apendikkusu 'appendix', takkusu 'tax', sakkusu 'sax', pahappusu 'perhaps'

The gemination seems to occur only in word final position. Thus 'boxing' and 'saxophone' are vocalized as bokuṡingu and sakisohon. More examples are: mikuṡingu/mikiṡingu 'mixing', makiṡimamu 'maximum', takuṡii 'taxi'.

Notice that the t of the sequence ts# never undergoes CG in that the cluster of the homoganic ts, which is permitted in the Japanese phonology, automatically becomes tsu word finally.

(4) Stress in Stem Final Syllable. The stress on the stem final syllable in an SW is also pertinent to CG. Thus 'editing' and 'classical', in which the stress does not fall on the vowel immediately preceding the consonant (i.e., éditing and clássical), are not subject to CG; they are vocalized as editingu and kurasikaru. In such SW's as kotteeji/koteeji 'cottage', offaa/ofaa 'offer', and batteri 'battery', the gemination of the obstruent may be attributable to a false impression of the sequences like -age, -er, and -ery as the derivational morphemes, or to spelling influence. Notice the SW's that end in -tion are exceptional to the above rule: adiṡon 'addition'; kondiṡon 'condition'; ediṡon 'edition'.

(5) Palatal Segment. Another factor that has not been well studied in the literature is the relevance of the palatal segment to CG. As will be seen below, the word or morpheme final ṡ regularly undergoes CG, while the non-palatal s does not.

raʒʒu 'rush', fiʒʒu 'fish', raʒʒingu 'rushing', daniʒʒu/danisu "Danish', inguriʒʒu/ingurisu 'English', buraʒʒu/buraṡi 'brush', kisu/kissu [10] 'kiss', kurosu 'cross', hosutesu 'hostess'

The variant danisu, for instance, has not undergone CG due to the non-palatal s. It is noted here that the palatal ṡ of buraṡi is underlyingly /s/, which is automatically palatalized before the palatal i by assimilation in Japanese. In fact, there are very few Japanese words ending in -ssu. The only word that we know of which ends in -ssu is hissu 'indispensible'. There is, however, a small set of words ending in -ʒʒu (e.g., iʒʒu 'a kind of', zaʒʒu 'a mixed breed', daʒʒu 'capture').

One possible explanation for ṡ-gemination may be that the palatal ṡ is, acoustically, of somewhat lower frequency (approx. 25000 Hz) [11] and longer duration as compared with the non-palatal s (approx. 3000 Hz), and is more resonant.

We note, however, that if an s-ending SW is followed by a derivational morpheme, then the stem final s is subject to CG:

doresu 'dress', doreṡi 'dressy', doreṡṡingu 'dressing'; gurasu 'glass', guraṡṡi 'glassy'; kompuresu 'compress', kompuressaa 'compressor'; pasu 'pass', paṡṡingu 'passing'

(6) Obstruent before (ə)n#. The obstruent preceding (ə)n# (i.e., syllabic nasal) is also responsible for CG: kotton 'cotton', fa$$on 'fashion', mi$$on 'mission', le$$un 'lesson'.

(7) Obstruent before Syllabic Lateral. The obstruent that precedes the lateral in word final position or at the morpheme boundary is geminated: bakkuru 'buckle', kyassuru 'castle', kappuringu 'coupling', bakkuringu 'buckling', takkuru 'tackle', appuru 'apple', kappuru 'couple', nikkeru 'nickle', pikkurusu 'picles', hassuru 'hustle'. It is interesting to note that if the first member of the l-cluster is an alveolar top, then CG is blocked. The gemination here seems to occur only when the consonant preceding the lateral is non-coronal. Consider the following data:

ritoru 'little', botoru 'bottle', padoru 'paddle', ratoru 'rattle', $atoru 'shuttle', sadoru 'saddle', ketoru/kettoru 'kettle'

The occurrence of CG with -C1# is further constrained to such an extent that when the vowel before -C1# is preceded by a consonant cluster or unstressed (thus forming a multi-syllable word), CG is also blocked. This is probably due to the stress shift that was caused by the added syllable(s) as seen in the examples below. For instance, in BW toripuru 'triple', the stress has shifted to the preceding syllable (i.e., from ri to to). Notice here that the initial C-cluster, which is not permitted in Japanese morphology, is broken by the paragogic vowel o.

kuripuru 'cripple', puri$ipuru 'principle', marutipuru 'multiple', aatikuru 'article'

We note, however, that in a word like 'pineapple', the p before the lateral is subject to CG in that the word is of two morphemes, thus not affected by the above rule.

Final Obstruent Tend to Be Unvoiced and Degeminated. CG is less regular and less frequent in its occurrence if the consonant is voiced; it fluctuates between the voiced doublet and the voiceless doublet, apparently to conform to the morpheme structure condition in Japanese that does not allow a voiced consonant geminate, as evidenced in the following examples that were enunciated by young Japanese college students:[12]

betto[13]/beddo 'bed', bikku/biggu 'big', hetto/heddo 'head', retto/reddo 'red', ekku/eggu 'egg', bakku/baggu 'bag', detto/deddo 'dead', patto/paddo 'pad', mittonaito/middonaito 'midnight', e$$i/e$$i 'edge', kitto/kiddo 'kid'

The gemination of the voiced labial b occurs only rarely; its occurrence is much more infrequent than that of the other voiced stops d and g. One of the reasons may be that the closure duration of the labial stop is longer than that of the alveolar stop or the velar stop [14], and that the vowel duration is shorter before labials than before alveolars or velars (Peterson and Lehiste: 1960; Lehiste: 1970). The inverse relationship between vowel duration and closure duration of the following stop appears to be maintained[15]. Hence the need for gemination of b is felt less in this context. See some data below:

bobu "Bob', pabu 'pub', hobu 'hub', jobu 'job', nobu/nobbu 'knob', mobbu 'mob', sunobbu 'snob', sukyabbu 'scab'

Ohso (1971) is correct in saying that there is a tendency for the younger generation to avoid geminating a voiced stop, especially the labial stop. This claim was well supported by our experiment [16]. The following are the results:

I. Voiced vs. Voiceless

	Younger Gen.		Older Gen.	
	Vd-CG	V1-CG	Vd-CG	V1-CG
'handbag'	51%	49%	77%	23%
'bed'	60%	40%	83%	17%
'hotdog'	59%	41%	94%	6%

The above list indicates that the younger generation is more sensitive to the morpheme condition rule in Japanese that prohibits the gemination of a voiced stop.

II. Geminated vs. Degeminated.

	Younger Gen.		Older Gen.	
	Gem.	Degem.	Gem.	Degem.
'Philippines'	7%	93%	36%	64%
'pub'	2%	98%	1%	99%
'plug'	6%	94%	3%	97%

Our experiment also shows that there is a tendency that both the younger generation and the older generation avoid gemination. The tendency is stronger with the labial stop.

CONCLUSION

We have discussed some of the factors that cause CG in the realization of English loanwords in Japanese. The study has shown that CG is closely related to such factors as intrinsic segmental duration, durational disparity between SL and BL, stress, morphological boundary, and syllable structure; these are all interrelated intricately. Moreover, the problem of weeding out spelling influenced pronunciation from phonologically conditioned pronunciation additionally reveals the intricate nature of CG.

We admit that the principles provided are not exhaustive and that some of the supporting phonetic evidences are not quite convincing. There may be some exceptions to the rules suggested as there are to most any rule. However, the generalizations made above should hold for the majority of the CG-related loanwords. The occurrence of CG is constrained in such a complicated manner that it is probably not possible to predict the exact contexts in which CG takes place. Further study is still needed in this area.

FOOTNOTES

[1] This does not include loanwords from Chinese. The number of English loanwords has greatly increased since 1945.

[2] These examples are from Higa (1979).

[3] See Kaye and Nykiel (1979).

[4] A large portion of the data listed here was collected from ten native Japanese speakers, who were randomly selected from various categories of employment (including college students and elementary school teachers), and whose ages ranged from 21 to 45. Some data come from numerous secondary sources such as Japanese dictionaries and magazines. We are grateful to Miss Tomomi Hasegawa of Gifu University for her assistance in collecting some valuable data for the present research.

[5] The word final o is a paragogic vowel added to the loanword ending in an alveolar stop. It is to be noted that Japanese words end, without exception, in a vowel or the nasal n, which is the only consonant allowed in word final position (e.g., pan 'bread', kantan 'simple').

[6] Note that the stressed English vowels are slightly longer than unstressed vowels (Klatt: 1976), whereas the Japanese vowel duration is independent of pitch stress (Homma: 1985).

[7] The sequences er and ar in SW's change to aa before a stop or word finally. Furthermore, the diphthongs are flattened to long vowels (e.g., ei to ee, ow to oo).

[8] Japanese is a mora-counting language. In Japanese, a mora functions as the unit of the phonological distance. For further details, see McCawley (1968: 131-4). Notice here that the basic syllable pattern of Japanese is an open syllable.

[9] It is noted that syllabification in English is not uniform, which Ohso (1970) also notes.

[10] In Japanese, CG is sometimes used as a device to make the word expressive (or emphatic). For example, kisu 'kiss' is also emphatically said kissu. See more examples: sugoku (non-emphatic): suggoku (emphatic) 'extremely'; totemo (non-emphatic): tottemo (emphatic) 'very very'.

[11] See Catford (1977). This information is also based on our spectrographic measurements.

[12] Iida and Yamamoto's dictionary (193) lists all the words in this section with the voiced geminates.

[13] Miller (1967: 243-4) claims that the origin of betto is from German Bett. It does not matter here whether the word is of German origin or not.

[14] Homma (1981), Shimizu (1985), and Stathopoulos and Weismer (1983) report that voiced stops are shorter in their closure duration than their voiceless counterparts.

[15] This relationship was also noted in Shimizu (1985).

[16] In the present project, 200 native Japanese speakers participated. Out of the 200 speakers, 100 were English major students in college, who were 20 years old, and the other 100 were mostly college professors, whose ages were in the range of 40 to 65.

REFERENCES

Arakawa, Soohei. 1967. Gairaigo Jiten (Loanword Dictionary). Kadokawa-shoten. Tokyo.

Catford, J.C. 1977. Fundamental Problems in Phonetics. Indiana University Press.

Higa, Masanori. 1979. Sociolinguistic Aspects of Word-Borrowing. Socio-linguistic Studies in Language Contact: Methods and Cases. William F. Mackey and Jacob Omstein (eds.). Mouton Publishers.

Homma, Yayoi. 1981. Durational relationship between Japanese stops & vowels. Journal of Phonetics.

_____. 1985. Accoustic Phonetics in English and Japanese. Kyoto: Yamaguchi Publication House.

Iida, Takaaki and Keiichi Yamamoto. 1983. Gairaigo Jiten (Loanword Dictionary). Tokyo: Shueisha.

Kaye, Jonathan and Barbara Nykiel. 1979. Loanwords and abstract phonotactic constraints. Canadian Journal of Lingluistics. 1.1-48.

Klatt, D.H. 1976. Linguistic uses of segmental duration in English: acoustic and perceptional evidence. Jouranl of the Acoustic Society of America 59, 1208-1221.

Kunihiro, Tetsuya. 1963. Gairaigo hyooki ni tsuite: nichi-ei on'intaikei hikaku. Nichi-Ei Ryoogo no Hikaku Kenkyuu Jissen Kiroku, pp. 27-48. Tokyo: Taishuukan.

Lehiste, Ilse. 1970. Suprasegmentals. MIT Press.

Lovins, Julie B. 1973. Loanwords and the Phonological Structure of Japanese. Ph.D. dissertation. University of Chicago.

McCawley, James 1968. The Phonological Component of a Grammar of Japanese. Mouton & Co.

Miller, Roy Andrew. 1967. The Japanese Language. University of Chicago Press.

Ohso, Mieko. 1971. A Phonological Study of Some English Loanwords in Japanese. M.A. Thesis. Ohio State University.

Peterson, G. E. and I. Lehiste. 1960. Duration of Syllable Nuclei in English. Journal of Acoustic Society of America. 32 (693-703).

Shimizu, Katsumasa. 1985. A Study on Vowel Duration in English. Nagoya Gakuin University Round Table on Language, Linguistics, and Literature 13.

Stathopoulos, E.T. and G. Weismer. 1983. Closure duration of stop consonants. Journal of Phonetics 11 (395-400).

КВАНТИТАТИВНОСТЬ КУМАНДИНСКИХ ГЛАСНЫХ

И.Я.СЕЛЮТИНА

Лаборатория экспериментально-фонетических исследований Института истории, филологии и философии СО АН СССР Новосибирск,СССР, 630090

ABSTRACT

The aim of the paper is to show the correlation of the vowel duration in Qumandy dialect of the Altai language with the acoustic and articulation characteristics, syllable structure of the word, quality of the pre- and post-vocalic consonants, type of the vocalic axis. The lengthening of the wide etymologically short vowel in open syllable before the syllable with narrow vowel till the duration of the contractive vowels is interpreted not as the reconstruction of the proto-duration but as the positional lengthening of phonemic character. This tendency is typologically common for the Qypchak language and can indicate the presence in Qumandy ethnical group of the Qypchak-turkic components.

ВВЕДЕНИЕ

Кумандинцы - одна из северных групп алтайцев, в этногенезе которых по данным антропологии /I/, этнографии /2/ и лингвистики /3/ наряду с тюркскими отмечаются и угро-самодийские этнические компоненты.

Предварительный слуховой анализ языкового материала на доинструментальном этапе исследования позволил выделить в языке кумандинцев I5 гласных фонем - классическую для тюркских языков восьмёрку кратких и семь долгих фонем, попарно противопоставленных по артикуляционным признакам ряда, подъёма, участия губ (табл. I).

Таблица I

Фонемы	Неогубленные широкие		Неогубленные узкие		Огубленные широкие		Огубленные узкие	
Заднерядные	a	a:	ъ	–	o	o:	ʊ	ʊ:
Переднерядн.	ɛ	ɛ:	i	i:	ø	ø:	y	y:

Правомерность фонологической трактовки квантитета гласных подтверждается наличием в языке кумандинцев квазиомонимов, семантическая дифференциация которых базируется лишь на противопоставлении кратких и долгих гласных: tar 'порох' - ta:r 'мешок'; er 'мужчина' - e:r 'седло'; kis 'режь' - ki:s 'кошма'; qol 'рука' - qo:l 'дыра'; kʊ̀ʰç 'кочуй' - kʊ́:ʰç 'крынка'; ʰçʊn 'вытирай' - ʰçʊ:n 'собрание'; tyr 'наматывай' - ty:r 'бубен' /4/.

В данной статье кумандинские гласные анализируются в аспекте их квантитативности. Цель исследования: I) определить на экспериментально-фонетическом материале, полученном методом пневмоосциллографирования /5/, ингерентную длительность гласных различного качества; 2) выявить особенности проявления потенциально существующей зависимости длительности гласных от фонетических условий - структурного типа слога и качественно-количественных параметров пре- и постпозитивных консонантов; 3) проверить квантитативные характеристики гласных на фонематическую признаковость; 4) проанализировать особенности реализации длительности гласных в корреляции с качеством вокальной оси словоформы.

ИНГЕРЕНТНАЯ ДЛИТЕЛЬНОСТЬ ГЛАСНЫХ

Сопоставление средних относительных длительностей кумандинских гласных /6/ в типологически сходных позиционно-комбинаторных условиях свидетельствует о существовании корреляции между типом гласного и его длительностью. Среди кратких фонем самой длительной является широкая переднерядная огубленная фонема [ø] , самой короткой - узкая заднерядная неогубленная фонема [ъ] ; в обобщённом виде распределение кратких гласных по их убывающей квантитативности можно представить так:

[ø] > [o] > [ɛ] > [a] >
II7,7 II2,6 I02,4 97,8
> [i] > [y] > [ʊ] > [ъ]
92,6 9I,9 77,9 76,2

В подсистеме долгих гласных фонем самыми длительными репрезентантами характеризуется широкая переднерядная неогубленная фонема [ɛ:], самыми краткими - узкая переднерядная неогубленная фонема [i:]:

[ɛ:] > [o:] > [ø:] > [a:] =
I82,5 I77,8 I72,4
= [a:] > [ʊ:] > [y:] > [i:]
I72,4 I67,8 I59,4 I46,6

Пограничная зона квантитативности между самой длительной из кратких фонем ([ø]) и самой короткой из долгих ([i:]) составляет 28,9% средней длительности звука (СДЗ).

Широкие гласные регулярно, во всех типах кумандинских слоговых структур, длительнее соответствующих узких коррелятов. Наиболее отчётливо эта закономерность проявляется в подгруппе огубленных фонем: широкая заднерядная огубленная фонема [o] длительнее своего узкого коррелята [ʊ] в среднем на 34,7%, широкая переднерядная огубленная фонема [ø] > узкого коррелята на 25,8% СДЗ. Слабее отмеченная тенденция реализуется в подсистеме долгих гласных фонем.

Для кратких гласных отмечена тенденция к доминированию длительности оттенков переднерядных фонем над заднерядными коррелятами. С абсолютной регулярностью эта закономерность проводится в подгруппе узких фонем: [i] > [ъ] в среднем на I6,4%, [y] > [ʊ] на I4,0% СДЗ. В подгруппе широких гласных выявленная закономерность в ряде случаев нарушается, с сохранением, однако, общего баланса в пользу переднерядных гласных.

Сравнение средних относительных длительностей репрезентантов огубленных и соответствующих им неогубленных гласных фонем в идентичных позиционно-комбинаторных условиях указывает на преобладание (в 75% случаев) длительности лабиализованных гласных над квантитативностью нелабиализованных коррелятов. Эта закономерность более регулярна в подгруппе широких гласных - как кратких, так и долгих: [ø] > [ɛ] в среднем на I5,3%, [o] > [a] на I4,8%, [o:] > [a:] на 5,4% СДЗ.

Итак, анализ ингерентной длительности кумандинских гласных позволяет констатировать следующее: I) широкие гласные регулярно длительнее своих узких коррелятов; 2) квантитативность аллофонов переднерядных фонем превалирует над темпоральностью соответствующих заднерядных гласных; 3) огубленные фонемы характеризуются большей фонической длительностью сравнительно с противопоставленными им неогубленными гласными.

Выявленная выше для кратких гласных последовательность распределения собственных, внутренне присущих каждому гласному длительностей - результат одновременной реализации в языке всех трёх закономерностей. Единственным нарушением симметрии системы является квантитативность фонемы [i] : в идеально логичной схеме [i] локализовалось бы после [ʊ] перед [ъ].

ЗАВИСИМОСТЬ ДЛИТЕЛЬНОСТИ ОТ ТИПА СЛОГА

Как показал анализ экспериментальных данных, зависимость фонической длительности кумандинских гласных от структурного типа слога по-разному реализуется в подсистемах кратких и долгих фонем. Как краткие, так и долгие фонемы характеризуются наименьшей средней относительной длительностью в приоткрыто-закрытых моносиллабах типа VC , V̄C ; наиболее длительные репрезентанты кратких фонем констатированы в призакрыто-открытых слогах типа CV , долгие же фонемы маркируются максимальной квантитативностью в призакрыто-закрытых слогах типа CV̄C. Распределение длительностей кратких и долгих гласных фонем в моносиллабах различной слоговой структуры можно отразить следующей формулой:

VC < CVC < CV < V̄C < CV̄ < CV̄C.

ЗАВИСИМОСТЬ ДЛИТЕЛЬНОСТИ ГЛАСНЫХ ОТ КОНСОНАНТНОГО ОКРУЖЕНИЯ

В зависимости от консонантного окружения отмечается определённая вариативность количественных характеристик кумандинских гласных. Наиболее стабильна корреляция гласных с качеством постпозитивных малошумных /l/ согласных: при типологическом тождестве препозитивного консонанта наиболее длительными являются оттенки гласных фонем перед последующим малошумным согласным. При этом средние относительные длительности широких гласных в моносиллабах типа CVL значительно превосходят СОД оттенков этих же фонем в препозиции к шумным согласным: для [a] эта разница в длительности максимального и минимального оттенков составляет в среднем 32,I% СДЗ, для [ɛ] - 39,8%, для [o] - 37,5%, для [ø] - 4I,8%, где максимально длительными являются репрезентанты соответствующих гласных в моносиллабах типа CVL ,минимально длительные - в моносиллабах типа CVS или CVC. Для узкого вокализма дифференциация фонической длительности гласных в зависимости от качества поствокала не столь контрастна, как для широкого, в среднем она составляет I2,3% СДЗ.

Влияние постпозитивных шумных согласных различного способа образования на квантитативные характеристики гласных выражено значительно слабее, чем воздействие малошумных, в большинстве случаев различие СОД оттенков гласных фонем перед шумными согласными различного качества составляет 2-3%, что практически индифферентно для восприятия.

Анализ количественных параметров гласных для выявления возможного влияния качества препозитивных согласных на квантитативность гласных не даёт оснований констатировать существование какой-либо закономерной связи между темпоральностью гласных и качеством препозитивных согласных.

ФОНЕМАТИЧЕСКАЯ ДЛИТЕЛЬНОСТЬ ГЛАСНЫХ

Учитывая выявленные выше закономерности в распределении фонической длительности гласных, при проверке количественных характеристик гласных на фонематическую признаковость представляется допустимым сопоставление длительностей фонем одинакового качества в строго идентичных позиционно-комбинаторных условиях (табл. 2).

Средняя относительная длительность кратких и долгих гласных в моносиллабах CVC*

Таблица 2

Тип слога	Фоне [a]	[a:]	[ɛ]	[ɛ:]	[i]	[i:]
TVT	95,1	I60,7	–	–	–	–
TVS	–	–	–	–	85,4	I46,6
TVČ	–	–	–	–	–	–
TVL	I23,4	I97,9	I26,I	I89,8	–	–
SVT	83,5	I52,2	–	–	–	–
SVL	–	–	II9,0	I72,7	–	–
ČVT	–	–	I09,8	I78,0	–	–
ČVL	I3I,4	202,2	II8,2	I95,9	–	–
LVL	II4,3	I49,8	I30,2	I92,0	–	–

Мы [o]	[o:]	[ʊ]	[ʊ:]	[ø]	[ø:]	[ɣ]-[ɣ:]
III,9	I65,0	75,6	I47,2	–	–	– –
99,7	I73,0	–	–	–	–	– –
–	–	–	–	96,0	I58,2	– –
I39,2	I99,3	86,7	I88,5	I42,4	I85,3	98,2-
I00,4	I69,3	–	–	I32,0	I58,4	I85,3
I3I,8	I86,7	–	–	II9,4	I87,8	– –
I26,0	I69,8	–	–	–	–	– –
–	–	–	–	–	–	– –
–	–	–	–	–	–	– –

* V – гласный, V̄ – долгий гласный, C – согласный, T – шумный смычный согласный, S– шумный щелевой, Č – шумный смычно-щелевой, L – малошумный согласный.

Результаты сопоставления приведённых данных не оставляют сомнений в правильности вывода, сделанного предварительно на слух, о релевантности для кумандинского вокализма признаков краткость-долгота. Реализации долгих фонем длительнее оттенков кратких коррелятов того же качества в аналогичных позиционно-комбинаторных условиях в I,4-2,I раза; средние относительные длительности долгих фонем отличаются от СОД кратких фонем на 50,0-87,I%. При этом зоны относительных длительностей (ОД) ни в одном из типов моносиллабов CVC не перекрывают друг друга, отстояние этих зон составляет I0,2-6I,7% СДЗ. Существование этой пограничной зоны длительности исключает возможность фонематического смешения гласных по количественному признаку.

КОРРЕЛЯЦИЯ ДЛИТЕЛЬНОСТИ ГЛАСНЫХ С КАЧЕСТВОМ ВОКАЛЬНОЙ ОСИ

Анализ количественных характеристик гласных в би- и полисиллабах свидетельствует о том, что в структуре многосложного слова реализуются те же закономерности, что и в односложных словах, но кроме типа слога и качества пре- и поствокала на длительность гласного в многослогах оказывают влияние тенденции, обусловленные характером вокальной оси.

В словоформах с вокальной осью, представленной качественно и количественно однородными гласными (taraq 'гребень', taraq-tar – мн.ч., taraqtardaq – исх.п.), проявляется тенденция к постепенному увеличению фонической длительности гласных к концу слова. По-видимому, это явление следует отнести к числу суперсегментных, как проявление словесного ударения, локализующегося на последнем слоге тюркского слова.

В бисиллабах с качественно неоднородной вокальной осью – с широким гласным в I-ом слоге и узким гласным во 2-ом – реализуется тенденция к удлинению широкого гласного в I-ом открытом слоге, причём это удлинение настолько существенно, что носит фонематический характер; следовательно, в данной позиции констатируется чередование краткой и долгой гласных фонем: tar 'порох' – ta:ram '(мой) порох', sat 'продавать' – sa:dap 'продав'.

В полисиллабах с качественно неоднородной вокальной осью также происходит существенное фонематическое удлинение широкого гласного любого (не только первого) открытого слога, непосредственно предшествующего слогу с узким гласным. Сопоставление средних относительных длительностей (СОД) широких **этимологически** кратких гласных в рассматриваемых структурах (I2I,I--I60,4% СДЗ) с СОД широких стяжённых долгих гласных (I30,I-I66,3% СДЗ) в идентичных позиционно-комбинаторных условиях свидетельствует об облигаторном чередовании широких кратких гласных любого открытого слога полисиллабов перед узким гласным следующего слога.

Удлинение широких гласных перед узкими – закономерность, характерная для ряда тюркских языков. Н.П.Дыренкова, говоря о локализации шорского ударения на последнем слоге и трактуя его как долготное, отмечает, в то же время, ряд исключений, когда ударение падает на I-ый слог и вообще на слог с широкими гласными, типа парбадым, келбессин и т.д. /8/. Н.К.Дмитриев указал на удлинение конечного гласного именных основ перед аффиксами личной принадлежности I-го и 2-го лица /9/. Р.М.Бирюкович констатирует аналогичное удлинение гласных и перед аффиксами 3-го лица /I0/.

Но во всех этих работах отмечаются лишь частные **моменты** общей закономерности. Наиболее адекватно эта универсальная для ряда тюркских языков особенность отражена М.И.Боргояковым, зафиксировавшим для хакасского вокализма удлинение широких гласных нормальной долготы до длительности стяжённых, т.е. долгих, перед слогом с узкими ы и i /II/. При такой трактовке удлинение конечных гласных именных основ перед аффиксами личной принадлежности, отмеченное Н.К.Дмитриевым для туркменского и Р.М.Бирюкович для чулымско-тюркского, а также отсутствующие в туркменском и якутском чулымско-тюркские долгие гласные в двусложных словах типа se:mis 'жирный', te:mir 'железо' можно рассматривать как частные моменты общей закономерности. Во всех этих случаях следует говорить не о восстановлении так называемой первичной долготы гласных, а о позиционном удлинении широких гласных перед узкими.

По особенностям реализации рассматриваемой тенденции язык кумандинцев несколько отличается от хакасского, где широкие гласные удлиняются до долгих только перед узкими ы и i и только в I-ом слоге. В кумандинском широкие гласные удлиняются также и в непервых слогах, причём не тслько перед ъ и ĭ , но и перед значительно менее частотными узкими ʊ и ɣ . В этом язык кумандинцев более сходен с шорским, где, по косвенным данным Н.П.Дыренковой, экспериментально подтверждённым для мрасского диалекта Н.В.Шавловой /I2/, удлинятся перед последующим узким гласным может широкий гласный любого открытого слога. Эта же закономерность отмечена В.И.Филоненко для балкарского языка /I3/, Л.П.Покровской для гагаузского /I4/, М.Рясяненом для тюркских языков Поволжья /I5/. Л.П. Покровская,отмечая позиционный характер долготы этих гласных в гагаузском, считает её нефонематичной. Однако такой трактовке противоречит следующее её же (Л.А. Покровской) замечание: "Согласно правилам орфографии они (т.е. позиционные долготы) не отражаются на письме, но в письменной практике ясно заметно стремление их отобразить" . Следовательно, носители языка интуитивно ощущают фонематичность рассматриваемых долгот. Удлинение широких кратких гласных перед узкими характерно также для казахского, киргизского, башкирского, чувашского, азербайджанского, т.е. является типологически общим для языков кыпчакских или сильно кыпчакизированных.

Итак, результаты анализа экспериментально-фонетического материала свидетельствуют о следующем:

I. Количественные характеристики гласных объединяются в две долготно-фонематические зоны – зону краткости и зону долготы. Такая стабильная инвариантная распределённость квантитативных характеристик гласных по двум достаточно чётко выдержанным зонам указывает на релевантность противопоставления по краткости-долготе и позволяет определить фонемы [a, ɛ, ɛ̆, i, o, ø, ʊ, ɣ] как к р а т к и е , а фонемы [a:, ɛ:, i:, o:, ø:, ʊ:, ɣ:] как д о л г и е .

2. Кумандинская система противопоставления гласных по длительности ещё не сформировалась окончательно. Широкие гласные чётко противопоставлены по квантитету, для узких же долгота исторически, по-видимому, развивается позднее.

3. В полисиллабах с качественно и количественно однородной вокальной осью реализуется тенденция последовательного увеличения длительности гласных к концу словоформы. Удлинение гласных финальных слогов при этом столь значительно, что позволяет трактовать эти гласные как долгие, констатируя в данной позиции чередование кратких фонем с долгими.

4. В полисиллабах с широким гласным в открытом слоге, предшествующем слогу с узким гласным, отмечается закономерное значительное удлинение этимологически кратких широких гласных перед узкими, также свидетельствующее о чередовании в данной позиции кратких широких гласных с долгими того же качества.

5. Отмеченное для кумандинского вокализма позиционное удлинение широких гласных открытых слогов перед слогами с узкими гласными, являющееся типологическим общим для языков кыпчакских или сильно кыпчакизированных, может свидетельствовать о наличии в кумандинском этническом образовании кыпчакско-тюркских компонентов.

/I/А.И.Ярхо.Алтае-саянские тюрки. Абакан, I947,с.I24-I25;Г.Ф.Дебец.Палеоантропология СССР. М.-Л., I948, с. I29-I30.

/2/Л.П.Потапов. Этнический состав и происхождение алтайцев. Л.,I969.

/3/Н.А.Баскаков.Диалект кумандинцев (куманды-кижи).М.,I972; И.Я.Селютина.Самодийские субстратные черты в консонантизме северных алтайцев. – Шестой международный конгресс финно-угроведов.Сыктывкар,I985.

/4/В статье использована транскрипционная система Л.В.Щербы на основе алфавита МФА с уточнениями В.М.Наделяева.– В.М.Наделяев. Проект универсальной унифицированной фонетической транскрипции(УУФТ).М.-Л.,I960. Он же. Артикуляционная классификация гласных. – В кн.: Фонетические исследования по сибирским языкам. Новосибирск, I980,с.3-9I.

/5/Пневмоосциллограф – модифицированный осциллограф, осуществляющий синхронную раздельную по каналам носа, рта и горла запись речевого сигнала на киноплёнке(фотобумаге) с фиксацией показателей отметчика времени.

/6/Средняя относительная длительность звука (СОД)=$(АД_1+АД_2+...+АД_n) : (СДЗ_1+СДЗ_2+...+СДЗ_n)$ x I00%, где АД – абсолютная длительность звука,n – количество словоформ с исследуемым звуком, СДЗ – средняя абсолютная длительность звука в словоформе.

/7/Сонорные,реализующиеся в звонких и глухих оттенках,трактуются как малошумные.

/8/Н.П.Дыренкова. Грамматика шорского языка. М.-Л.,I94I, с.22-23.

/9/Н.К.Дмитриев.Долгие гласные в туркменском языке.– Исследования по сравнительной грамматике тюркских языков.М,I955,с.I90.

/I0/Р.М.Бирюкович. О первичных долгих гласных в чулымско-тюркском языке. – Советская тюркология, I975, № 6, с.55-67.

/II/М.И.Боргояков. Об образовании и развитии некоторых долгих гласных в хакасском языке. – Уч. записки Хакасского НИИЯЛИ. Абакан, I966, вып. XII, с. 8I-98.

/I2/Н.В.Шавлова. Словесное ударение в нижнемрасском говоре шорского языка. Автореф. канд. дис. Алма-Ата, I983, с. 9-I0.

/I3/В.И.Филоненко. Грамматика балкарского языка. Нальчик, I940, с. I2-I3.

/I4/Л.П.Покровская. Грамматика гагаузского языка.Фонетика и морфология.М.,I964.

/I5/М.Рясянен. Материалы по исторической фонетике тюркских языков. М., I955.

THE NATURE OF THE SO-CALLED NON-SYLLABIC VOWELS IN THE UKRAINIAN LANGUAGE

N.I.Totskaya

Dept. of Philology
Kiev Shevchenko State University
Kiev, the Ukraine, USSR, 252030

The report, based on the data of the experimental phonetic analysis /X-ray cinematography and spectography/ is devised to prove that the so-called non-syllabic vowels in the Ukrainian language /u̯/, /i̯/ are acoustically, articulatory and functionally different from the vowels /u/, /i/. It is suggested that /u̯/, /i̯/ are typical representatives of the sonants /w/, /j/, and that the terms non-syllabic vowels, or half-vowels are unsatisfactory and do not describe the true nature of the sounds.

One of the characteristic features of the system of sounds in the Ukrainian language is further vocalization of the maximally voiced sonants /w/, /j/ and their change into half-vowels or so-called non-syllabic /u̯/, /i̯/. On the other hand, the unstressed vowels /u/, /i/ are weakened in some positions and also change into non-syllabic /u̯/, /i̯/ Thus we may say that the non-syllabic vowels /u̯/, /i̯/ create a link between vowels and consonants.

The existense of the sounds /u̯/, /i̯/ in Ukrainian is well-known, attempts were made to study their origin and functioning as well as their acoustic and articulatory nature /A.A.Potebnya, P.G.Zhitetsky, N.F.Nakonetchny, I.P.Suntsova, P.P.Kostruba, M.A.Zhovtobrukh, V.S.Perebyinos and others/. However the phonetic nature of these sounds has not been duly investigated and it is still to be proved whether they are vowels or consonants and whether it is necessary to establish a separate type of sounds - "half-vowels" or "non-syllabic vowels".

The aim of this report is to present comprehensive analysis of the acoustic and articulatory nature of the sounds /u̯/, /i̯/ in modern Ukrainian as revealed in their functioning.

The data for investigation were obtained by means of spectography, X-ray cinematography and auditory observation.

Phoneticians agree that the sounds /u̯/, /i̯/ occur in the following positions: in the beginning of a word preceeding a consonant or a group of consonants /u̯dowa/, /u̯poru/, /u̯prawa/, /u̯las/, /i̯du/ /i̯mu/ in the end of a word or syllable after a vowel as a result of vocalization of the consonants /w/, /j/ /dauno/, /buu/, /stau̯ka/, /l'ij/, /toj/, /l'i̯ika/, /soi̯ka/ ; in other cases, i.e. in the beginning of a word or syllable or in the middle of a word in the intervocalic position, their main allophones are the sounds /w/ or /j/ (/woda/, /towar/, /jama/, /majak/). Thus, it becomes evident that the phonemes /w/ and /j/ may be realized as allophones of both vowel and consonant types, i.e. they are ambivalent in nature.

The nature of sound traditionally defined as /u̯/, /i̯/ can be made clear only by means of articulatory and acoustic collation with the similar vowels /ú/, /u/, /í/, /i/ as well as with the consonants /w/, /v/, /j/, in which the phonemes /w/ and /j/ may be realized.

The results of the analysis show certain differences, so each sound will be treated separately.

X-ray cinematographic stills show that in the beginning of a word the articulation of /u̯/ preceeding a consonant, as in /uzuu/, differs in principle from /ú/ and /u/ in three aspects: a/ the sounds /ú/ and /u/ as vowels are characterized by a distinct movement of the back of the tongue to the soft palate; the sound /u̯/ as the consonant /w/ does not demand any strict localization of the back of the tongue, normally it acquires the shape necessary for the pronunciation of a preceding or following sound; b/ the distance between medial incisors in pronouncing /u̯/ is shorter than in the case of /ú/ or /u/, thus approaching to /w/ or /v/, or, to be more exact, this distance changes from /w/ to /v/, i.e. from a bilabial to a labio-dental sound; c/ in pronouncing /ú/, /u/ the lips are rounded and protruded more in the case of /ú/ and less in the case of /u/; in pronouncing /u̯/ the position of lips is changeable: first bilabial, then labio-dental; the lips are slightly rounded, but not protruded and not tense. Schematic diagrams clearly show that the sounds of the /u/ class are characterized by a wide lip opening, and, consequently, the absence of the typical for /w/- sounds constriction.

The same can be seen in the end of a word or syllable. Thus, it may be clearly seen that in the word /uzuu/, where /u̯/ follows /ú/, the lips are less protruded after its articulation, the distance between them narrows, the back of the tongue moves somewhat down and forward, as in the case of the non-stressed /u̯/ the position of lips changes from bilabial to labiodental.

Spectrograms give the possibility to broaden the scope of investigation. It turns out that /w/ preceding vowels may function in a number of allophones. The sound, which is the most frequent in this position, has only the fundamental tone, it has no formant structure, typical for vowels, no noise components, typical for noise consonants, i.e. this is a sonant. Another type is a noise consonant: besides the fundamental tone, there are other noise components of different intensity in the sphere of high frequency. The third type is the sound with an unstable articulation: it is a sonant at the beginning, and a noise consonant at the end. The fourth type is represented by very rare occurrences, when /w/ is substituted with a sound with a formant structure, i.e. of a vowel type.

The same variation can be observed not only in the end of a syllable after a vowel, but contrary to our notions, in the intervocal position as well, with /w/ in the latter case occurring in 36 per cent of instances with a more or less expressed noise components. In some instances these are sounds characterized only by the fundamental tone or weak traces of the formant structure.

Auditory observation shows that in the colloquial style, when the tensity is weakened, the noise components are less prominent. Naturally, this may be regarded as strengthening of sonority alongside with the weakening of articulation in the phonation focus.

The above analysis makes it evident that the sound represented in transcription as a non-syllabic vowel /u̯/ is in almost all instances more similar to consonants, sonants in fact, than to vowels. Not only articulation and acoustic structure testify to the fact, but also the tendency in careful pronunciation to strengthen tensity of the speech organs in the phonation focus.

The specific characteristics of the articulatory nature of the sound /i̯/ are more difficult to define by means of X-ray cinematography. Nevertheless the comparison of the first sound in the word /jama/, traditionally represented as /j/, with the stressed /í/ and the unstressed /i/ shows that the lips are the most widely open in the case of /í/, and the least widely open in the case of /j/, with /i/ occupying the intermediate position. The back of the tongue occupies the highest position in the case of /j/, it is less high in the case of /í/ and still less in the case of /i/. Evidently this is the result of greater or smaller tensity of the speech organs in the phonation focus of these sounds. Unfortunately in this case the character of the constriction cannot be defined by means of X-ray cinematography. Still there is little doubt that the articulatory initial sound considerably differs from /í/ and /i/, possessing more consonant qualities.

The same differences can be observed if we compare the articulation of /i̯/ in the word-final position after a vowel with the articulation of /i/ and with the articulation of /j/. The position of the back of the tongue in the case of /i̯/ is similar to that of unstressed /i/. Unfortunately radiograms don't display the character of labial and linguo-palatal constriction.

Spectography gives a more complete notion of the sounds in which the phoneme /j/ is realized, displaying a wide range of their acoustic qualities.

Thus, in the beginning of a word before a vowel and in the intervocal position the following variants are realized: 1/ The sound, which has the fundamental tone only. 2/ The sound which, apart from the jundamental tone, has traces of the formant structure, specifically those of F_1. These qualities make it possible to qualify both types of sounds as sonants, the second type being the sound with a more vividly expressed sonority, which, however, doesn't develop into vocality, as both sounds have no distinct formant structure. 3/ The sound, which, apart from the fundamental tone, has more or less distinct noise components in the highest frequencies - i.e. this is a typical voiced noise consonant. 4/ There are instances when in the intervocal posttonic position (/rozwal'ajec'ca/) /j/ devoid of noise components, is obviously weakened, reduced and fuses with the neighbouring vowels. 5/ In the word-final position, far from the stressed syllable (/dobroji/) the phoneme /j/ may have the qualities of a typical voiced noise consonant. Thus we may conclude that the majority of realizations in this position are sonants, some are noise consonants,

the remaining realizations are sporadic, depending mainly upon the style of pronunciation.

Distinct variation may be observed also in the end of a syllable after a vowel. The most frequent realization here is the sound which consists of the fundamental tone only. Sometimes traces of the formant structure of the previous sound may be observed. Finally, in very rare instances, spectrograms show the picture, slightly reminding of unstressed [i̯]. When the sound in the word-final position is characterized by a changeable structure, another variant may occur, which begins with a voiced and ends with a voiceless noise consonant (/jij/, /st'ij/).

The picture will be complete if we dwell on the problem of the phonological status of non-syllabic [u̯] and [i̯]. Functional approach made it possible to prove the consonant character of [u̯], [i̯], /P.P.Kostruba/. This conclusion was confirmed by applying the principle of complementary distribution /V.S.Perebyinos/: [u̯], [i̯] can't be allophones of the phonemes /u/, /i/, but they undoubtedly are allophones of the consonants /w/, /j/, with which they are in the relations of complementary distribution.

Summing up the results of our analysis we come to the following conclusions.

The range of realization of the sonants /w/, /j/ is wide enough, from the voiced sonant (traditionally of the type /u̯/, /i̯/) to the voiceless noise consonant and even - in careless pronunciation - the reduced sound of an indefinite quality. Of the voiced noise consonants [b], [d] may be fully voiced half-voiced (with different degrees of the loss of the quality), they may alternate with voiceless consonants, may be reduced to the loss of their distinctive features.

The sonants /w/, /j/, that occupy the intermediate position between vowels and noise consonants in one part of their allophones approach vowels, in the other allophones they are close to consonants.

The so-called non-syllabic vowels [u̯], [i̯] are by their articulatory and acoustic nature not vowels, but resonant consonants, occurring not only in the beginning of a word before a consonant or in the end of a word or syllable after a vowel, but also before a vowel, as well as in the intervocal position - i.e. practically in any position.

Naturally the question arises what allophones of the sonants [w], [j] are to be considered typical. The investigation shows that it is the allophones [u̯], [i̯] that are the typical representatives of the phonemes /w/, /j/. It is these allophones that preserve the essential features of the sonants /w/, /j/. In this case the term "non-syllabic vowels" as well as the notion itself can hardly be considered justifiable. If the sonants are considered without "non-syllabic vowels" there remains practically nothing apart from comparatively rare noise allophones. This term as well causes confusion in the definition of vowels, which are characterized by the existence of the formant structure. As it has been shown, the so-called non-syllabic vowels [u̯], [i̯] do not possess the formant structure. Moreover, the admission of the existence of the so-called non-syllabic vowels or halfvowels brings about the necessity to divide all the sounds in the Ukrainian language into vowels, consonants and half-vowels, which does not correspond to the existing reality neither on the phonemic nor or the phonetic levels.

If the sonants [w], [j] are more than normally vocalized, they can alternate with the voiceless vowels [u̥], [i̥] (e.g.: /wranc'i/ - /uranc'i/; /wpered/ - /upered/). The opposite is also possible - the alternation of [u̥], [i̥] with [w], [j]. This is the characteristic feature of the Ukrainian language.

ON THE PROBLEM OF ACOUSTIC PECULIARITIES OF STOPS IN SOME LANGUAGES OF THE CAUCASIAN AREA

Ivane Lezhava, Nana Gamkrelidze

Laboratory of Experimental Phonetics
Tbilisi State University
Tbilisi, Georgia, USSR, 380028

The point argued below is that the position of the vocal cords in the articulation of voiced stops in initial position can be interpreted not only on the basis of articulation and physiology but also by the paradigmatic characteristics of the system. These conclusions are drawn as a result of the spectrographic analysis of the stops of the Abaza, Avar, Adyghe, Hunzib, Ingush, Chechen, Lak, Udi, Georgian, Zan, Svan, Lezgian and Ossetic languages.

The basic model of the stops of the Caucasian languages represents a triple (three-member) system: voiced-aspirated-glottalized. This system, due to the existence of the force correlation, undergoes transformations and results in a four-member system or a triple system of a different type /1/.

In spite of general similarity, or even of the complete adequacy of some of the systems of stops in the Caucasian languages, distinctions between them show themselves during the processes of production of some features.

According to the analysed material, three degrees of voicing of the voiced consonants of the Caucasian languages are distinguished: the Kartvelian (Georgian, Zan Svan) languages are characterized by a low degree of voicing (the occurrence of voiced consonants in the initial position with the voiceless phase of closure). Initial voiced stops of the Avar and Abaza languages represent the medial stage. Abazian voiced -b, d, g, g°, ǯ°, ʒ°,-consonants are produced with the voiced closure and -g̃, ʒ , ǯ -consonants - with the devoiced phase of closure. An Avar initial voiced stop in connected speech is, as a rule, represented by its devoiced variant, but in an isolated word, the end of the closure may become voiced. Voiced stops of the other languages are produced with the phase of voiced closure in initial position.

Low degree of voicing of Svanian voiced stops results in the item acoustically similar to specifically pronounced glottalized sounds. In producing the Svanian glottalized sounds either a reduced vibration of the vocal cords or a vertical vibration of the closed vocal cords seems to take place in a number of cases. Apparently, this fact can to some extent account for the phonetic changes that have taken place in some languages.

In such triple system of stops, as Udian, where non-aspirated consonants are produced in place of glottalized sounds, voicing of voiced stops is complete.

In the speech of the elder generation of Vartashen dialect of the Udi language (on the territory of Georgia, Oktomberi-village), the existence of glottalized stops and the pronunciation of initial voiced consonants with a voiced or devoiced phase of closure is optional. Whereas, in the speech of the younger generation voiced consonants are used only by a devoiced variant and non-aspirated voiceless stops are substituted by glottalized ones.

The existence of glottalized stops in the speech of the inhabitants of Oktomberi village might be due to the influence of the Georgian language. It is less probable that the oldest state is preserved. A high degree of voicing of voiced stops is characteristic of the system of the stops of the Ossetic language, which acquired glottalized stops in the Caucasian environment (this refers to dialect of the Ossetic language which is spoken by the Ossetes living on the territory of the Georgian SSR.)

Thus, in spite of the Georgian environment, in this case, the degree of voicing is preserved. The larger number of Ossetes in comparison with Udians may be taken as the extralinguistic cause of this fact, and the existence of the "fourth row" (strong) stops which are remainders in the system of stops and have little duration of noise in comparison with the aspirated sounds, may be taken as its linguistic cause.

In initial position Lak voiceless strong stops are reproduced with a decreased duration (in their non-aspirated form) in comparison with the aspirated sounds. In

intervocalic position the duration of the phase of the closure of a strong stop is about twice as much and the duration of noise is about 1.5 times less in the corresponding phases of an aspirated stop. In Avar a strong stop in initial as well as in intervocalic position differs from an aspirated consonant mainly in its long duration of noise.

The Avar t̄ in intervocalic position shows other characteristics received by the devoicing of -d- or by the succession of two -t-s. It is reproduced with the perceptible increase of closure and decrease of noise.

As noted above, the degree of voicing of the Avar voiced stops is lower than that of the Lak language.

We may suppose that in Lak and Avar languages there exists an interdependent relationship between different representations of strong consonants and different degrees of voicing of voiced consonants. In Lak the non-aspirated reproduction of an initial strong stop must favour the formation and preservation of voicing or, vice versa, the existence of voicing must determine the acoustic picture of a strong consonant. The same relation exists in Avar. It is probable that a long duration of noise of a strong consonant determines the low degree of voicing, or, vice versa, a low degree of voicing determines the long duration of noise of a strong consonant.

Thus,it may be concluded that in triple as well as in four-member systems of stops, the existence of a non-aspirated consonant conditions complete voicing of an initial voiced stop. In the systems of stops with glottalized consonants, complete voicing of the initial voiced stop is not obligatory. Weakening of intensification, disappearance or appearance of one feature in the sound system may cause alternations of other features.

Reference

I.Структурные общности кавказских языков. -М.:Наука,1978.

MULTIDIMENSIONAL ANALYSIS OF THE SIMILARITY OF PITCH CONTOURS

GRAZYNA DEMENKO

Acoustic Phonetics Research Unit
Institute of Fundamental Technological Research
Polish Academy of Sciences
Noskowskiego 10, 61-704 Poznan, Poland

ABSTRACT

In order to find the relations between the physical and perceptual analysis of fundamental frequency, a number of listening tests were performed and evaluated by means of Multidimensional Scaling. The experimental materials consisted of utterances with eight different intonation patterns. On the basis of results obtained from automatic pitch pattern recognition, such cases were selected as would represent (1) a 100% recognition (2) fair recognition (about 50% correct) (3) poor recognition (about 20% correct). The listening panel judged the proximity between the elements in each case including two replications of each of two patterns. The purpose of the experiment was (a) to establish the perceptual dissimilarities between the patterns (b) to create a basis for the classification and (c) to compare the results of an objective and a subjective analysis.

1. INTRODUCTION

The analysis of prosodic features takes a significant position in an acoustical and a perceptual description of the speech signal. The F0 parameter (the fundamental frequency) is the subject of much theoretical and experimental work. Experimental investigations may be performed at the perceptual or the physical level. A selection of just one of them does not ensure proper analysis procedures. Perceptual experiments may be objected to on the grounds of subjectivity. On the other hand, purely instrumental analysis may lack a clear relation to linguistic entities. As it is generally accepted that variations of fundamental frequency produce, at the perceptual level the sensation of tone height, a psychoacoustic analysis of this parameter appears to be very much to the point. Temporal variations of fundamental frequency are due to a number of effects that vary themselves during an utterance. It is essential, for the analysis of this parameter, to define which of the many possible sources of variability are effective in a given case.

In [3], the various sources of variability of F0 were briefly discussed. If it is desired that most of the manifold variability sources be kept out, the experimental material should include only simple utterances. An analysis of more complex melodic structures requires a prior discrimination of the functional units of intonation. The present work attemps to find possibilities of evaluating the physical and the perceptual similarities between various simple pitch curves and to classify the curves on the basis of a limited set of prototypical natural Polish utterances.

2.PREPARATION OF THE EXPERIMENTAL MATERIAL.

The Polish phrase "Dobrze" = approx."all right" was uttered by a phonetician with 8 different intonation patterns.The utterances were recorded at intervals of approx. 5 s. The patterns (treated as prototypes) were reproduced over loudspeakers to be immediately and without reflection imitated by the test person who was always asked just to repeat what he or she had heard, as naturally as possible, with their own natural voice, without any attempt to mimick. 15 native speakers of Polish (10 males and 5 females) were used as test subjects. Three of them had previously been exposed to professional phonetic training. The reproduction of each of the 8 Prototypes: (1) Low Rise, (2) Full Rise, (3) High Rise, (4) Low Fall, (5) Full Fall, (6) Level. (7) Low Rise Fall, (8) Full Rise Fall was performed in several sessions, altogether 10 times by each subject.

3. A MULTIDIMENSIONAL STATISTICAL ANALYSIS.

A fundamental problem in any recognition procedure is the selection of characteristic features.A method which is optimal with respect to data description uses eigenvectors of the covariance matrices (the Karhunen - Loeve method). It was used for data reduction in F0 curves. e.g., by ATAL ([1]). But the aim of recognition is a discrimination of classes, so better possibilities are offered by subspaces constructed on the basis of discriminant vectors.

The problems of discriminant analysis are presented in a number of publications, e.g. [5]. The aim of a discriminant analysis is to find a subspace in which the total dispersion of the data collection will be maximum relative to the within-class dispersions.

It was assumed that using the discriminant analysis it would be possible to examine differences between F0 curves, to define the features necessary for their correct discrimination and to establish the possibility of their classification.

In order to eliminate differences caused by varying pitches of the individual voices, frequency normalization was performed. The logarithm of the lowest value was substracted from the logarithm of successive frequency values within a curve. Then, the difference between the means for the frequency variation ranges of the given voice producing the prototypes was added or subtracted leading to the desired relation among the reproductions as well as between these and the Prototypes.

In order to normalize for time, as well as reducing data, each utterance was divided into 8 parts within which average frequency was calculated as the reciprocal of mean period length. It was accepted that the Prototype utterances and their 10 replications by each of 3 of the imitators (the phoneticians) will form the classes to be examined. The pitches of these voices differed: the lowest frequency for the two male voices was 65 Hz and that for the female voice, 160 Hz. Each of the individual classes was thus represented by 30 replications. Fig.1 depicts a classification tree over the mean vectors of the classes under examination. The values of the Hotelling T^2 statistic are placed over each of the connecting lines. By comparing these with the critical value at the 5 percent significance level, (which was 88.35) it was found that all distances between the classes are statistically significant. The performed analysis leads to the following conclusions:

(1) the classes under examination may be defined in a 2-D space with 90 percent correct distances between them or in a 3-D space with 99 percent accurancy in the distances

(2) the differences between the individual classes are all statistically significant.

3. CLASSIFICATION.

As the features corresponding to the discriminant variables represent an optimal set with respect to recognition, a description of the F0 curves in terms of these features appears desirable.

The first discriminant variable is interpretable as the slope of a straight line passing through the initial and the terminal point of the time and frequency normalized F0 curve.

The second characteristic of the set under examination was defined as the initial frequency value of the curve. Although a satisfactory description of the curves was obtained with only two variables, a third variable (see above) will slightly improve the classification. It is related to the degree of convexity or concavity of the curve. A still more precise description may be obtained when a fourth feature is introduced, viz. the location of the extremum.

One of the basic methods used in deterministic classification is referred to, in recognition literature, as the "perceptron algorithm", with the decision functions• generated from patterns provided for the computer by an iterative learning algorithm. The coefficients of the decision function have here been defined as follows:

It was assumed that there exist M decision functions having the property that if $x \in \omega_i$, then $d_i(x) > d_j(x)$ for all $j \neq i$, x being the vector to be recognized and ω_i being the class ωi. Let us consider M classes $\omega_1 \ldots \omega_M$ and assume that in the kth iterative step during the learning stage the pattern x belonging to class ω_i is presented to the computer . The decision functions

$d_j[x(k)] = w_j'(k)\, X(k)$ are calculated,

and if

$d_i[x(k)] > d_j[x(k)]$ for $j = 1,2,..M$ and $j \neq i$,

then the weighting vector w_j remains unaltered in the next iterative step:

$W_j(k+1) = W_j(k)$ for $j = 1,2..M$.

Otherwise, the weighting vector is altered in accordance with the relation

$W_i(k+1) = W_i(k) + cX(k)$

else

$W_i(k+1) = W_i(k) - cX(k)$

where c is a constant. If the classes are linearly disjoint, then the algorithm is convergent in a finite number of iterations for an arbitrary initial weighting vector.

Our learning set included all the replications produced as imitations of the Prototypes by the three phonetically trained subjects. For each speaker,3 replications of each of the eight classes were selected at random and subjected to the recognition procedure. Fig. 2 presents the results which were in agreement with the assumed classes 80 percent of the time. This suggests that the algorithm should be modified by using a greater number of features. As the classes turned out not to be in fact linearly disjoint, an alternative type of decision functions may be preferable. Fig.2 also shows the results of recognition of the entire collection of 1200 curves using a different method, viz. quadratic statistical discriminant functions. The method and the results will not be here discussed (see the companion paper by W.JASSEM presented at this Congress) except for mentioning that 8 features were used there. But it is noteworthy that though the deterministic algorithm yielded distinctly poorer results, both methods divided the test subjects into identical three groups of very good imitators (LR, JI, WJ, WI) good ones (AM, HK, KK, BS, MC) and bad ones (TK, RC, MK, CW, BI, PD).

4. PERCEPTUAL ANALYSIS.

The advances in methods of computation and optimization of the recent 15-20 years permitted the development of a method of evaluating the results of perceptual experiments known as Muldidimensional Scaling [4]. Its aim is to find a configuration of n elements such that the distances between them should correspond to subjective dissimilarities between observed objects. A monotonic relation between the distances and the dissimilarities is required. The concept of stress is introduced to reflect the measure of non-monotonicity, i.e., of the error in the approximation to the experimental data. Except for degenerated systems, the stress is minimum for the optimum configuration. The quality of the configuration is generally described as very good if the stress is 5 percent or less, good if between 5 and 10 percent and acceptable up to 20 percent. An extensive study of the psychological process involved in the perception of tone in speech was preesented by Gandour [2] On the basis of results obtained in Multidimensional Scaling, the author accepted two features as being characteristic: the mean frequency and the direction of pitch movement. He confirms the stability of these features and concludes that other dimensions are difficult to interpret.

The purpose of the listening experiment to be reported here was to seek the answer to the following questions: (1) Are some of the different intonations perceptually similar? (2) Do the listeners consistently use the similarity measures ? (3) Is there a systematic relation between perceptual similarity and some physical features of the pitch curves?

By reference to the results of automatic recognition of the intonations (the deterministic model), listening tests were prepared which consisted of the utterances of one very good imitator (WJ), one good imitator (MC) and one bad imitator (TK, see above). A panel of 20 listeners (all university students) judged the similarities between pairs of stimuli. 2 replications were randomly selected for each of the three voices and each of the 8 intonation patterns producing for each voice a collection of 136 stimulus pairs. The listeners judged the similarity between the members of each pair on a scale of "0" to "4",with an increase of the rating reflecting the measure of similarity. 28 pairs of stimuli were administred in one session. The measure of similarity between any two stimuli was defined as the sum of the ratings obtained from all listeners.

The results of the test are presented in Fig.3, with the second selected replication indicated by a prime. In a 2-D space it can be seen that for voice WJ the replications form distinct clusters, that for voice MC the utterances 2, 2', 3, and 3' form a single cluster whilst the other replications group together, and that for TK there are five clusters: (1) 2, 2', 3, 3', (2) 4, 4', 7, 7', (3) 8, 8', 5, 5', (4) 1, 1', (5) 6, 6'. In order to show how these perceptual results are related to the physical properties of the stimuli Fig. 4 a presents the 10 replications of patterns 2 and 3 as produced by MC whilst Fig.4 b shows the replications of patterns No. 4, 5, 7 and 8 as produced by voice TK. It is clear from Fig. 4 that the intonations that are confused in perception are also indistinguishable as F0 curves. The two perceptual **dimensions** obtained in the present study may be described as relating to the steepness of the curve (the first dimension, i.e. the strongest distinctive feature) and to the terminal pitch (the second dimension, i.e. the weaker distinctive feature).

CONCLUSIONS

1. Both the automatic and perceptual analysis permitted a classification of the F0 patterns.
2. The set under examination can be described using a few features. The first two are statistically and perceptually most significant.
3. A final automatic classification of the intonation curves requires more stringent methods.
4. Perceptual classification would be improved by considering differences between individual listeners (INDSCAL).

REFERENCES

1. ATAL, B.: Automatic Speaker Recognition Based on Pitch Contours, JASA, vol.52, No. 6, 1687 - 1697, 1972.
2. GANDOUR, J.T.: Perceived Dimensions of 13 Tones: A Multidimensional Scaling Investigation, Phonetica, vol. 35, No.3, 169 - 180, 1978.
3. JASSEM, W., DEMENKO,G.: On extracting linguistic information from F0 traces, Studies of Intonation in Discourse, (C.Johns-Levis ed.) London, 1984.
4. KRUSKAL, J.B.: Nonmetric multidimensional scaling: a numerical method, Psychometrika, vol.29, NO.2, June, 115 - 129, 1964.
5. LACHENBRUCH, P.A.: Discriminant analysis, Hafner Press, New York, 1975.

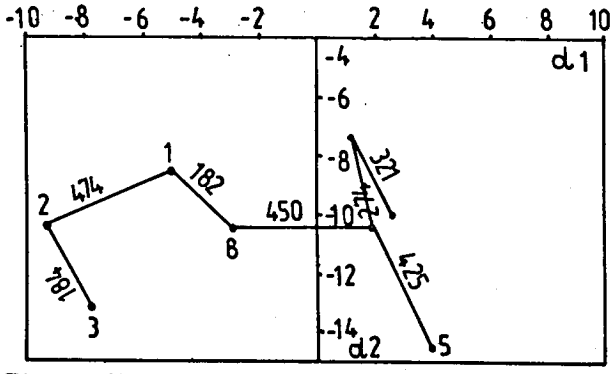


Fig.1. Mean vectors of the 8 classes in a coordinate system of discriminant variables.

Fig.2. Error scores for F contours (a) in the deterministic algorithm (blank areas) and (b) using quadratic discriminant functions (shaded areas).

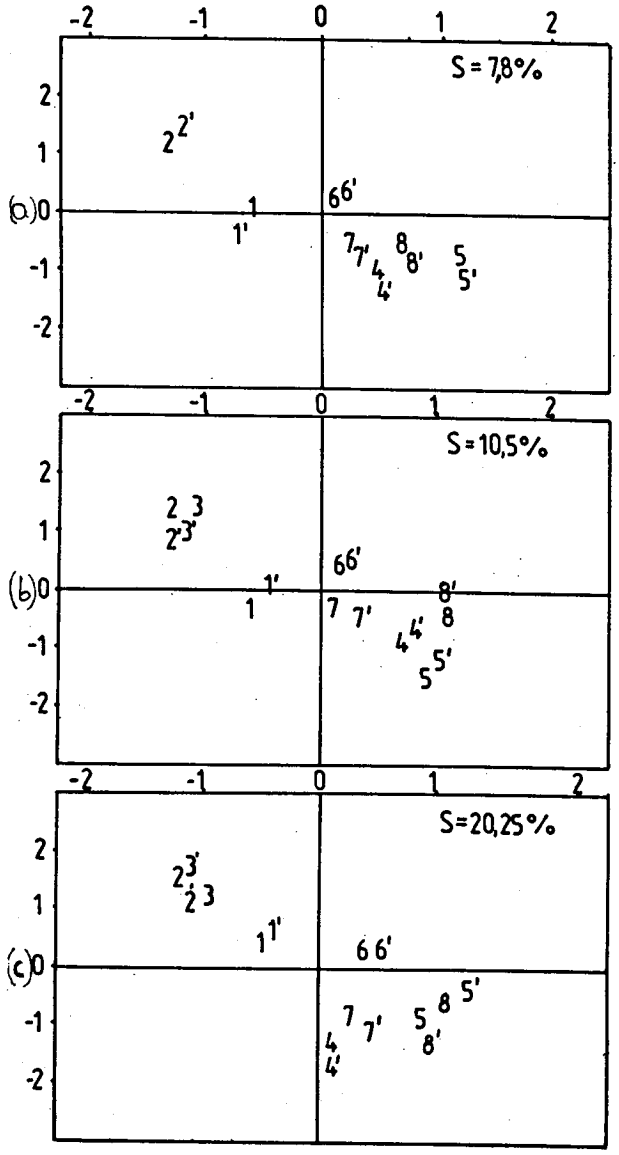


Fig.3. Results of Multidimensional Scaling
(a) voice WI
(b) voice MC (c) voice TK

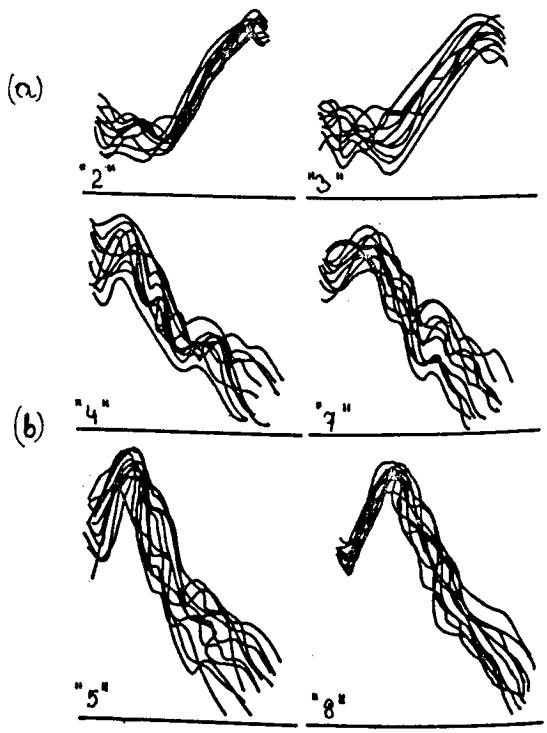


Fig.4. Replications of patterns
(a) "2", "3". Voice MC
(b) "4", "7", "5", "8". Voice TK

COMPUTER-ASSISTED CLASSIFICATION OF BASIC POLISH INTONATIONS

WIKTOR JASSEM

Acoustic Phonetics Research Unit
Institute of Fundamental Technological Research
Polish Academy of Sciences
Noskowskiego 10, 61-704 Poznan, Poland

ABSTRACT

An experiment was performed to explore, at a basic level, acoustic differences among FØ contours as related to linguistic and perceptual distinctions among intonation patterns. Each of 8 distinct pitch patterns was reproduced, in three sessions 10 times by 10 male and 5 female speakers of Polish. The FØ contours were treated as vectors in an 8-D space. Quadratic and linear discriminant functions were used for an automatic classification of the 1200 vectors with scores of over 80% correct. The misassignments were largely due to missing distinctions in the imitations. It is suggested that not all linguistic distinctions in intonation are categorical. The discriminant functions also permitted a study of similarities and dissimilarites among the different patterns.

THE PURPOSE OF THE EXPERIMENT AND THE DATA

The relations between the properties of an FØ curve as a representation of an acoustical, i.e., physical event and the corresponding linguistically distinct intonation pattern are largely unknown. The present study attempts to come to grips with the basic issue of describing some simple FØ curves so as to be able to assign them automatically to perceptually -- and -- presumably -- linguistically distinct classes.

A phonetician (WJ) recorded 8 versions of the Polish phrase "Dobrze." (/dobʒe/, approx. "OK"), each with a different intonation, viz. Low Rise (LR), Full Rise (FR), High Rise (HR), Low Fall (LF), Full Fall (FF), Level (L), Low Rise-Fall (LRF) and Full Rise-Fall (FRF), with pauses of 5 s. Both with respect to distribution (in discourse) and "meaning" (in a broad sense of the term), the intonations are all different. So it could be assumed that they might be treated as linguistically and perceptually distinct. 10 male and 5 female speakers of Polish listened to these utterances -- the prototypes -- and used the pauses to reproduce (repeat) them with the same "tone of voice". The reproductions (or imitations) were recorded in three different sessions, the first and the last being one month apart, and included 10 replications of each prototype by each speaker. All the 1200 new recordings were analyzed using a period-length meter and a minicomputer. Time normalization was obtained by dividing each utterance into 8 equal fragments and calculating average frequency in each fragment. Thus, the intonation contour of each utterance was represented by a sequence of 8 numbers. The raw data were also frequency normalized (after conversion to a log scale) by putting the mean for each individual voice from all his or her 80 utterances at zero and the variance at one (statistical standarization). This eliminated differences between speakers and allowed one average pattern to be obtained from 150 tokens for each of the 8 FØ contours as shown in Fig. 1.

STATISTICAL TREATMENT

After time and frequency normalization, each of the 1200 utterances was mathematically treated an an 8-element vector, i.e., as a point in an 8-dimensional space. The elements of each vector were the normalized successive frequency values. For each vector, 8 quadratic and 8 linear discriminant functions ere calculated to decide, in two ways, to which of the assumed eight classes: LR,FR,HR,LF,FF,L,LRF or FRF the vector belonged. This was indicated by the highest value of the discriminant function. Also, by observing the decreasing values of the remaining discriminant functions -- DFs --, the relative similarity of each utterance to each of the averaged patterns could be stated. The two kinds of DFs were: (1) The estimator of the quadratic discriminant function, EQDF, and (2) the estimator of the linear discriminant function, ELDF, of the following forms:

EQDF

$$\hat{u}_{ij}(x) = \frac{1}{2}\left[\frac{N_i - p - 2}{N_j - 1} D_j^2(x) - \frac{N_i - p - 2}{N_i - 1} D_i^2(x) + \ln\left(|S_j|/|S_i|\right) - h_4(p_i N_i, N_j) + h_3(p_i N_i, N_j)\right]$$

ELDF

$$\hat{e}_{ij}(x) = \tfrac{1}{2}[h_4(p,N,K)D^2_j(x) - D^2_i(x) + K^{-1}(p,N,K)h_3(p,N_i,N_j)]$$

In the above expressions, x is the observed vector, N is the sample size (here, 150 everywhere), p is the number of dimensions (here, 8), K is the number of classes (here, 8) and S_i , S_j are within-class covariance matrices:

$$D^2_i(x) = (x - \bar{x})' S_i^{-1} (x - \bar{x}) \quad , \quad i,j = 1,\ldots,k \quad j \neq i$$

The forms of the functions h_1, h_3 and h_4 are somewhat involved and are dealt with in [1] . They ensure that the estimators are unbiassed.

Tables 1 and 2 present results of classification obtained by observing the highest value of the DF for every utterance-vector.

		LR	FR	HR	LF	FF	L	LRF	FRF	
	LR	92	1	2	0	0	5	0	0	
a	FR	3	83	15	0	0	0	0	0	a
a	HR	5	17	78	0	0	0	0	0	a
E	LF	0	0	0	84	9	0	5	3	E
J	FF	0	0	0	7	89	0	0	5	J
L	L	5	0	0	0	0	95	7	0	L
L	LRF	0	0	0	12	2	0	82	4	L
d	FRF	0	0	0	1	12	0	4	83	d

classified

Table 1. Results of classification with EQDFs. The figures are percent scores.

		LR	FR	HR	LF	FF	L	LRF	FRF	
	LR	87	1	3	0	0	9	0	0	
a	FR	4	77	19	0	0	0	0	0	a
a	HR	3	25	72	0	0	1	0	0	a
E	LF	0	0	0	81	9	1	7	2	E
J	FF	0	0	0	7	82	0	0	11	J
L	L	6	6	0	0	0	94	0	0	L
L	LRF	7	0	0	12	3	2	80	3	L
d	FRF	0	0	0	1	11	0	4	83	d

classified — presented

Table 2. Results of classification with ELDFs. The figures are percent scores.

It can be seen from Tables 1 and 2 that (1) For all 8 patterns, EQDFs give better classifications than do ELDFs. (2) There is some confusion among the three Rises, between the two Falls and between the two Rise-Falls. (3) The Level and the Low Rise are sometimes mutually confused. (4) There is mutual confusion between the Falls and the Rise-Falls.

The overall results are 85.7% correct classification with the EQDFs and 81.8% correct with the ELDFs.

When the results of the classification of the individual vectors were compared between the two DFs, it was found that in 78.5% of the cases both gave correct and in in 10.9 % both gave the same incorrect assignment. The two methods gave different classification results 9.6% of the time.It is clear therefore that neither one nor the other kind of hypersurfaces separating the eight classes could be perfectly fitted to the entire data.However, a large proportion of the discrepancies between the results obtained by using the two DFs was due to the fact that the final decision was practically a random choice between two of the eight classes. As mentioned above, of the eight EQDFs and eight ELDFs it is those with the highest value that indicate the final assignment. We shall consider two cases here. For one imitation of a Full Rise (voice MC), the following DF values were obtained:

	EQDF	ELDF
LR	- 5.02	-10.50
FR	0.35	- 5.03
HR	1.56	- 4.39
LF	-41.55	-45.74
FF	-59.06	-57.53
L	-26.95	-16.54
LRF	-33.56	-43.35
FRF	-50.82	-54.09

Both functions have the highest values at HR, so both ways the particular expected FR was classified as HR. But in both columns, the difference between the values in the FR and HR rows are distinctly smaller than any other differences. So the ultimate decision between HR and FR is frail. In another case an HR imitation was classified as FR by the quadratic, but as HR (i.e., correctly) by the linear function:

	EQDF	ELDF
LR	- 7.43	-10.87
FR	1.41	- 4.70
HR	1.13	- 4.20
LF	-44.21	-46.85
FF	-59.06	-58.89
L	-26.95	-16.58
LRF	-33.56	-41.94
FRF	-50.82	-51.25

Again, the differences between the two highest values are much less than those between any of the remeining ones. Thus, even a correct decision is not convincing. Indeed, the two utterances were represented by the follwoing vectors (raw data, successive average frequencies in Hz):

(1)[220, 218, 210, 206, 258, 274, 300, 320]
(2)[209, 210, 205, 235, 273, 286, 308, 313]

There is nothing to indicate that the two

sequences of FØ values (or the two corresponding FØ contours) represent two different Rises. Many of the misassignments were of this kind, which is a strong indication that the misclassifications were largely due to an overlap between the 8 classes of utterance-vectors.

INTERSPEAKER DIFFERENCES

When the results of the classification were considered separately for each speaker, the following scores were obtained (percent error for EQDF, with ELDF results in parentheses):

Numbers indicate percent error

WJ Ø
IŁ Ø
LR 1 (1)
JI 4 (7.5)
BS 11 (16)
MC 11 (17)
AM 14 (17.5)
KK 14 (21)
HK 15 (20)
MK 20 (24)
PD 21 (25)
BI 25 (29)
TK 25 (22.5)
CW 25 (30)
BS 26 (38)

The speakers can be seen to have performed quite unequally. The top four speakers were phonetically trained. The remaining ones were all naive speakers. Should the 8 assumed classes of utterance-vectors be completely distinct at the linguistic level, one should have expe ted better individual scores. On the other hand, should they only be perceptually distinct after phonetic training, there would have been less variation in the scores of the 11 untrained subjects. A conclusion that suggests itself from these results is that though the 8 classes can be distinguished at the linguistic level, the differences between some of the classes are not entirely categorical.

SIMILARITIES BETWEEN THE CONTOURS

When the values of the DFs are arranged from the highest to the lowest, the relative similarity of each token to the eight patterns can be judged, the second highest DF indicating the most similar and the last, the most dissimilar pattern. The strength of the similarity and the dissimilarity in our entire materials may be evaluated by considering the number of times that the particular class (pattern) was indicated by the second-highest and the lowest DF. We shall here take into account the quadratic functions only. The results may be summarized as shown in Table 3.This Table contains, in the successive columns, the following:

1. The recognized pattern
2. The most similar pattern
3. The number of cases in which the pattern indicated as most similar actually occurred as the second-highest EQDF
4. The most dissimilar pattern.
5. The number of cases in which the pattern indicated as the most dissimilar actually occurred as the last EQDF. It is to be understood that other patterns occurred in the second and in the last places less frequently than indicated in the Table.

TABLE 3.

1	2	3	4	5
recog.	sim.	freq.	dissim.	freq.
LR	L	77	LRF	94
FR	HR	119	FRF	95
HR	FR	110	FF	77
LF	LRF	88	FR	94
FF	FRF	70	L	59
L	LR	110	FF	60
LRF	LF	61	HR	123
FRF	FF	66	LR	113

The following conclusions can be drawn from the results summarized in Table 3: (1) All the similarities are reciprocal. (2) The dissimilarities are mostly not reciprocal. (3) There is strong similarity between HR and FR, beween L and LR, and there is somewhat weaker similarity between the Falls and the corresponding Rise-Falls (Low with Low and Full With Full). (4) There is strong dissimilarity between the Rise-Falls and the Rises. (5) There is distinct dissimilarity between FF and L.

The similarities and dissimilarities among the 8 patterns may be studied in some more detail by considering also the third highest and the second-last DF. The results of such a study can best be shown by a three-dimensional bar graph like the one in Fig. 2., which by way of an example, refers to the 150 cases of (assumed) HR. The horizontal axis refers to the order of the DF. The highest-value DF, which indicates the assignment to a class, is No.1. The second highest, indicating the strongest similarity, is No.2.The third highest DF, referred to by No.3 shows second-order similarity.. Positions 4,5 and 6 are not very informative. No.8 is the strongest dissimilarity and No. 7 the second-order dissimilarity. The Figure shows that both FR and LR are similar to HR and that FRF and also FF are dissimilar to it.

REFERENCE

[1] G. DEMENKO, W. JASSEM & M. KRZYŚKO: Classification of basic FØ patterns using discriminant functions (forthcoming).

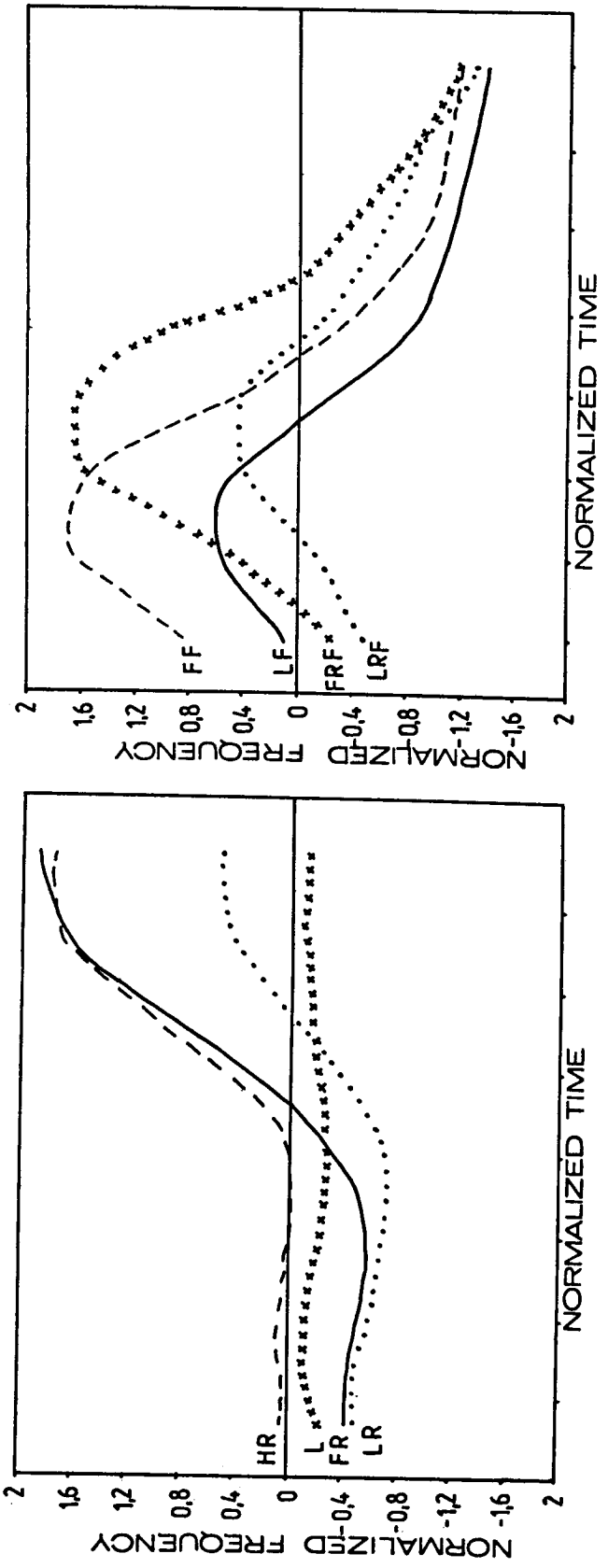


Fig. 1. The eight averaged F_o patterns.

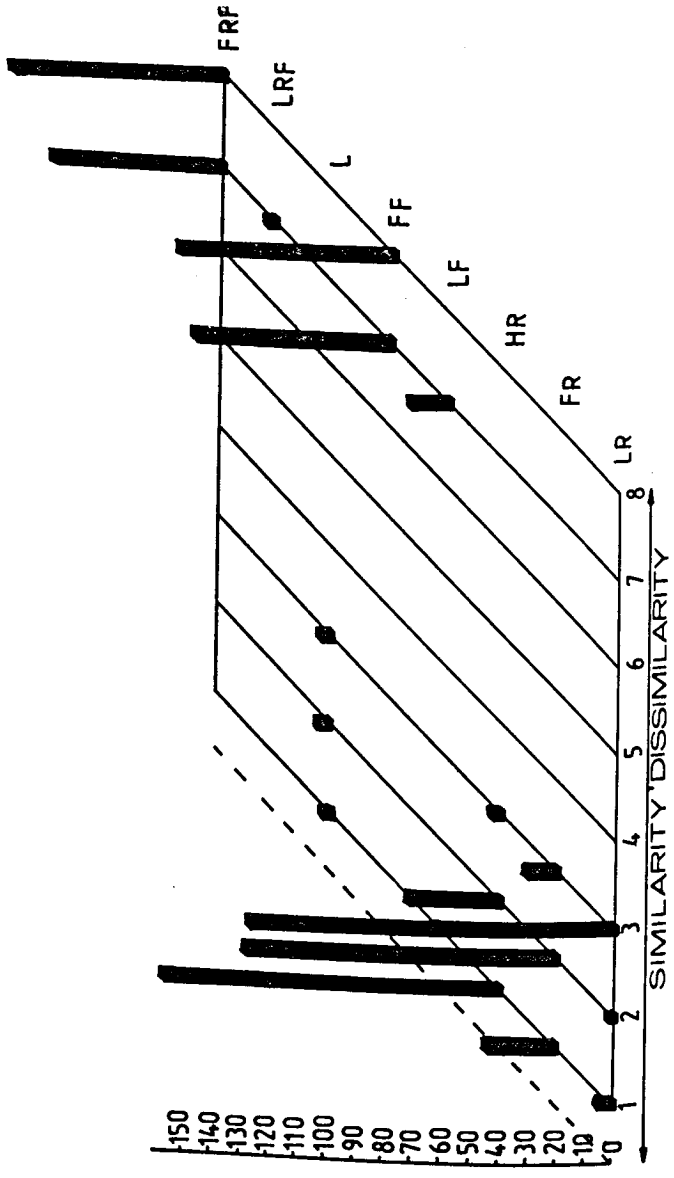


Fig. 2. The 150 tokens of HS with their similarities and dissimilarities.

SEPARATE PITCH AND RHYTHMIC PATTERNS IN SYNTHETIC SPEECH AND MUSIC

JAN TRO

The Norwegian Institute of Technology, Acoustics
N-7034 Trondheim-NTN, NORWAY

ABSTRACT

For Text-to-Speech synthesizers normally the text is manually prepared to add some prosodic information. This is rather time consuming and requires an expert user of the system. Even if speech and music seem to be closely related as acoustical signals, we need different procedures to add melodic and prosodic information into a music and a speech synthesizer, respectively. The strong relation of syllable duration, pitch and sound level in Norwegian dialects seem to complicate the procedure of adding humanlike prosody to synthetic speech.

This paper deals with aspects of similarities between speech and music signal. The aim is to contribute to the simplification of adding Norwegian prosodic features to speech synthesizers.

INTRODUCTION

Prosodic parameters as syllable duration, pitch and sound level are closely linked variables in Norwegian dialects. This fact complicates the procedure of adding humanlike prosody to synthetic speech.

As long as the intonation in some Norwegian dialects is very important to establish the correct meaning of an utterance, we need simple methods for adding sufficient prosodic information to synthesizers.

If we consider both music and speech as elements in the process of human communication, both signals may be described by a set of acoustical parameters in the frequency and the time domain. It is very easy to discuss musical details (pitch, rhythm, level) as near independent parameters.

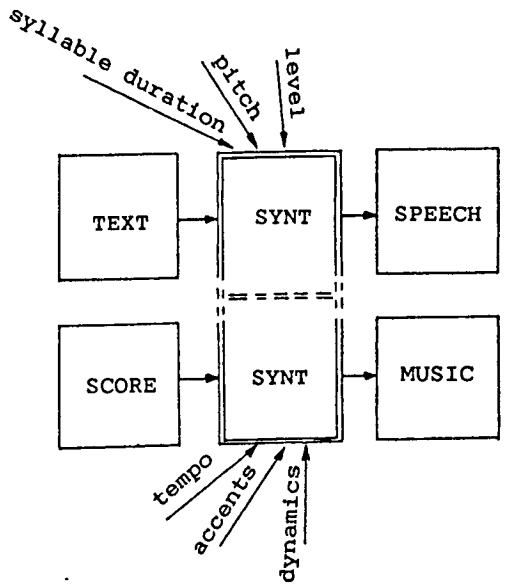


Figure 1: Sound synthesis with additional information.

Even if speech and music seem to be closely related as acoustical signals, we need different procedures to add melodic and prosodic information into a music and a speech synthesizer, respectively.

Could it still be possible to add prosodic information factor by factor, - in order to obtain increased intelligibility, understanding and naturalness?

SPEECH AND MUSIC

It is not a quite new idea to compare features in speech and music. In 1916 Alnæs [ref. 1] used the standard musical notation in his book to described the variation of pitch, - with minor success.

Nevertheless there is obviously a need for some kind of graphic notations in order to describe prosodic information in detail. The more common used system with dots and lines for unstressed and stressed syllables seems to give a suitable precision in the description of pitch variation. This system is still very close to the musical notation with the traditional five lined staff with dots (tones) of different duration.

Rakowski [ref. 3] has considered music as one part in the process of human communication. In this way it is possible to discuss music on the basis of information theory. The distinction of sensory features which differentiate musical intervals seems to be bases on the same principle as with vowels of speech. That means we are dealing with a set of cognitive prototypes of tone color in the case of vowels and prototypes of interval size in the case of musical intervals. The reason why the number of phonological units of speech and music is not very high (30-40 phonemes, 12 equally tempered intervals per octave), is probably more because of the limited capacity of memory than due to the discriminatory properties of hearing.
This approach is indeed very promising when comparing speech and music signal.

However, it is possible to make a shorter way to this comparison saying that the intonation of speech, the "melody of speech", is the aspect which comes closest to music [ref. 7]. This is an acceptable assumption as long as "music" is restricted to comprise melodic lines only. Let us have a closer look at basic parameters in melodic lines.

Pitch
The limited collection of notes in traditional western melodic lines consist of a set of fixed tones positioned on the well-known five lined staff. It is very tempting to use a similar notation system for the pitch variation of syllables. One model is proposed by Sirnes [ref. 4]. However, some problems will arise due to the fact that music notation is normally designed for instruments with fixed pitch. All your practicing and rehearsals will lead to an even closer connection to the accepted collection of tones. This is a genuine situation for the performance of classical western music. The attempt to use a set of fixed tones in speech will give the impression of trying to recite in a singsong manner.

Sirnes has proposed some modifications for his model in order to implement nonlinear effects (nonsymmetrical syllable transitions, avoid stationary pitch, etc.). This will probably contribute to increased naturalness of the synthesis.

As far as music concerns, removing the fixed pitch is fatal to the perception of melody. Even the fixed intervals is a basic property of music in contradiction to normal speech.

Is it possible to deal with pitch as an independent parameter?

It is not possible in music because the combination of time and pitch information is necessary for the correct perception of melodic lines. Although the intelligibility of speech will remain unaffected, the lack of time information will dramatically decrease the naturalness of Norwegian dialects. Analysis of Norwegian native dialects show strong bindings of syllable pitch and duration as prosodic elements. We have to remember that pitch changes may even change the meaning of the utterance.

Duration
Deviations in the time domain will affect music and speech differently. In melodic lines both relative time and absolute duration of notes are very important in order to obtain a proper rhythmic percept. Any changes in the microstructure of the notes (short/long attack time, decay time, sustain time and release time) will normally affect the perception of timbre keeping the melody still intact.

Deviations of syllable start time and duration will not have a corresponding fatal effect. As minor deviations will lead to decreased naturalness, only larger deviations may result in decreasing intelligibility.

Is it possible to deal with tone/syllable duration as an independent parameter?

The strong bindings of pitch and duration in the constructions of melodic lines is well-known and do not invite us to get rid of the pitch information. However, if we need to analyse or synthesize the rhythmic patterns it is a good idea to concentrate on the time domain exclusively. This holds for speech synthesis as well.

If we remove the pitch information (meaning the voiced sounds) the utterance may still be accepted as natural, namely as whisper! In this way it is possible to deal with syllable duration in forming rhythmic patterns of utterances without being confused by the pitch contour. This principle has been tested and reported by Ottesen [ref. 2].

Intensity
The recognition of melodic lines are nearly affected at all by variation in the tone intensity. However, as part of the dynamic musical structure it is very important.

In Norwegian dialects syllable intensity plays a minor role compared to pitch and time information in the prosodic structure.

Some valuable information on the construction of the syllable envelope curves and the transition from one curve to another, may be found in the area of musical synthesis.

Speech quality
The discussion and definition of speech quality is of vital importance for the evaluation of synthetic speech. This discussion has to include aspects of intelligibility, sense of utterance, and naturalness. We have to define what is an acceptable, expected, and a sufficient level of quality.

As we have seen, the quality of spoken Norwegian dialects is affected by variations in pitch, duration and intensity. Even feelings may be expressed by some changes in the prosodic features. In some cases large prosodic differences (like in different Norwegian dialects) may be evaluated as natural. But only minor prosidic variations may cause in a fatal decrease in naturalness when it is spoken by a non-native tongue. This means that adaptational processes will influence the evaluation of quality.

It is important to establish references for all the different levels of synthesized speech quality.

LISTENING TESTS

Experiences of musical listening tests on pitch and rhythm has been an advantage in designing our laboratory tests.
Through different experiments we try to isolate independent prosodic elements which can easily be added separately to the speech synthesizer in order to obtain more humanlike speech sounds.

Listening tests has included recognition of melodic lines, separation of pitch and time information in melodic lines, effects of prosodic variations in a Norwegian dialect, and pitch changes as a result of intensity variation.
Results from these tests will be presented.

REFERENCES

[1] Alnæs, I.:
Norsk sætningsmelodi,
(Norwegian Prosody)
Kristiania, 1916. (in Norwegian)

[2] Ottesen, G.:
Adding natural prosody to a phoneme synthesiser,
11th PCPhS, Tallin, 1987.

[3] Rakowski, A.:
Acoustics of music and acoustics of speech. Some common factors.
Proceedings of 23rd Acoustical Conference on Physiological and Psychological Acoustics, Acoustics of Speech and Music. Ceske Budejovice, 1984.

[4] Sirnes, G.:
Analyse og syntese av norsk setningsprosodi. (Analysis and synthesis of Norwegian Prosody) Thesis.
The Norwegian Institute of Technology
Trondheim, 1986. (in Norwegian)

[5] Tro, J.:
Aspekter ved lytting til levende og teknisk gjengitt musikk. (aspects of Listening to Live and Recorded Music)
ELAB memo 44-AN85030.
Trondheim, 1985. (in Norwegian)

[6] Vanvik, A.:
Norsk fonetikk. (Norwegian Phonetics)
The University of Oslo,
Oslo, 1979, (in Norwegian)

[7] Vanvik, A.:
Reflections on the relation between speech and music.
In Fretheim (ed): Papers from a Symposium. Nordic Prosody II.
Trondheim, 1980.

ADDING NATURAL PROSODY TO A PHONEME SYNTHESISER

GEORG E. OTTESEN

Acoustics Research Center, ELAB
N-7034 Trondheim-NTH, NORWAY

ABSTRACT

This work investigates the possibility of increasing the quality of synthetic speech by adding some timing information. The rhythm of the syllables is tapped onto the keyboard of a computer. The vowel and consonant durations are modified by rules to fit the rhythmic pattern given. The result is compared to speech synthesis based on standard phoneme lengths and to synthesis with phoneme lengths aligned with actual speech. Listening tests are performemd on Norwegian sentences synthesised in whisper.

INTRODUCTION

Text-to-speech synthesisers and phoneme synthesisers are not used for public services in Norway. They are judged to sound very unnatural. The main shortcoming is the lack of natural prosody. Norwegian has a complex prosodic system. There are two distinct word tonemes, the timing pattern is mainly stress based, and the overall sentence prosody is strongly dependant on the syntactic and semantic structure. The natural prosody can not be derived automatically from text without advanced methods of sentence analysis.

In many applications the text can be prepared by adding some prosodic information. This work investigates a method for adding the timing information to sentences by tapping the rhythm of the sentences onto the keyboard of a computer. The sentences are synthesised in whisper, which isolates the timing information from the pitch information. The synthesis of pitch contours will, therefore, not be discussed in this paper.

METHOD OF TIMING

Experiments on the rhythm of German speech indicate a close relationship between the perceived rhythm of speech and the point of onset of the vowel in each syllable [1]. This fact is used to add a natural timing to a speech synthesiser by tapping the rhythm of the syllables onto the keyboard of the computer. Simple rules are given to align the synthetic speech with the keystroke sequence.

The synthesiser is a 4-formant phonetic synthesiser with Norwegian phonemes. Each phoneme corresponds to 1, 2 or 3 phonetic elements in the synthesiser. Vowels consist of one element, diphthongs of two, and unvoiced plosives of three elements. The durational resolution of each element is 10 ms. Each phonetic element has a default duration which is the mean duration of that element in actual speech.

The rhythm of the syllables is tapped on the keyboard of a personal computer by using two fingers to operate two keys; this method gives a better timing than tapping with one finger only. The sentence is spoken by the person simultanously with the tapping. Each tap corresponds to the starting point of the vowel in the syllable. The vowel or diphtong and the following consonants make up the time interval between two keystrokes.

Initial experiments show that a linear scaling of the elements of each syllable is not acceptable. The burst of the unvoiced plosives /p,t,k/ cannot be stretched significantly without losing the impression of a burst. A major increase in the duration of the short vowels creates confusion between short and long vowels, as in the words /kane/ and /ka:ne/. Other elements can be prolonged by a factor of 3 between slow and fast speaking rate. These observations lead to the introduction of a stretching factor for each phonetic element. This factor is called the time warping sensitivity, s_i, of the phonetic element, i. Each phonetic element is given an additional length, ΔT_i, which is proportional to the default duration, T_i, to the time warping sensitivity, s_i, and to the difference of actual syllable length, T, and the sum of the default lengths:

$$\Delta T_i = \frac{s_i \, T_i \, (T - \Sigma T_i)}{\Sigma s_i T_i}$$

Figure 1 gives an illustration of this alignment.

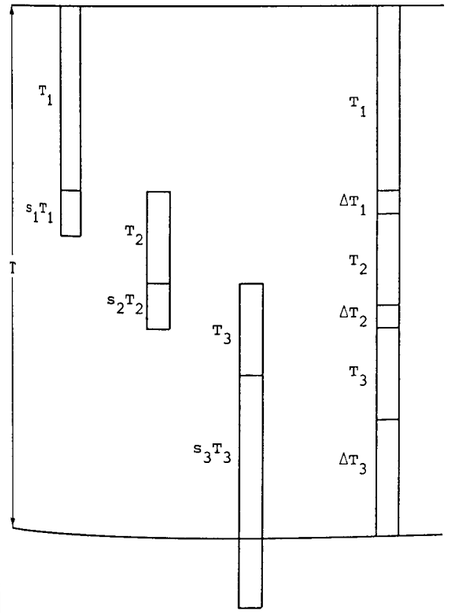


Figure 1. Alignment of phoneme durations.
T - time gap between vowel onsets
T_i - phoneme default duration
s_i - time warping sensitivity
ΔT_i - additional duration

The numerical range of the warping sensitivities is arbitrary; a constant factor does not change the timing. The range is chosen to be [0, 100]. A warping sensitivity of zero means that the element can not be stretched, one of 100 means that the element is maximally stretched. Phonemes with similar properties are ordered in groups with the same sensitivity. Long vowels have maximum time warping sensitivity. Nasals and the closure phase of unvoiced plosives have medium sensitivity, short vowels have low sensitivity and the burst of the unvoiced plosives have minimum sensitivity. The values are adjusted to give an acceptable pronounciation of single words spoken slowly and quickly.

AUTOMATIC TIMING

The work on manual timing of synthetic speech is a step towards an automatic timing procedure. The next step will be to mark the stressed syllables of the sentence and let that form the basis of the synthesis of prosody.

The duration of consonant clusters presents an additional problem. It is well known that consonants in clusters normally are shorter than single consonants. Our first algorithm for automatic timing assigns a syllable interval which is a linear combination of the sum of the element default durations, and a syllable default duration. Using a syllable default duration only would mean making the intervals between the vowel onsets equal. A simple listening test was performed to find the best balance between these two factors. A Norwegian sentence with large consonant clusters was randomly presented with different timing parameters, and the listeners were asked to choose the sample with the most natural rhythm. The test was repeated with several sentences and several listeners. The preferred timing consists of 80% of the element default durations and 20% of a syllable default duration.

LISTENING TEST

As this work only considers the timing of sentences, all the synthesis is made in whisper. Three different methods of sentence timing are compared:

1) Manual timing of vowel onsets,
2) Automatic timing based on element default durations combined with a syllable default duration,
3) Element timing aligned with human speech.

The test is designed to measure the perceptual distance between these three methods. The listeners are asked to pick the sentence with the most natural rhythm from a presented pair. Results will be presented at the conference.

REFERENCES

[1] W. Heinback:
Rhythmus von Sprache: Untersuchung methodischer Einflüsse.
Proceedings of DAGA'85, Stuttgart

SPEECH SYNTHESIS OF SENTENCE FOCUS IN ENGLISH DECLARATIVES

S.J. EADY and B.C. DICKSON

Centre for Speech Technology Research, University of Victoria,
P.O. Box 1700, Victoria, B. C., V8W 2Y2, Canada

ABSTRACT

This paper describes an algorithm that has been developed to synthesize sentences containing focused words in varying locations. Synthetic speech is generated by concatenating LPC-encoded words and phrases of English. The method involves the assignment of one of three accent levels to each syllable of all encoded vocabulary items. The highest accent marking for a given word depends on its grammatical category. The accent assignments are then used to determine the relative pitch for each word in a synthesized sentence. A method is described for using this system of pitch assignment to synthesize sentence focus in English declaratives. The pitch assignment algorithm is currently being used in applications requiring automated speech output.

INTRODUCTION

The use of synthesized speech for applications such as computer-assisted language instruction [1] and automated information delivery [2] requires a synthesis system that is capable of generating utterances in which the location of sentence focus may vary. For example, the English sentence "I LIKE TO PLAY FOOTBALL" would be produced in a different way for each of the following contexts:

1. What do you like to play?

2. Do you like to play football or watch football?

3. Do you like or dislike playing football?

For each of these questions, a different word would be focused in the declarative response. The ability to convey this difference in the location of focus is an important aspect of a speech synthesis system.

Previous work on the production of sentence focus in English [3-5] shows that the focusing of a word results in changes to the durations and pitch patterns of the sentence components. In this paper, we describe how the acoustical patterns of sentence focus are generated using a word-concatenation synthesis method [6,7].

SPEECH SYNTHESIS METHOD

Synthesized speech is generated on a microcomputer using a Texas Instruments TMS-5220C speech synthesis chip. A control program is used to provide the synthesis chip with a series of quantized values for pitch, energy and ten LPC reflection coefficients. These parameter values are stored as preprocessed vocabulary items corresponding to individual words or short phrases. English sentences are synthesized by concatenating these preprocessed vocabulary items in a specified order and then applying rules to produce appropriate pitch patterns and to eliminate spectral discontinuities at word boundaries.

Vocabulary Production

Vocabulary items required for a particular application are first embedded in carrier sentences and read by a male speaker whose voice is recorded on a digital PCM recorder. Each item is then digitized (at a 10-kHz sampling rate with 12-bit resolution) and excised from its sentence environment. The digitized vocabulary items are then analyzed using the autocorrelation method of LPC [8] to derive values of energy, pitch and 10 LPC reflection coefficients at 20-msec intervals. These parameters are quantized for output on the synthesis chip.

Each encoded vocabulary item is then edited to eliminate any spectral discontinuities, and to provide a uniform energy maximum and a neutral pitch contour. During the editing process, every syllable is assigned one of three levels of "accent". An accent marker is assigned to the "nuclear" frame of each syllable (usually the frame containing the highest energy in the syllable). In general, the primary accent level is assigned to the highest stressed syllable of nouns, adverbs and adjectives, whereas verbs are marked with secondary accent, and the third level of accent is assigned to function words, such as conjunctions, prepositions and modal verbs.

Word Concatenation

At the time of synthesis, encoded vocabulary items are concatenated to form complete sentences of English. The input to the system is standard English spelling augmented by diacritics to specify sentence type (i.e., statement or question), the identity of any focused words, and the location and duration of any pauses within a sentence. The system verifies the existence of each word in the list of encoded vocabulary items, and the requested items are joined together in the order specified. At this point, the encoded vocabulary items are modified by a number of phonetic liaison rules (described elsewhere [7]) and by a set of rules for pitch assignment.

Pitch Assignment

The pitch assignment rules utilize the accent levels assigned to each vocabulary item to adjust the pitch level of each word in a sentence. For the purposes of the pitch assignment rules, the sentence is divided into two major components, called the "head" and the "tonic" (following Halliday [9] and Young and Fallside [10]). The head comprises all syllables from the start of the sentence up to and including the penultimate primary-accented syllable. The tonic is made up of the final primary-accented syllable and any other following syllables in the sentence.

Head. Within the head, the pitch rules utilize the accent level assigned to each vocabulary item to determine the highest pitch with respect to a predetermined fundamental frequency (F0) topline, midline and baseline. Each of these lines has a gradual F0 declination over the duration of the sentence head. These three lines are used to determine the relative pitch level of each word in a synthesized sentence. Vocabulary items containing a primary accent have their maximum pitch on the topline; those with secondary accent are set to the midline; and words with only tertiary-accented syllables fall to the baseline. This method maintains the overall shape of the pitch contour that was given to each vocabulary item during the original recording procedure, while at the same time modifying the relative pitch level for each word with respect to other vocabulary items in the sentence.

Tonic. A special series of rules applies to the last primary-accented syllable in a sentence. This syllable, known as the tonic, is set apart from the sentence head and is not bound by the limits of the topline and baseline. Instead, it is given a distinctive pitch contour depending on the sentence type. If the tonic occurs at the end of a declarative sentence, it is given a falling pitch contour. For sentences requiring an interrogative intonation pattern (i.e., yes-no questions), the tonic is given a rising pitch contour.

The difference between the tonic pitch contours for declarative and interrogative sentences is shown in Figure 1. The pitch display at the bottom of the figure illustrates the difference between the declarative tonic (on the words PREFER and SOCCER) and the interrogative tonic (on the word FOOTBALL).

Synthesis of Sentence Focus

The strategy for synthesizing sentence focus is based on recent studies [3-5] showing that focused words in declarative English sentences are characterized primarily by two factors. These are an increase in the duration of the focused item and a relatively low, flattened pitch contour for that part of the sentence following the focused word. In attempting to synthesize sentence focus, we have found that the second factor (i.e., the lowering of the post-focus pitch contour) is more important than the first for focus perception.

The method that we have used to synthesize sentence focus utilizes the accent levels that are assigned to the words of an utterance, as described above. The strategy is to assign the focused word an accent level of 1, and to reduce any primary-accented syllables that follow the focused word to a level of 2. The result is to shift the tonic syllable of the sentence so that it occurs within the focused word. Consequently, the focused item coincides with a local peak in the pitch contour and is followed by a relatively low flattened pitch for any words that follow the focus. The effect of this algorithm on a sentential pitch contour is illustrated in Figure 2. This figure shows how the pitch contour for a sentence is modified when the focus is shifted from the end to the middle or to the beginning of the utterance.

Informal listening tests indicate that this strategy of modifying only the pitch contour of a sentence is successful in eliciting the required focus perception. In some cases, it is also useful to provide an increase in the duration of the focused word, and this option has been incorporated into the speech synthesis system. In general, however, the modification of the sentential pitch contour is usually sufficient to produce the desired effect.

CONCLUSION

The method described here for synthesizing sentence focus in English declaratives has been developed for use in a word-concatenation synthesis system. In this system, the pitch contour for a synthesized sentence is determined by preassigned accent levels associated with each word. The location of focus within a sentence is overtly specified at the time of input. The sentence focus algorithm exploits the presence of the preassigned accent levels and modifies the sentential pitch contour to produce the desired focus effect.

The word-concatenation synthesis system is presently being used in applications requiring automated speech output. The pitch assignment rules described here are also being incorporated into a demisyllable-based text-to-speech synthesis system that is currently under development.

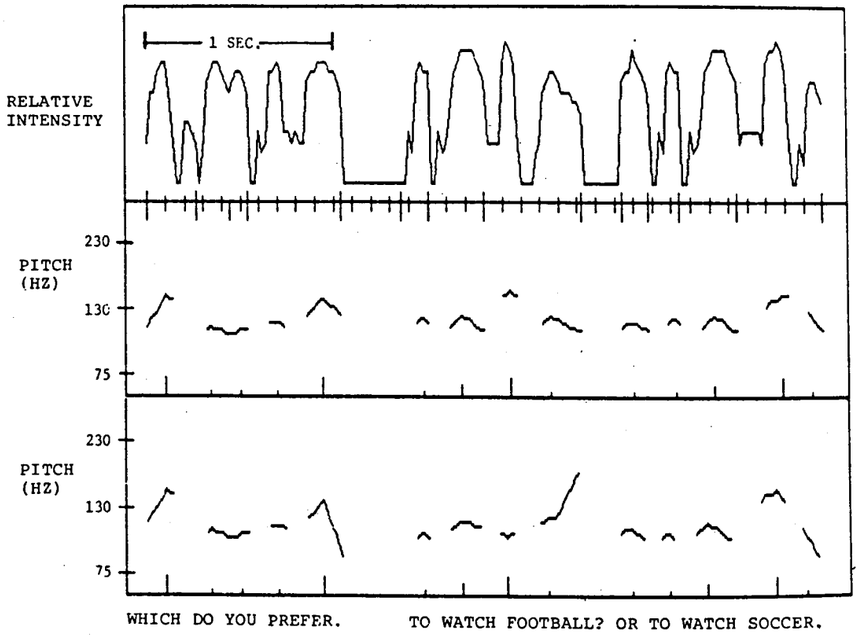


FIGURE 1: Display of relative intensity and pitch contours for a sentence synthesized by the word-concatenation method. The uppermost pitch display shows the pitch contours for the concatenated vocabulary items prior to the application of the pitch assignment rules. The bottom pitch display shows the modifications produced by the pitch assignment rules. Note the vertical markers located along the bottom of each pitch display, which indicate the accent level that has been assigned to each syllable. The tallest markers denote syllables with primary accent; markers of intermediate length are for secondary accent; and the shortest markers designate syllables with tertiary accent.

ACKNOWLEDGEMENT

This work was funded in part by the Science Council of British Columbia and by the Natural Sciences and Engineering Research Council of Canada. We thank Allan Wynrib, Jocelyn Clayards and Sue Urbanczyk for technical assistance.

REFERENCES

[1] Eady, S.J., Dickson, B.C. and Walraven, J. (1987). "Use of speech synthesis for language instruction on a microcomputer," Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, in press.

[2] Dickson, B.C., Eady, S.J., Clayards, J.A.W., Urbanczyk, S.C. and Wynrib, A.G. (1987). "Use of speech synthesis in an information system for handicapped travellers," Proceedings of the 11th International Congress of Phonetic Sciences, in press.

[3] Cooper, W.E., Eady, S.J. and Mueller, P.R. (1985). "Acoustical aspects of contrastive stress in question-answer contexts," J. Acoust. Soc. America, vol. 77, pp. 2142-2156.

[4] Eady, S.J. and Cooper, W.E. (1986). "Speech intonation and focus location in matched statements and questions," J. Acoust. Soc. America, vol. 80, pp. 402-415.

[5] Eady, S.J., Cooper, W.E., Klouda, G.V., Mueller, P.R. and Lotts, D.W. (1986). "Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments," Language and Speech, in press.

[6] Olive, J.P. and Nakatani, L.H. (1974). "Rule-synthesis of speech by word concatenation: A first step," J. Acoust. Soc. America, vol. 55, pp. 660-666.

[7] Eady, S.J., Dickson, B.C., Urbanczyk, S.C., Clayards, J.A.W., and Wynrib, A.G. (1987). "Pitch assignment rules for speech synthesis by word concatenation," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, in press.

[8] Markel, J.D. and Gray, A.H. (1976). Linear Prediction of Speech (New York).

[9] Halliday, M.A.K. (1967). Intonation and Grammar in British English (Paris).

[10] Young, S.J. and Fallside, F. (1980). "Synthesis by rule of prosodic features in word concatenation synthesis," International Journal of Man-Machine Studies, vol. 12, pp. 241-258.

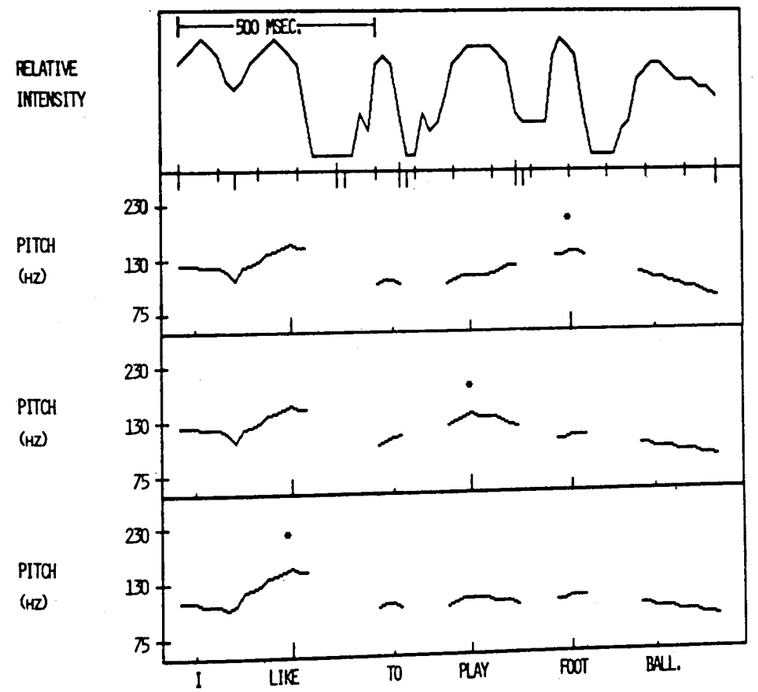


FIGURE 2: Display of relative intensity and pitch contours for three synthesized versions of the same sentence. The asterisk in each pitch display indicates the location of focus in each version, as described in the text. As in Figure 1, the vertical markers located along the bottom of each pitch display indicate the accent level assigned to each syllable. Note how the focused word in each version always corresponds with a primary-accented syllable (as denoted by the tallest accent markers), whereas all post-focus words have lower accent levels.

INCREASEMENT OF NATURALNESS IN SYNTHETISIZED SPEECH

EINAR MEISTER
Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

MART ROHTLA
Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

MAIDU RAUDSEPP
Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

ABSTRACT

An algorithm for synthesizing pitch contours is presented. It is shown that the relationship between the fundamental frequency and the frequency of the first formant of stressed vowel should be simulated. The problem of naturalness of synthesized speech is discussed.

INTRODUCTION

The text-to-speech systems become more and more used nowadays. The demands on the quality of the synthesized speech differ greatly. It has been found out that under certain conditions the synthesized speech has even advantages over natural speech because of its machine-like quality attracting the attention of the listener [1]. Although the systems vary in their capability and structure, for those who use them the most important thing is their intelligibility and the level of naturalness.

One of the main drawbacks of the synthesized speech is prosody, i.e. the changes of intonation, intensity and duration which are not enough controlled. According to many authors the investigation and the modelling of intonation is the key-problem in increasing the quality of the synthesized speech.

When investigating the changes in the pitch of natural speech, two characteristic components could be brought out: firstly, slow and large changes of the pitch (intonation contours) and secondly, fast and small changes of speech (the "fluctuation" of pitch). Both the components are intrinsic for the speech signal and their absence causes losses in the quality of the synthesized speech.

The movement of the pitch has been investigated by many authors already for a number of years. Various algorithms for the modelling of the intonation contours have been created [2,3,4] and the micromelody in different contexts has been studied [5,6].

When describing the intonation of Russian, the intonation contours of syntagma are widely accepted units [7]. Syntagma is the minimal prosodical unit which could still be divided into the following functional parts: (i) precentre, (ii) centre, (iii) postcentre. A special role in a syntagma is played by the so called intonation centre (by the intonation centre we mean the most important word of the syntagma) because the changes of the pitch in the centre are the most important feature in distinguishing different intonation types.

In speech the intonation is organically connected with other components of the signal. There are two types of components in a speech signal. On the one hand, there are components controlled by the articulatory program and on the other hand, components depending on the structure of articulatory organs. It may be assumed that the naturalness of human speech depends on interdependent parameters of the signal, i.e. a change in a certain parameter brings about a change in some other parameter. In other words, there exists a principle of "integral unity" between the parameters of the signal.

The aim of our research is to increase the naturalness of the synthesized speech through the modelling of the intonation contours; to point out the possible causes of machine-like sound of the synthesized speech and to introduce the connections between the parameters of a speech signal according to the principle of the "integral unity".

THE ALGORITHM FOR THE SYNTHESIS OF THE INTONATION CONTOURS

Under the synthesis of the intonation contours we mean the creating of the control parameters .for the pitch generator which are based on the duration of the synthesized message, punctuation marks, word stress and sentence stress.

Pitch generator

The control parameters for the pitch generator are the fundamental frequency and the time of transition from one frequency to another (duration). The typical range of the fundamental frequency for the male voice is 80-180 Hz. During the synthesis this range is divided into 8 levels; as for the time of transition, it is sufficient to have 4 meanings in the range of 50 up to 300 ms. It is also possible to choose the form of glottal pulse.

The models of intonation contours

There are various descriptions of intonation contours in Russian [8]. In choosing the models for the present paper we used the descriptions given in [7,9]. At the present moment the declarative, the interrogative, the nonterminal and the exclamatory models have been realized.

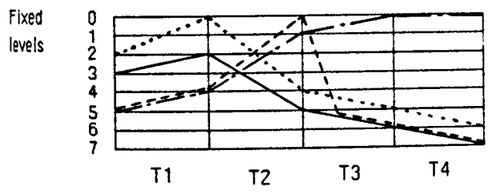


Fig. 1. The models of intonation contours.
—— a declaration,
--- an interrogation,
—·— a nonterminal,
····· an exclamation,
T1 - precentre,
T2 - centre,
T3 - postcentre,
T4 - the end of the syntagma.

Initial data

When developing the algorithm for the generation of intonation contours the following initial data was taken into consideration:
- the input text should be supplied with word-stress marks (') and with sentence-stress marks ('');
- in order to distinguish between different types of intonation the following punctuation marks are used at the end of a syntagma:
  [.] - a declaration,
  [,] - a nonterminal,
  [?] - an interrogation,
  [!] - an exclamation;
- the input data for the algorithm is a sequence of elements where each byte carries information about the duration of the element, about the stress and about the end of the syntagma;
- the possibility of changing the existing models and of adding other models of intonation contours corresponding to other punctuation marks;
- when describing the intonation contours no knowledge about programming is needed;
- the minimizing of the working time of the algorithm and the capacity of the memory;
- the realization of the algorithm using a 8-bit microprocessor I8080.

The description of the algorithm

The algorithm functions in 3 steps:
(i) the punctuation mark determines the type of the model of intonation contours of the syntagma and the durations T1, T2, T3, T4 are computed;
(ii) depending on the stress-mark the duration of the segments is determined during which the fundamental frequency changes with a constant rise;
(iii) the approximation of the fundamental frequency with linear cuts, i.e. the control parameters of the F0 generator are computed.

The algorithm needs about 1 KByte memory and works in real time.

Auditory estimation

In order to estimate the efectiveness of the algorithm, sentences consisting of one and two syntagmas with all types of intonation contours were synthesized. The type of intonation was in most cases distinguished correctly by the listeners and the intonation contours were said to correspond satisfactorily to those of human speech. At the same time the synthesized speech as a whole still had a machine-like sound.

Thus, in order to increase the naturalness of the synthesized speech it is not enough to control only one of the parameters.

THE PROBLEM OF NATURALNESS

The text-to-speech systems used by us /10/ and many other authors contain two main blocks - the model of vocal tract and the block of control on the basis of a microcomputer.

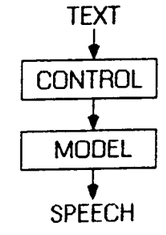


Fig. 2. The structure of the text-to-speech systems.

In case of such a structure of the system, the optimal correspondence between the control program and the technical aids has been found. The more exactly vocal tract is modelled, the harder it is to control it; on the other hand, a too simple model cannot provide the sufficient quality of speech.

We have used the classical model of formant synthesis:

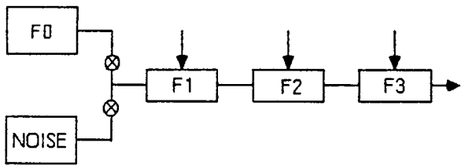


Fig. 3. The model of the vocal tract.

The control parameters of the model are established on the basis of a text which is provided with punctuation marks and stress-marks. This process consists of the following steps:

- transforming the orthographic text into a phonematic text;
- establishing the time structure;
- computing the intonation contour;
- taking into account the phenomena of coarticulation.

The model is characterized by the possibility of controlling all the parameters separately, i.e. the parameters of the model are not interdependent. Thus, incorrect control is possible.

The control errors can be eliminated in the process of determining the control parameters when the quantity of the knowledge and the rules of controlling the synthesis, both able to take the influence of the phonetic context into account, are big enough. But this brings about the enlarging of the memory of the system and the demands on the speed of operating the information.

Another way of eliminating the errors is to set uncontrollable by the program connections between the parameters of the vocal tract model. The existence of such connections is proved by many authors (more thoroughly are considered the relations between the intrinsic pitch of the vowels and the openness of the vocal tract; the changes of the pitch frequency in vowels of CVC context; the changes of the shape of the glottal pulse in the process of articulation). Such phenomena are rooted in the peculiarities of the vocal tract, and are not controlled by the program of articulation. People are used to such phenomena and regard them as compulsory. It may be assumed that not taking these connections into account is one of the reasons of the machine-like sound of the synthesized speech.

This can be proved by introducing into the vocal tract model the additional connections which imitate the above mentioned phenomena.

As the frequency of the first formant is the best determinant of the openness of the vocal tract, the connection between the frequency of the first formant and the pitch frequency is introduced. The experiments carried out by using the formant synthesizer showed that the machine-like sound diminishes if a dependence in accordance with

$$F0 = f0 * K/F1$$

is introduced, where F0 is the frequency of the pitch generator, F1 is the frequency of the first formant of a vowel, f0 is the computed frequency of the pitch and K is the coefficient of connection (K=500).

The dependence should be introduced for stressed vowels, for in all the other cases F0=f0.

In order to check the validity of this introduction, an experiment was carried out with a group of naitive listeners. Sentences consisting of one and two syntagmas and with different intonation types (10 sentences per each type) were synthesized using the formant synthesizer. The sentences were synthesized first without the connections between F0 and F1 (sentences A) and then with introducing the connection (sentences B). The listeners heard the sentences in pairs of optional order (AB or BA). The task of the listeners was to estimate which variant in a pair was more natural.

The results of the experiment prove firmly that the intorduction of the above mentioned connection is justified, for in all the cases the listeners chose variant B as more natural of the two.

The effectiveness of the introduction depends on the choice of the value of coefficient K. From the point of view of naturalness, the optimal value of K appeared to be K=500. When K<<500 the effect was not noticed, when K>>500 the structure of an intonation contour is violated and an effect of deep emotional excitement could be detected.

DISCUSSION

Although the obtained results demand to be studied thoroughly, already at the present stage the structure and the principles of encoding the speech signal can be detected. The fact that the introduction of the connections between the pitch and the first formant affect the quality and the naturalness of the synthesized speech was to be expected. There are a number of indirect data ponting to the fact that all living organisms and the signals they send out are subject to a general principle of "integral unity".

According to this principle information is encoded in a definte number of signs which are arranged in an order according to the importance and reliability and where the previous sign determines the region and corrects the strategy for searching for the following signs in the order of importance. All the signs can change within the corridor which width and the trajectories of the movement are determined by previous signs [11].

The introduced connections between the first formant and the pitch correspond to the principle of "integral unity" and can be explained by the fact that the frequency and the impulses of the pitch depend on the tension of the vocal cords, on subglottal pressure and pressure over the vocal cords. The tension of the vocal cords and subglottal pressure directly control the muscles of the larynx and the diaphragm. The pressure over the vocal cords coordinates the maxilla, the tongue and the lips. The less opened the vocal tract is, the higher is the pressure over the vocal cords and the fundamental frequency. In case of fluent speech the openness of the vocal tract changes quickly and the real fundamental frequency fluctuates around the intonation contours, depending on the position of the maxilla, tongue and lips.

CONCLUSION

Intonation is an essential factor when trying to increase the quality and the naturalness of the synthesized speech but in order to get rid of the machine-like sound it is not enough to model the intonation correctly. In a speech signal all the signs are distributed in the whole frequency-duration range and the machine-like effect is caused not only by the monotony of the changes of the pitch but are also characterized by the form of the glottal pulses, intnsity, duration structure etc. In order to completely free the synthesized speech from the machine-like sound it is necessary to control not only the changes of pitch around the intonation contour but also all the other signs. But for that it is inevitable to study the connections between the different parameters so that the optimal strategy of control can be worked out, i.e. which of the parameters are controlled by a program and to which the connections according to the principle of "integral unity" can be applied.

REFERENCES

[1] G.Kaplan, E.J.Lerner "Realism in synthetic speech" IEEE Spectrum, April, 1985, p.32-37.

[2] J.Pierrehumbert "Synthesizing intonation" J. Acoust. Soc. Am. 70(4), Oct. 1981, p.985-995.

[3] D.O'Shaughnessy "Linguistic Features in Fundamental Frequency Patterns" J. Phonet. 7, 1979, p.119-145.

[4] K.Hirose, H.Fujisaki "Analysis and Synthesis of Voice Fundamental Frequency Contours of Spoken Sentences" IEEE, 1982, p.950-953.

[5] C.H.Shadle "Intrinsic fundamental frequency of vowels in sentence context" J.Acoust. Soc. Am. 78(5) Nov. 1985, p.1562-1567.

[6] B.Lyberg "Some fundamental frequency perturbations in a sentence context" J. Phonet. 12, 1984, p.307-317.

[7] Е.А.Брызгунова Звуки и интонация русской речи" Москва, 1969.

[8] Н.Д. Светозарова "Интонационная система русского языка" Ленинград, 1984.

[9] N.D.Svetozarova "The Inner Structure of Intonation Contours in Russian", - Auditory analysis and perception of speech.- London, New York, San Francisco, Academic Press, 1975, p.499-510.

[10] О.Кюннап, А.Отт "Управляемый микропроцессором синтезатор речи" В кн.: Автоматическое распознавание слуховых образов - 12, Киев, 1982, стр. 410-411.

[11] M.Rohtla, K.Lindvere "Extraction of features from acoustic signals" EPP, Tallin, 1977, p.89-91.

PHONETIC CONSIDERATIONS FOR THE SYNTHESIS OF FEMALE VOICES

CAROLINE G. HENTON

Linguistics Program, University of California, Davis,
Davis, CA 95616, U.S.A.

ABSTRACT

Synthesized female voices are scarce and lack naturality, but they are growing in demand. Acoustic and sociophonetic criteria are supplied for the improvement of female voices, and a ranking of importance suggested.

"Synthesis is going to be the next barn-burning technology," was an informed forecast three years ago [9]. It is further predicted that by 1992, the combined American and European markets for electronic speech synthesis and recognition devices will approach $5 billion [8]. While speech synthesis is a priority in speech technology research, and several commercial packages (e.g. DECtalk, Calltext) are producing successful male voices, distinctly fewer felicitous female synthetic voices are available. The widespread appearance of synthetic female speech is slow. Why is this? Are female voices not to be included in the barn-burning, or contribute to the multi-billion sales?
Outlined here are phonetic and social reasons for the paucity of synthetic female voices. There follows discussion of some acoustic specifications of female voices which are relevant to synthesis. Recent research pertaining to female voice quality is reported and a ranking of these various factors proposed.

BACKGROUND

Phonetically, the female voice has been largely ignored for two reasons. The first is as a result of minimal productional data. A cross-language survey of phonetic studies conducted 1952-1985 [14], which ostensibly provided 'representative' adult acoustic data, shows that among 42 studies, 40.5% assembled solely male speakers. 21.4% incorporated more males than females. Only one study (2.4%) incorporated more females than males. Studies of females alone are a meagre 4.8%. The first point, then, is that in acoustic phonetic research, the female voice has been either excluded or minimalized.
Secondly, female voices have been rejected acoustically (and hence, disregarded in phonetic theory) owing to inadequacies in analytic hardware. That should be obvious to anyone who has wrestled with interpreting spectrograms of female voices. Until recently, the sound spectrograph has been the most frequently-used tool in acoustic speech analysis, and other instruments (such as narrow-band spectrum analyzers) are still imperfect in analyzing females' speech. Criticisms of the problematicity of formant frequency determination for female speakers, using spectrography, are made by Ladefoged [25] and Ladefoged and Bladon [26].
The apparent source of the 'problem' of female speech appears in an article by Johansson et al. [21]: "Comparatively little is known about the characteristics of the female voice as compared with the male voice. The background is the high fundamental frequency range of the female voice which makes formant frequency estimates uncertain, and hence, information on the voice source unsafe." The logic may be chopped, and the association of formants with the voice source misleading, but the message is clear: the female voice is puzzling because it is not the same as a male's. This issue has also caught the attention of Klatt [23]. Reviewing the efficacy of spectrograms from which to draw acoustic conclusions, he states, "As far as speech ...research is concerned, it is not inconceivable that the sound spectrograph has had an overall detrimental influence over the last forty years by emphasizing aspects of speech spectra that are probably not direct perceptual cues." He presents spectrograms of the same utterance produced by a man, woman and child: "The woman and child speak with a much higher fundamental frequency, have a more breathy voice quality, and also have shorter vocal tracts, implying higher formant frequencies...". These traits are discussed further, below; meanwhile Klatt asserts that (p.83), "...it seems to be generally believed that the speech patterns of men and women could be made to look more similar if minor modifications were made to the sound spectrograph...Yet, here we are, nearly forty years later, and the sound spectrograph machine essentially has not changed." Such a situation surely reflects androcentric reality in the obscuration of females' voices. Here, though, the plea is not even for the spectrographic commensurability of male and female voices, but simply for the *measurability* (and hence reproducability) of female voices.
The implication from many such comments about female speech is that there is something intrinsically more difficult analytically, or just deviant, about female voices. The assumption is incorrect, but too few authors have thought to blame the design of the technology rather than females for producing analytical problems. Female voices only appear more 'difficult' because of the limitations of some present instrumentation. They are not more 'difficult' to the human ear: females are not any less intelligible than males, and may even be more so, although evidence seems somewhat variable [6], [12], [23], and [14: 312 ff.].
It is possible to infer then that a great deal more could be known about female voices, if the technology were improved for processing speakers with higher fundamental frequencies, namely the 'unquantifiable' females. Unfortunately (but entirely in keeping with the distribution of women in scientific positions generally), few phoneticians and even fewer technologists are female. So there is little grassroot motivation for improving females' analyzable lot.

WHAT NOW?

Notwithstanding such a negative background, demand is increasing for synthetic female speech. Naval pilots, for example, apparently react best to the voice of a young woman when warning them of upcoming obstacles or potential problems in the cockpit [8] It is clear that female voices are going to be needed more, for 'smoothing' and other messages.
So what *do* we know about female voices to contribute to their better synthesis? Sociophonetic evidence of the kind described in [14], [15] and [16] indicates a broad range of acoustic, perceptual and social factors as being influential in the synthesis of female voices. We will now briefly outline some of those categories. Starting with the generation of an appropriate glottal waveform, we then address formant frequency values for vowels, possible sex-specific factors in consonant production, and lastly, but perhaps most importantly, suprasegmental considerations and types of voice quality associated with female speech.

Glottal source characteristics. Naively it might be thought that to synthesize a convincing female voice pitch, it would be sufficient to simply double that of a male, increasing from, say 120Hz. to 240Hz. Several studies have shown, however, that there are marked differences in the glottal gestures of females and males [24], [31]. The latter show the glottal sound source of a normal adult can vary within a wide range, in respect to F0 and rms intensity, the appearance and shape of the waveform, and the phase and intensity spectra. Most important for the argument here, is the fact that all these variants can be influenced by the variables speaker sex, voice register and linguistic context. In addition, the female glottal waveform tends to have a less steep closing phase and a more rounded 'shoulder' at the end of that phase, and consequently, a higher ratio of open-to-close time which could result in more glottal leakage or weaker excitation of higher harmonics. Generating an appropriately varying female glottal waveform is thus vital for natural-sounding synthesis.

Formant frequencies. Male/female differences in formant frequencies of vowels have been reported widely elsewhere [4],[14]. Details do not bear reiteration here. Cross-linguistic findings from seven languages/dialects [14] may be summarized: while a male-female auditory normalization of approximately one Bark appears appropriate for F1 and F2, there are also indications that different speech communities need different amounts of normalization. That is to say, in some communities females and males appear to speak more like/unlike each other than their vocal anatomies would predict. A socially-conditioned element in speech production is thus posited. Hence the amount of physically-based input to the voice signal as compared to the socially-learned component must be weighed carefully when synthesizing speech.
In addition, the spectral tilt of female and male vowels might yield further evidence of sex-differentiation. we might predict that the angle of tilt of the spectrum, as it increases in frequency, would be somewhat steeper for females than for males. Indications are [31] that the decrease is approximately -12dB per octave in male vowels, but -15dB in female vowels. Inter-sex differences of formant bandwidth may also appear, with females' bandwidths being wider than males [11]. It is not known yet whether this difference is perceptible.
Sex-appropriate consonants. Whereas vowels exhibit important sex-specific cues, the evidence for specifically female consonants is less convincing. Generally, consonants have been explored less thoroughly: where male-female differences

have been noticed, it has been for ethnographic or tangential reasons. The relative linguistic function of vowels and consonants indicates that vowels exhibit the most individual-speaker traits. So speaker-sex (along with tone, affective state etc.) is more likely to be shown by vowels. Paralinguistic information such as exclamations, expression of pain, anger and so forth is also most often conveyed by vowel-like sounds. Consonants, conversely, are more language-specific, conveying linguistic information. Certain consonants, e.g. front fricatives and many stops, are unlikely to have vocal tract resonance properties which would show acoustic spectral differences of a sex-linked kind. Sex-specific behaviour in fricatives has been examined to a limited extent [3], [19], [35]. They indicate that different fricatives seem to carry differing amounts of sex-specific information. Weeninck [39] suggests that male-female differences in plosive bursts are also not anticipated. Other consonants are unlikely to show a sex-linked difference because there is already much individual speaker variation in their production. This argument applies to nasals, where individuals' nasal structures and degree of constriction may vary greatly. When speakers do appear to use consonants sex-specifically, they do so in an apparently unpatterned way: thus, females' realizations may differ from males' according to place of articulation, manner of articulation, direction of the airstream, voicing, or any combination of these four parameters. It is probable, then, that, sibilants apart, manipulating consonants to improve synthetic female speech will be unrewarding (unless a language uses consonantal variation as a specific sex-marker - which European languages, on the whole, do not).

Suprasegmental markers. Pitch, tone, vowel length, intensity, hesitancy, stress (word and sentence), rhythm and intonational tunes are rich areas for the exhibition of femininity/masculinity. Many argue that suprasegmental features are the primary cues of speaker-sex. Surprisingly little empirical, objective research into these areas has been initiated, however. Pitch has received some attention. Hollien and colleagues [17], [18] have indicated that the speaking fundamental frequency (SFF) of males may be lower than differences in stature would predict. This tendency is further borne out by Henton [13] for British English, and the conclusion must be that voice pitch is to some extent learned, and subject to sociocultural expectations. Such behaviour should not be overlooked in synthetic voices. Still, no researchers seem to be asking whether, for example, females habitually use a higher/lower pitch than vocal anatomy would predict; or are certain long-term SFFs associated with one sex, as interacting with socio-economic status; or do speakers of either sex alter their SFF by differing amounts according to register, or speech style? Preliminary answers to these sorts of questions, which are after all vital to successful synthesis, are beginning to appear incidentally [33], but this is hardly a broadscale line of enquiry.

Crystal [7] cites anecdotal observation of male-female suprasegmental differences such as glissando effects, complex-tone usage, breathiness and moving to falsetto involved in the production of 'simpering' voice in English. Aronovitch [1] found that his parameters of 'Intensity average; intensity variance; rate; F0 average; F0 variance and sound-silence ratio' were used differentially by raters for the two sexes. His goal, however, was to investigate if there are any connections between voice quality and stereotyped judgements, and not to examine diverse suprasegmental behaviour of the sexes. Unfortunately, his findings have yet to tantalize other researchers into further work along these lines.

Intensity has been examined to a limited extent. One study [29] revealed that males speak with greater average intensity in interpersonal communication than did females, although both sexes address a member of the opposite sex with greater intensity than a same-sex member. Another investigation [34] showed that women raise their intensity more than men when compensating for external noise. There is a brief summary in [38] of the few studies of sex-linked verbal fluency in adults. A comment by Smith [37:125] is still valid however: there are "only two paralinguistic features (viz. loudness and speech fluency, my parentheses) for which there is even a modicum of data."

For word and sentence stress, there is an array of intuitive remarks. Jespersen [20] claimed that "exaggeration of stress" is a conspicuous characteristic of female speech; and his intuitive successor, Lakoff [27:56] states that "Women speak in italics." As far as we are aware, no empirical investigation of stress in female speech is known as yet. Nor is there any of vowel length.

Sex-based intonation tunes *have* attracted attention, but too often in an unsystematic way, with little use of experimental measurement and too great a reliance on subjective judgement: [5], [22], [30]. Better results are provided by Pellowe and Jones [32], who showed that females use more rising tones than men. Women also exploited a greater variety of intonational tunes. Elyan's results [10] lend more support to this observation. More extensive exploration of these tendencies, across accents, is still required. Male speech is characterized by monotonicity [2], while on the other hand female speech has greater *intonational dynamism* [30], varying more in respect of width of pitch-range; more frequent and rapid movements within the range; greater amplitude changes, and selection of differing tunes. McConnell-Ginet [30] hypothesizes that the female habit of changing pitch and loudness frequently may have great communicative importance, both attracting and keeping the listener's attention. As that is a common obstacle in speech synthesis, paramount attention should be given to such suprasegmental behaviour in females.

Voice quality. A major contribution to the paralinguistic component of speech is made by habitual vocal settings, or voice qualities, such as harshness, breathiness or creak. Perceptual correlates of voice qualities have been studied incidentally [1], but essential baseline information about the sociophonetic production and sex-specific use of different voice qualities is slow to emerge. Recently, though, Henton and Bladon [15], [16] have explored sex-related incidence of two common voice qualities: breathiness and creak. Both studies were conducted using a large corpus (80 speakers) of two accents of British English.

For breathiness, it was found, at least in open vowels, females employ significantly greater amounts of this voice quality, thus supporting previous tangential observations, [7], [23]. Implications for the diagnosis of normal versus pathological speech in females, as well as for speech synthesis are discussed further in that paper [15].

Creak, on the other hand, appears to be a marker of male speech in British English (and, by informal observation, in American English, too). Summarizing the findings of the creak study [16], we found that (a) creak was an utterance-final phenomenon, with a linguistic function of pre-pausal demarcation; (b) speaker-sex predicts drastically rate of creak: creak may be seen as a robust marker of male speech; (c) creak may be employed to different degrees in different accents to indicate hypo/hyper-masculinity; (d) creak is used habitually by overwhelming numbers of non-pathological male speakers. The incidence of creak thus varies with sex and accent, as well as with utterance-position. Male voice synthesis will obviously benefit from a sprinkling of creak, with female creak limited to utterance-final position (unless the desired voice is purposely androgynous).

These explorations make us ask more widely whether other voice qualities are sex-indicative. Are, for example, nasality, rhoticity, or pharyngealization typical of one sex or the other? Since voice quality appears to carry strong associations in the assessment of personalities [1], it seems essential to exploit this fact in speech synthesis. A female voice should therefore include greater breathiness, creak before relinquishing a turn and, speculatively, be more nasalized for appropriate accents.

SUMMARY AND PERCEPTUAL RANKING

After reviewing these various physical, segmental and suprasegmental aspects necessary for synthesizing female voices, it seems incumbent to provide a ranking of these features, according to perceptual salience. The ranking, in order of importance, is: voice quality, pitch, suprasegmentals, vowels, consonants. The synthesis of current synthetic female voices may not have adopted these criteria and so continue to sound unnatural. The increased production of convincing female voices, hand-in-hand with eliminating the female voice as "one of the mysteries of the universe" [36] is undoubtedly a profitable goal for speech synthesis.

[1] C.D. Aronovitch, The voice of personality: stereotyped judgements and their relation to voice quality and sex of speaker. Jour. Soc. Psych., 99: 207-20, 1976.
[2] S. Bennett & B. Weinberg, Sexual characteristics of pre-adolescent children's voices. Jour. Acoust. Soc. Amer., 65: 179-89, 1979.
[3] A. Bladon, The use of auditory modelling for speaker normalization in speech recognition. In P. Mermelstein (ed.) Montreal Symposium on Speech Recognition. 1986.
[4] R.A.W. Bladon, C.G. Henton & J.B. Pickering, Towards an auditory theory of speaker normalization. Lang. & Speech, 4: 59-69, 1984.
[5] R. Brend, Male-female intonation patterns in American English. In B. Thorne & N. Henley (eds.) Language and Sex: Difference and Dominance. Newbury House: 84-87, 1975.
[6] F.R. Chen, Acoustic Characteristics and Intelligibility of Clear and Conversational Speech at the Segmental Level. S.M. thesis, Dept. Electrical Engineering, MIT, 1980.
[7] D. Crystal, The English Tone of Voice. Arnold, 1975.
[8] B. Darrow, Research spurs development of talking machines. Design News, 12 March, 1984, p.110.
[9] L. Drinkwater, of Wang, quoted in Industry Week. October 15, 1984, p.82.
[10] O. Elyan, Sex differences in speech style. Women Speaking, 4, 1978.
[11] G. Fant, Temporal fine structure of formant damping and excitation. In J.J. Wolf & D. Klatt (eds.) Speech Communication Papers, Acoust. Soc. Amer: 161-66, 1979.
[12] U. Goldstein, An Articulatory Model for the Vocal Tracts of Growing Children. D.Sc. thesis, MIT, 1980.
[13] C.G. Henton, Normalization: fundamental problems. Proc. Inst. Acoust., 6: 267-73, 1984.
[14] C.G. Henton, A Comparative Study of Phonetic Sex-specific Differences Across Languages. D.Phil. thesis, University of Oxford, 1986.
[15] C.G. Henton & R.A.W. Bladon, Breathiness in normal female speech: inefficiency versus desirability. Lang. and Commn., 5: 221-27, 1985.
[16] C.G. Henton & R.A.W. Bladon, Creak as a sociophonetic marker. In L. Hyman & C. Li (eds.) Language, Speech and Mind: Studies in Honor of Victoria A. Fromkin. Croom Helm, forthcoming, 1988.
[17] H. Hollien & B. Jackson, Normative data on the speaking fundamental frequency characteristics of young adult males. Jour. Phon., 1: 117-20, 1973.
[18] H. Hollien & T. Shipp, Speaking fundamental frequency and chronologic age in males. Jour. Sp. and Hear. Res., 15: 155-59, 1972.
[19] F. Ingemann, Identification of the speaker's sex from voiceless fricatives. Jour. Acoust. Soc. Amer., 44: 1142-44, 1968.
[20] O. Jespersen, Language, Its Nature, Development and Origin. Allen & Unwin, 1922.
[21] C. Johansson, J. Sundberg & H. Wilbrand, X ray study of articulation and formant frequencies in two female singers. Quar. Prog. Stat. Rep., KIT Stockholm, 4: 117-34, 1982.
[22] M.R. Key, Linguistic behavior of male and female. Linguistics, 88: 15-31, 1972.
[23] D. Klatt, Speech processing strategies based on auditory models. In R. Carlson and B. Granstrom (eds.) The Representation of Speech in the Peripheral Auditory System. Elsevier: 181-96, 1982.
[24] D. Klatt, Detailed spectral analysis of a female voice. Abstract in Jour. Acoust. Soc. Amer., 80, Suppl.1: 597, 1986.
[25] P. Ladefoged, Three Areas of Experimental Phonetics. Oxford University Press, 1967.
[26] P. Ladefoged & R.A.W. Bladon, Attempts by human speakers to reproduce Fant's nomograms, Sp. Comm., 1: 185-98.
[27] R. Lakoff, Language and Woman's Place. Harper Colophon, 1975.
[28] W. Majewski, H. Hollien & J. Zalewski, Speaking fundamental frequency characteristics of Polish adult males. Phonetica, 25: 119-25, 1972.
[29] N.N. Markel, L.D. Prebor & J.F. Brandt, Biosocial factors in dyadic communication: sex and speaking intensity. Jour. Pers. and Soc. Psych., 23: 11-13, 1972.
[30] S. McConnell-Ginet, Intonation in a man's world. In B. Thorne, C. Kramarae & N. Henley (eds.) Language, Gender and Society. Newbury House, 1983.
[31] R.B. Monsen & A.M. Engebretson, Study of variations in the male and female glottal wave. Jour. Acoust. Soc. Amer., 62: 981-93, 1977.
[32] J. Pellowe & V. Jones, On intonational variability in Tyneside speech. In P. Trudgill (ed.) Sociolinguistic Patterns in British English. Arnold: 101-21, 1978.
[33] O. de Pinto & H. Hollien, Speaking fundamental frequency characteristics of Australian women: then and now. Jour. Phon., 10: 367-75.
[34] W. von Raffler-Engel & J. Buckner, A difference beyond inherent pitch? In B. Dubois & I. Crouch (eds.) The Sociology of the Languages of American Women. Texas, 1976.
[35] M.F. Schwartz, Identification of speaker sex from isolated, voiceless fricatives. Jour. Acoust. Soc. Amer., 43: 1178-79, 1968.
[36] R. Shuy, Sociolinguistic research at the Center for Applied Linguistics: the correlation of language and sex. International Days of Sociolinguistics. Institutio Luigi Sturzo: 849-57.
[37] P. Smith, Language, the Sexes and Society. Blackwell, 1985.
[38] B. Thorne, C. Kramarae & N. Henley (eds.) Language, Gender and Society. Newbury House, 1983.
[39] D.J.H. Weeninck, Literature overview on perceptual and physical normalization of speaker variation. Proc. Inst. Phon. Sciences, Univ. Amsterdam, 8: 5-17, 1984.

STUDIES OF PHONETIC FUNCTION APPLICABILITY TO AUTOMATIC QUALITY TESTING OF SPEECH PROCESSING SYSTEMS AND TRANSMISSION CHANNELS

V.N. SOBOLEV

All-Union Correspondence Electrotechnical Institute of Communications, Moscow, USSR 123855

ABSTRACT

The quality of a speech channel or codec is tested by comparing the phonetic functions at its input and output. The results of experimental verification of this method are presented, its hardware implementation is discussed.

INTRODUCTION

An objective method of transmission quality estimation is essential in designing new speech processing equipment, such as vocoders and speech waveform coders, as well as in maintenance of existing communication channels. At present, this is performed by averaging the subjective responces of a large number of listeners. Such statistics yield consistently accurate results, but are time- and labor-consuming. An objective quality measure would facilitate detection of distortion causes and thus improve the efficiency of designing new speech coders. In maintenance this would provide automatic monitoring of channel quality for timely replacement or adjustment of faulty units.

The problem of developing, study and use of methods of objective speech transmission quality measurements has been addressed by a number of authors. The methods discussed include correlation techniques [1,2], segment and subjective signal-to-noise ratios [3,4], isopreference method [5], various physical methods (e.g. as in [6]), log likelihood ratio measure [7], and the Itakura-Saito measures [8]. However, some of them are rather complicated in actual practice, others are insensitive to certain types of distortion, and yet others are not always adequately accurate. Thus, the problem of automated objective quality testing retains its urgency and further research for new approaches is justified. Presented here is an attempt to introduce a unique transmission quality index based on comparing phonetic function at a system's input and output.

THE METHOD

The phonetic function

$$P(\omega, t) = \int_0^{\infty} \exp(-\tau/T) \cdot \log \frac{S(\omega, t)}{S(\omega, t-\tau)} d\tau$$

where $S(\omega,t)$ is the modulus of the speech signal's f(t) short-term spectrum, was first introduced by A.A.Pirogov [9] and is successfully used for phonem recognition [10]. The feasibility of evaluating speech transmission quality with this function has been demonstrated [11]. The logarithmic term in this equation describes the increment in information amount during the time interval τ at a t moment of time and frequency ω. This phonetic function is based on the Weber-Fechner psychophysiological law and takes human ear adaptation effects into account, i.e. the logarithmic relation between sound perception and intensity of audial analyzer excitation along with the human ear tending to percieve subsequent sounds against a background of impressions from preceeding sounds. The phonetic function describes only the dynamics of speech signal spectrum variations in the time domain.

Comparing phonetic functions at a codec input and output yields the following equation to describe a criterion of speech transmission quality:

$$q = \int_{\omega_1}^{\omega_2} w(\omega) \int_{t_1}^{t_2} P(\omega,t) \circledast \tilde{P}(\omega,t) \cdot dt \cdot d\omega,$$

where: $w(\omega)$ is the weight function to deform the frequency axis according to the Koenig scale; ω_1 and ω_2 are the lower and upper frequencies of the band under study; t_1 and t_2 limit the comparison time interval; and symbol $\circledast$ depicts the comparison operation. Speech signals differ more from one another in spectrum distribution time variations, rather than in spectra themselves and therefore comparing signals by their phonetic functions is both justified and feasible.

EXPERIMENTAL VERIFICATION

The experiments included comparing the phonetic functions at inputs and outputs of speech waveform codecs and comparing these results with those of articulation tests. DM and PCM devices were tested at various transmission rates with syllabic intelligibility ranging from 30 % to 80 %. Measurements were computerized, with the test signal in the form of a tape record of phonetically balanced speech of 42 seconds duration, from four different dictors. The computer input signals were the logs of the speech signals short-term spectra, $\log S(\omega_i,t)$ and $\log \tilde{S}(\omega_i,t)$, from the outputs of band-pass analyzers, rather than the initial speech signals f(t) and $\tilde{f}(t)$. The center frequencies of the 16 band-pass filters range from DC to 7 kHz and are distributed according to Koenig's scale, thus realizing the $w(\omega)$ weight function. The phonetic functions were computed by recursion formulas

$$P(\omega_i, n\cdot\Delta t) = (T/\Delta t)\log S(\omega_i, n\cdot\Delta t) - H(\omega_i, n\cdot\Delta t)$$

$$H(\omega_i, n\cdot\Delta t) = \exp(-\Delta t/T)\cdot H(\omega_i, (n-1)\cdot\Delta t) + \log S(\omega_i, n\cdot\Delta t)$$

along with the quality index q, with various comparison techniques, the best of

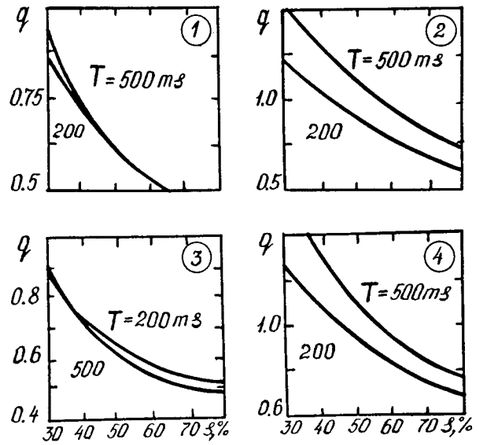


Fig. 1. Quality index q vs. syllabic intelligibility s (with different methods of comparing phonetic functions)

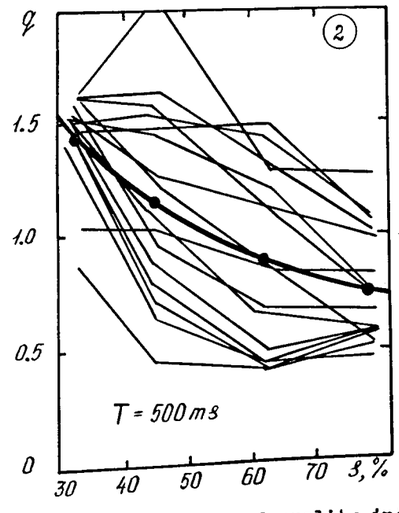


Fig. 2. Individual channel quality indices $q_i(s)$ (broken lines) and the averaged function q(s) (solid curve)

which turned out to be:

$$P①\tilde{P} = |P - \tilde{P}|; \quad P②\tilde{P} = (P - \tilde{P})^2;$$

$$P③\tilde{P} = \frac{|P - \tilde{P}|}{\max(|P|,|\tilde{P}|)}; \quad P④\tilde{P} = \frac{(P - \tilde{P})^2}{\max(P^2, \tilde{P}^2)}.$$

These computations resulted in monotonic functions which relate the quality index q to the syllabic intelligibility s; examples are shown in Fig.1. The highest

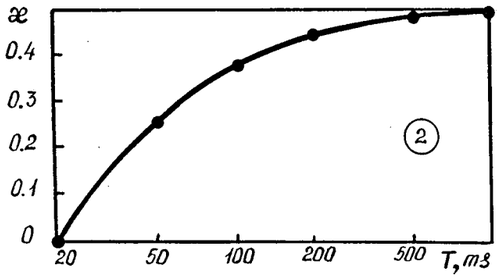


Fig. 3. Average slope of q(s) curves vs. averaging time T

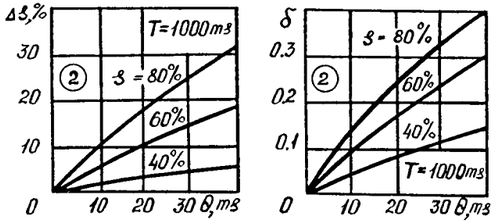


Fig. 4. Absolute (Δs) and relative (δ) syllabic intelligibility measurement error vs. time shift between signals being compared

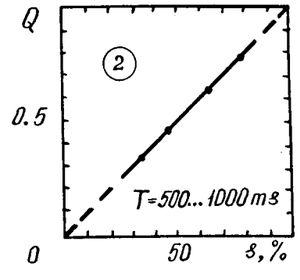


Fig. 5. Quality index Q and syllabic intelligibility s

sensitivity was obtained with the second comparison technique. Relations

$$q_i = \int_{t_1}^{t_2} P(\omega_i,t) \quad ② \quad \tilde{P}(\omega_i,t)dt$$

for individual band-pass filters are shown in Fig.2 in broken lines along with the resulting relation

$$q = \sum_{i=1}^{16} q_i$$

shown as a bold curve, confirming the expedience of covering a wide frequency band. Fig. 3 shows the average slope of q(s) curves, calculated by $æ = (q(33) - q(78))/q(33)$, vs. the averaging time constant T. The highest measurement sensitivity was obtained at T from 200 to 1000 ms.

The effect of a time shift between f(t) and $\tilde{f}(t)$ was studied, with measurements showing the monotonicity of q(s) curves being maintained at time shifts up to 50 ms, but at steadily decreasing slope angles and accompanied by an upward displacement of the curves. This means that the presence of a time shift between the signals being compared impairs the measurement sensitivity and leads to underrating the measured quality index. Measured absolute Δs and relative δ errors of automatic syllabic intelligibility s measurements are shown in Fig.4 as functions of the time shift θ for systems of various transmission quality.

These measurements are complicated by the inverse proportionality between q and s; a modified criterion $Q = (r/q) - \varphi$ (r and φ are empirical coefficients) proves to be more convenient. When using the second comparison technique, the best fit was obtained by setting r = 0.71 and φ = 0.13 for time constants T from 500 to 1000 ms; Q(s) under these conditions is as shown in Fig. 5, which confirmes the correspondence between Q and s.

HARDWARE IMPLEMENTATION

From the above it follows, that automatic measurements of speech transmission quality can be provided by arrangement shown in Fig. 6, which functions as follows. The log of the output signals from two band-pass analyzers are fed to the inputs of channel integrators. In each of the channel adders the integrator output signal is subtracted from its input signal, thus producing a signal proportional to the log of the ratio of the short-term spectrum at this time moment to the integral of this spectrum over the preceeding

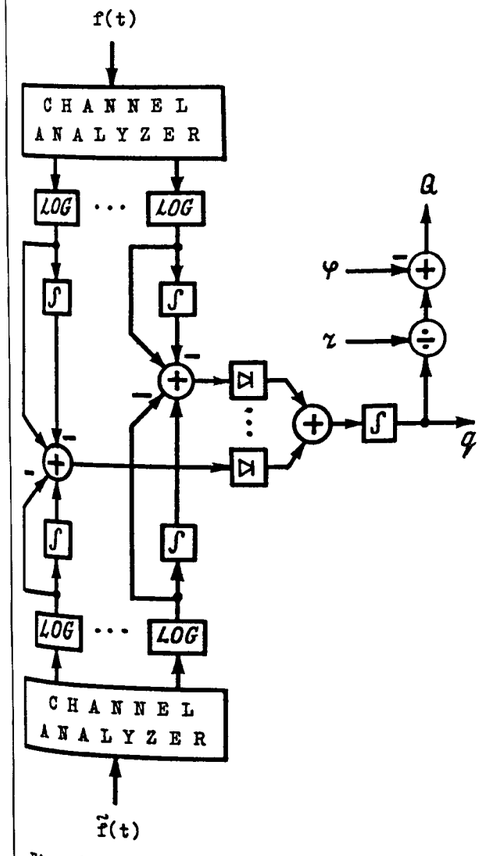


Fig. 6. Skeleton diagram of measurement circuit arrangement

time interval. The set of such difference signals in all channels represents the speech signal phonetic function. Since the difference signals from the upper and lower sections of this circuit arrive at the channel adders in antiphase, the output of these adders is the difference between the phonetic functions of speech signals f(t) and $\tilde{f}(t)$. The channel adder output signals are squared, summed and averaged in a group integrator over a significantly longer time interval. The output voltage, q, of this group integrator is then converted into Q. This latter quantity is the measure of transmission quality via a codec or channel under measurement.

REFERENCES

1. М.А.Сапожков. Акустический журнал, 1956, вып. 3, с. 279 - 284.
2. Т.Е.Зайцев. Электросвязь, 1958, № 10, с. 38 - 46.
3. C.Scagliola. BSTJ, v.58, Jul.-Aug. 1979, pp.1369-1394.
4. M.Nakatsui, P.Mermelstein.JASA, v.72, No.4 (1982), pp.1136-1144.
5. W.A.Munson, J.E.Karlin. JASA, v.34 (1962), pp.762-774.
6. H.J.M.Steeneken, T.Houtgast. JASA, v.67 (1980), pp.318-326.
7. R.E.Crochiere, J.M.Tribolet, L.R.Rabiner IEEE Trans.Acous., Speech Signal Processing, v.ASSP-28, No.3, June 1980, pp.318-323
8.B.-H.Juang. On using the Itakura-Saito measures for speech coder performance evaluation.- AT&T Bell Lab.Techn.J., v.63, No.8, Pt.1, 1984, pp. 1477-1498.
9. А.А.Пирогов. Электросвязь, 1967, № 5, с. 24 - 31.
10. Вокодерная телефония / под ред. А.А. Пирогова - М.: Связь, 1974.
11. В.Н.Соболев, Г.В.Титова. Автоматическое распознавание слуховых образов. Тезисы докладов Всесоюзного семинара АРСО-8, с. 87 - 89, 1974.

THE SYNTHESIS-BY-RULE DEVELOPMENT SYSTEM WITH EXPERT CAPABILITIES

ARVO OTT

Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

IMRE SIIL

Dept. of Software
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

ABSTRACT

A flexible speech synthesis development system is described. It is a production system in which the two components - the declarative and the procedural knowledge base must be created by the domain expert. A simple rule language, interactive graphics, acoustical AB tests and explanation capabilities of the system are at the disposal of the expert. The production system runs on the I8080 microprocessor in real time.

INTRODUCTION

Futher progress in speech synthesis obviously depends on powerful and flexible development tools [1,2].
It is useful to reduce the role of the speech synthesis engineers in the process of obtaining various linguistic knowledge and give the linguists the possibility to explicitly model and immediately apply their professional knowledge. Besides, the synthesis system must be observed from two different aspects - the synthesis system, meant for real time, on-line implementation, and the system as a tool for investigations to examin the adequacy of phonetic or phonologic descriptions.
Selecting a suitable representation for the domain knowledge is one of the major problems to be encountered in building a knowledge-based system aimed at disposal of experts. Several high-level rule languages have been created for speech synthesis [1,3,4,5]. The general methodology on which these languages are based, is in principle the same. It is close to the technique of the production systems, known in the field of artificial intelligence. Indeed, in speech synthesis the terminology of production systems and expert systems has been used too [5].
The production rule system, representing the knowledge of speech synthesis control has been under development since 1982 in the Institute of Cybernetics of the Estonian Academy of Sciences. These studies were aimed at creating a flexible speech system, using terminal-analog speech synthesizer. Also, there was the task to minimize the calculation resources of microprocessor I8080, used to control the synthesizer.
We will focuse on representation of knowledge needed on different control levels of the synthesizer. All discussions have been made, taking into account the technical limitations of the realized synthesis system on the one hand, and at the same time to give the system maximum flexibility and to make minimal ad hoc solutions on the other hand.

1.VOCAL TRACT MODEL

The configuration of the vocal tract model of the formant synthesizer FS-05 was chosen and determined by a set of experiments with digital model, realized on a general purpose computer ES-1010. The resulting serial/parallel formant model is somewhat similar to the model, used in synthesizer OVE 3. There are 3 turnable (F1,F2,F3) and 2 fixed formant filters in vocal branch, a turnable resonator (FF) in fricative branch, fixed resonator in nasal branch, 5 switches for amplitude control (AV,AH,AF, AN,MO), fundamental frequency control FT and 3 transition times for pitch, formant frequencies and amplitudes (TP,TF,TA) [6]. Every control parameter is determined by one byte per 10 msek.

2.KNOWLEDGE BASE

Knowledge base is a specialized body of knowledge (facts, relationships and rules) embodied in computer memory.
Aquisition and maintenance of a domain-specific knowledge body is a critical problem for all knowledge-based systems [7]. Just putting an initial knowledge base together in a suitable representation for experts seems a formidable task. Moreover, the system must offer powerful and at the same time quite simple tools from the point of view of nonprogramming linguists for keeping the knowledge base accurate and current.
Our system works with linguistic knowledge encoded in the bases of declarations and production rules. The production rule has the form:

IF <condition> THEN <action>

The knowledge base is structured in the way which takes into consideration both logic of fonematic description and the need to process in real time the descriptions obtained. The knowledge representation and the structure of knowledge base to be filled should be sufficiently comprehensible to the domain expert. He will be familiar with the fundamental structure, organization and use of production rules, but may understand it only at the conceptual level and not in terms of performance program.
The knowledge base consists of three main parts (see Fig.1). The first part stores parametric descriptions of phonematic units and the rules determining positional variations and coarticulation of units. The second one is for knowledge about speech prosody forming. In the third part the explanations of correspondences between symbols to determine the system of spelling, allowed in the input of the synthesizer are maintained. The vocabulary of proclitics and enclitics for phonetic word forming is foreseen as well.

ELEM correspondences between control parameters and phonematic units phonematic unit groups	MODULE B	**RULE** control parameters transformation rules
ELEM	C	**TIME** duration rules (pitch rules) (intensity rules)
CLIT vocabulary of proclitics and enclitics **ABBREV** vocabulary of abbreviations **DEF** correspondences between phonematic units and graphemes grapheme groups	A	**ORFRULE** grapheme to phonematic unit rules accent rules etc.

Fig.1 Knowledge base structure

The knowledge base contains initial pieces of knowledge - a set of graphemes, a set of internal and external representations of phonematic units and the control parameters which define the acoustics of these units. In other words the graphemes and initial sound representations must be determined to enable the unit stream in the control process.
The linguist gets the possibility to augment initial knowledge base according to his phonologic conception using some convenient formalism. The system has meta-level knowledge about every subbase in the knowledge base structure and it can help as an assistant of expert to fill out subbases with specific knowledge.
The special rule language was described by us in [8]. The translators of the rules forms from the rule text the compact intermediate rule tables for interpreters in kernel modules (see Fig.2).

3.KERNEL MODULES

The knowledge base is used by 3 functional modules of the system which perform transformations defined in production rules:

A - grapheme to phonematic unit (phoneme) transformations (process P1 on the Fig.3);
B - phonematic unit to terminal unit (allophone) transformations (process P3);
C - prosody transformations (process P2).

These modules constitute the kernel of the system. The modules are maximally independent - they are connected by unit strings which are observable by editing and explanation modules. The modules can be used separately, for example the module A was used as a phonetic transcriptor in tools of linguistic studies.

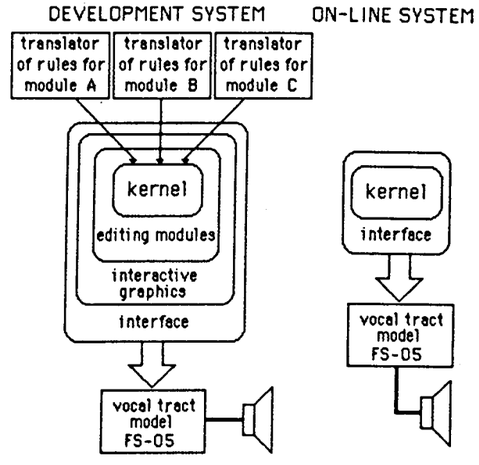


Fig.2 Structure of the development and on-line systems

Every module uses corresponding knowledge subbases and has the interpreters of rules and interpreters of declarations. The sub-

bases of declarations to module A are for:

-default correspondences between graphemes and phonematic units, grapheme grouping (base DEF);
-abbreviations (base ABBREV);
-proclitics and enclitics (base CLIT) to form phonetical word

Module B uses declarative base of default control parameters of phonematic units (ELEM).
The condition part of the record of the rule base ORFRULE for module A describes the situation in 16 bytes analysis window. This window acts as a shift register, where in addition to the grapheme codes are the indicators of the grapheme groups. The action part of production in ORFRULE makes structural changes in the string of phonematic units, derived, using the base of declarations DEF. The action part can change also the contents of analysis window - it can determine some additional flags in the indicator of grapheme grouping. For instance, 11 types of actions are needed to carry out all structural changes, specified in the rule system for Russian.
The production rule for module B (RULE base) determines the changes in control parameter domain, depending on the adjacent phonematic units. The formalism of rules in module B is simpler than that, used in parametric rules of the system SRS [4]. The analysis window of production rules for B is 3 reperesentations of phonematic units which are the pointers to the nodes of the unit (phoneme) tree. To use the tree structure describing the properties of the phonematic unit we can minimize the size and time of work of the B module rules.
In the third module C the production rules are used at present only to determine the time model of the speech - the pitch and loudness rule system is under development. Left hand side of the timing rule is similar to the condition part of the rules for module A.
To determine the segmental durations we use the formula:

$$D=D_1f_1+\ldots+D_nf_n$$

where: D - segmental duration; $f_1..f_n$ - factor, fixed by the condition part of the timing rule; $D_1..D_n$ - value of factor, determined by the action part of the rule.

D_if_i is determined by one production rule and may be interpreted for example as a factor determining the speech tempo or the inherent duration of the segments etc.
The intention was to use the same speech synthesis kernel program for both - the speech synthesis development system and the applied system. The development system has in addition a set of editing and explanation modules to examine and change the units in different stages of synthesis (see Fig.2). The last modules trace the rules evoked in transformation processes and the intermediate results of every process stage.

4.UNITS REPRESENTATION

The speech synthesis control algorithm treats different representations of units in synthesis. It is quite clear that in the process of development the internal representation of these units is hardly observed - an expert who develops rules of speech synthesis must use various means to produce the external representation of units.
The expert must have a possibility to define his own abstractions - for instance how to mark the phonematic units of speech synthesis or in which terms to fix the groups of units.
Figure 3 describes the internal and external structure of the units in speech synthesizer FS-05.

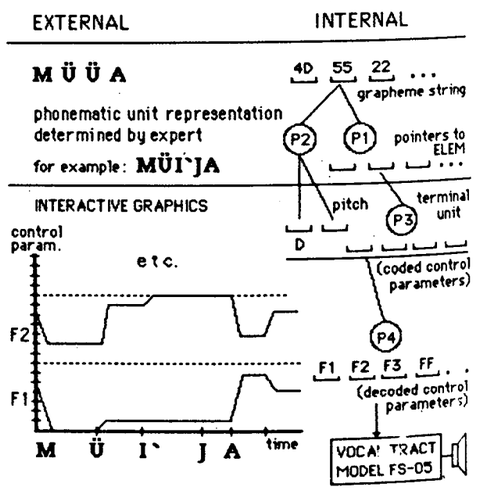


Fig.3 External and internal representation of the units

P1-P3 indicates the processes A,B,C in the kernel which was characterized above.
In synthesis process the phonematic units are the addresses to the ELEM base. As the kernel program of the synthesizer has to work in real time and must use minimal memory, the phonematic unit is internally represented by 1 byte and the control parameter base of these units (ELEM base) is situated in 256 byte ROM. For this purpose the control parameters, describing different sounds were coded and packed into one or several 4 byte control words and decoded only at the end of the control process. The maximal lenght of the ELEM base is 64 records of units (if every unit is determined by one control word). The graphical description of the parameter segment, determined by one control word is given in Fig.4.

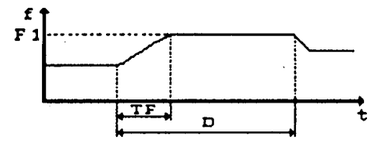


Fig.4 Graphical representation of the parameter F1 of the control word. D - duration of the segment, TF - onset transition time;

It is the task of an expert to decide - if he chose more complex units (units, described by more than one control word), phonematic units will be accordingly less than 64.
To simulate the needed variations of sounds on the acoustical level the control parameters must be changed. For testing and determining these changes in the development system the control parameters can be represented in a graphical form.
Using interactive computer graphics the domain expert can work with concrete phonetic descriptions. The graphical representation of the speech fragment is in essence the explicated input specification for the vocal tract model. It is possible to modify control parameters immediately in the course of perceptual experiments.

5.IMPLEMENTATION AND USE

Most of the production systems are realized using LISP or PROLOG language, which easily allows to describe the condition-action rules. Nevertheless, we support the viewpoint of [9] that the expert systems (and also production systems) will find more real use if they are programmed in some common language. Indeed, for example LISP needs a lot of programming resources and is usually slow. Especially for the task, described in this work, we find it important to program some parts of the production system in microprocessor ASSEMBLER.
The interpreters of productions (inference engine) were programmed in I8080 ASSEMBLER and are exactly the same for the development system, based on the microcomputer and the system for real applications (see Fig.2).
The kernel program was used to drive different vocal tract models: FS-05, formant model on the signal processor I2920 etc.
The development system runs on the personal computer LABTAM under CP/M-80 operating system and needs about 40 Kbyte of memory. Created rule language have been used by domain experts. For example the set of Russian text-to-phonematical unit rules were fixed by the phonetician of Moscow State University. This rule system with interpreter, declarative tables for 20 abbreviations and 41 clitics takes only 2 Kbytes of ROM and 512 bytes of RAM in I8080 microprocessor system.
Also the parametric rules were selected both for Russian and Estonian speech synthesis, using the aid of interactive computer graphics. The speech synthesis algorithm for Russian - the kernel program with its knowledge base and interpreters takes 6 Kbyte of ROM and 1 Kbyte of RAM of I8080 system.
The system has proved to be a powerful and flexible rule development tool which requires little resources of an ordinary microprocessor.

REFERENCES

[1] S.R.Hertz, J.Kadin, K.J.Karplus "The DELTA rule development system for speech synthesis from text" Proc. of the IEEE, 1985, vol.73, No.11, p.1589-1601
[2] J.Allen "A perspective on man-machine communication by speech" Proc. of the IEEE, 1985, vol.73, No.11, p.1541-1550
[3] R.Carlson, B.Granström "A text-to-speech system based on a phonetically oriented programming language" STL-QRSR 1/1975.
[4] S.R.Hertz "From text to speech with SRS" J.Acoust.Soc.Am., 1982, vol.72, No.4, p.1155-1170
[5] A.Aggoun et al. "Prosodic knowledge in the rule-based SYNTHEX expert system for speech synthesis" NATO ASI series, 1985, vol.F16, Springer Verlag, p.495-516
[6] О.Кюннап, А.Отт "Управляемый микропроцессором синтезатор речи" В кн.: Автоматическое распознавание слуховых образов - 12, Киев, 1982, стр.410-411
[7] R.Davis, D.B.Lenat "Knowledge-based systems in artificial intelligence", 1982, McGraw-Hill Int. Book Co.
[8] A.Ott, I.Siil "Real time speech synthesis - development and employment" Computers and artificial intelligence, Bratislava, 1987, Vol.6, No.2
[9] T.Mannel "What's holding back expert systems?" Electronics, 1986, No.28, p.59-65

RECOGNITION OF FRENCH VOWELS BY EXPERT SYSTEM SERAC

Anne BONNEAU

Centre National d'Etudes
des télécommunications
22300 Lannion (France)

Mario ROSSI

Institut de Phonétique

13100 Aix-en Provence (France)

ABSTRACT

This paper concerns two methods aiming to the automatic recognition of French vowels in continuous speech. The first part presents the results obtained by an algorithm based on the detection of context- and speaker-independent acoustic cues for the fine identification of the vowels. The second part concerns the preliminary results obtained for the detection of the features open/close and front/back, by context-independent cues and partially speaker-dependent cues (the frequency ranges on which certain rules operate are adapted to the sex of the speaker). The limits of the two methods are discussed. It is suggested that recognition of the vowels should be performed using a mixed strategy: an "invariant feature" recognition module, to classify the vowels, followed, for each vowel class, by a specific module which would partially be speaker- and context- dependent.

INTRODUCTION

The detailed recognition of vowels independently of the context and of the speakers is difficult in languages like French which has a rich vocalic vowel system (See Fig.1 the French vocalic triangle). French is generally considered as having four distinctive degrees of opening; nasality is distinctive; it has a series of front unrounded vowels and a series of front rounded vowels.
Speaker- and context-independent cues can however be used for the recognition of the most robust features of the vowels, allowing a gross classification into large vocalic classes. The description of the words by gross features is useful (at least from a computational point of view) to access the lexicon and to select a subset of words to be later verified against the signal. Such a scheme involves a mixed strategy for the recognition of the vowels. After the selection of a subset of vowels sharing the same gross feature(s), done by the use of context- and speaker- independent rules (related to the existence of invariant cues), a vowel in the subset is choosen by a specific module which uses speaker- and context- dependent knowledge. Such a mixed strategy was suggested to us by a careful examination of the results (i.e. vowel confusions) done the algorithm developped by Rossi. The algorithm and its evaluation will be presented in the first part in this paper. At the present state of knowledge, it is not really feasible to identify all the vocalic features by speaker- and context independent rules. The second part deals with the detection of the open/close and front/back features.

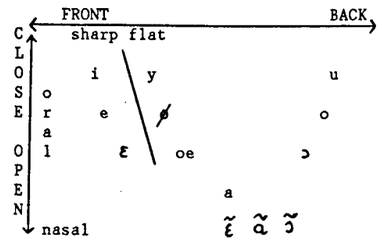


Figure 1: French vocalic triangle

I. VOWEL RECOGNITIONBY ROSSI'S ALGORITM

I.1 Algorithm and results

Speech is first process using a fourteen channel vocoder. The rules apply to the central region of the vowel. Recognition is done on a binary basis, using a tree-type structure. The main acoustic parameter is the vector of energies in each channel of the vocoder. The cues generally make comparisons between the energy levels of two specific frequency ranges in the spectrum.
For example, one of the rules is:

if EK1 >= (EK3 + EK4) then OPEN 1 = TRUE

where EK1, EK3 and EK4 represent the energy level in the first, fourth and fifth channel. The rules have been formulated by the study of 320 CVCV isolated words.
For conveniency, for testing, the rules have been implemented in the SERAC system, an expert system developped at CNET in connection with the Artificial Intelligence group.

The performance of the system has been evaluated on 20 sentences (by two male speakers), 50 connected numbers (by two male speakers) and 300 isolated numbers spoken (by six male speakers). for eacn vowel, one to three candidates are proposed. In other words, the list of candidates for each vowel never contains more than three hypotheses. In average, in 75 percent of the cases, the right vowel is one of the candidates included in the list. In 52 percent of the cases, the first candidate is the right solution (See Table 1).

TEST-CORPUS	SPEAKERS males only	LIST OF CANDIDATES % CORRECT	FIRST CANDIDATE % CORRECT
.SENTENCES (20)	2	79	55
.CONNECTED NUMBERS (50)	2	76	53
.NUMBERS (300)	6	72	50
TRAINING CORPUS: 320 CVCVC WORDS			

Table 1: Percentages of correctly recognized vowels

I.2 Discussions

Confusions occur mainly between:

(1) the back close vowels /u/ and /o/;
(2) /oe/ and its nasal counterpart /ɶ̃/;
(3) /a/ preceded by /k/, and /oe/ preceded by /r/;
(4) the three nasal vowels /ɛ̃/, /ɑ̃/ and /ɔ̃/.

Some of the errors are therefore probably due to the fact that the mid region is not stable and includes transitional movement, and to a lack of information in the very low frequencies. We think also that binary cues are not well adapted at this stage of recognition, and Klatt [4] has already spoken of the undesirability of forcing an early decision.

Although the algorithm passes directly from the cues to the vowels without a clearly defined intermediary level which would be a feature recognition module, it is interesting to evaluate the confusions appearing between vowels of opposite classes: between open and closed vowels, between front and back vowels, or oral and nasal vowels, etc...
Let us do a number of remarks on the assignement of a degree of opening and of a degree of backing to some of the vowels.
(1) The degree of opening of mid vowels -/e,ɛ,o,ɔ,ø,œ/- is not often easy to determine. In some cases, the contrast between the /e/ and /ɛ/, /o/ and /ɔ/, and /ø/ and /oe/ is distinctive. In many cases, the speaker tends to use the open vowels (/ɛ/,/ɔ/,/oe/) when there are embedded in closed syllables; on the contrary, they tend to close (/e/,/o/,/ø/) them in open syllables. Further more, despite such tendencies, it is often the case that the ear is not able to decide whether or not the vowel is open or close. Therefore the error rates presented in this paper do not take into account errors arising between the mid vowels. Two degrees of opening only (close and open) are considered.
(2) We adopt the same restrictions for the vowels /a/, /ɛ̃/, /oe/,/ø̃/, which were not a priori classified on the front-back axis.
The results, detailed in Table 2, show that there is a few number of confusions between open and closed vowels and between front and back vowels (with the restrictions explained above).

	F	B		O	C		OR	N		FL	S
F	95	3	O	99	4	OR	82	36	FL	43	89
B	5	97	C	1	96	N	18	64	S	57	11

Table 2: CONFUSION MATRIX BETWEEN VOCALIC CLASSES
Only the first vowel candidate is taken into account. The horizontal axis represents the recognized vowel class. The results are given in percentages.
F: front, B: back, O: open, C: close,
OR: oral, N: nasal, FL: flat, S: sharp

To summarize, the vowel recognition module, as developped by Rossi and presented here shows an ability to identify half of the vowels and to correctly classify all the vowels in the main vocalic classes. Context-sensitive rules and speaker adaptation seem necessary to perform finer vocalic distinctions (since there are 48 percent of errors).

From these results, we suggest a recognition process which would be:
(1) relatively speaker- and context-independent for the robust features recognition. By relatively, we mean that we do not exclude a-priori a separation between large classes of speakers - such as men, women-, or to introduce in some cases contextual rules.
(2) speaker- and context-dependent for the detailed recognition of vowels.

We are currently working on a feature recognition module. We present below our methodology and our preliminary results.

II. FEATURE RECOGNITION

The methodology remains the same as that adopted previously by Rossi with two exceptions which take into account our previous remarks:
(1) the cues are only evaluated on a single spectrum sampled in the middle of the vowel in order to minimise the influence of context or that of a poor delimitation of the boundaries of the vowel.
(2) the cues are no longer binary.

The training corpus is made of 160 logatoms of the CVCVCVC-type, where the three vowels and the three consonants are identical-, and V represents 13 French vowels and C the 16 French consonants. The data have been automatically segmented using a segmentation module. Since automatic segmentation is by no way perfect, some of segmentation errors have been manually corrected. The tests were carried out on two speakers, one male and one female.

We have tested about a dozen of rules for the open/close and front/back distinctions. Statistical tests were used to select the most discriminatory ones. Each candidate is given with a confidence score which falls between 0 and 1, according to the values of the cue.
As said previously, the degree of opening of the mid vowels is not determined a priori. These vowels are not taken into account during tests on opening cue validity and, consequently, during the evaluation of the discrimination rate of this cue. During the recognition phase, the mid vowels are automatically classified by the program as either open or closed depending on whether their values match those of open or closed vowels: this method gives an objective criterion for distinguishing between these allophones. The same strategy is adopted for /a, ɛ̃, oe, ɸ/ which are not a priori classified on the front-back axis.
We will successively present the results obtained on the training corpus itself and on another corpus.

a) Results on the training corpus.

Figure 2 shows the histograms of the acoustic correlates of the two features: the "open 1" and the "front 1" cues. Each of the two features can be identified by a single cue with an error rate lower than or equal to 3%.

Such results can be further improved in two ways:

- ADAPTING SOME FREQUENCY RANGES TO THE SEX OF THE SPEAKER:
This error rate can be brought down by 1% by adapting the frequency range considered by the rule for the "front 1" cue, depending on whether the speaker is a male or a female.

- ADDING A RULE FOR THE /i/-/u/ DISTINCTION:
For certain vowels which sometimes have a weak second formant (front /i/ and back /u/), or a very low one (/u/), the "front 1" cue isn't always well correlated to their place of articulation. A secondary cue adapted to the identification of /i/ and /u/, allows the elimination of a large number of incertainties or errors between these two vowels. In order not to lower the scores obtained with the first cue, only the values of the second cue permitting a sure identification of the feature -i.e when the confidence score is maximal- are used.

To summarize, three cues are enough to identify the open/close and front/back features with an error rate of 1% on the training corpus: a open/close cue, a front/back cue, adapted to separate two set of values for males and females, and another front/back cue only used in cases of certitude.

b) Preliminary results on the test-corpus.

The corpus is made of numbers, the tests were carried out on seven men and seven women. We suggest two ways to evaluate the performances of our set of rules:
- the error rate made when the most probable candidate is considered.
- the error rate done for the recognition of the vowels for which the confidence score is 1 (i.e maximal). Together with this rate error rate is indicated the number of vowels for which this score is obtained. It is, of course, important to get as many as possible vowels with maximal identification score, and to make the least possible errors on those vowels which have obtained maximal score.

FEATURE	ERROR RATE		NUMBER OF CANDIDATES WITH T=1
	T>.5	T=1	
FRONT/BACK	1	0	80
OPEN/CLOSE	1	0	40

Table 3: RECOGNITION RATES FOR FEATURES
T represents the confidence score.

The results confirm the results obtained on the training corpus (see table 3). The errors on the open/close feature only concern the /u/-vowel in the number "douze" (/duz/): back /u/ is identified as front. The same vowel is also responsible for the rather weak number of candidates for open vowels given with a maximal confidence score. We are presently looking for simple solutions to solve the /u/-problem. The great articulatory variations of /u/ have already been noted in dental context in French as well as in other languages [5]. The satisfactory results obtained for the front/back cue let us hope that it will be as effective on larger corpus and for a greater number of speakers.

CONCLUSION

We have proposed in this paper à mixed strategy for the recognition of French vowels : speaker- an context-independent for the recognition of the open/close and front/back features, and speaker- and context-dependent for the recognition of vowels. The aim of the features recognition module is to perform a reliable, prior classification of the vowels. This module can be used independently, principally for accessing the lexicon, or can be connected to vowel recognition modules. On a corpus made of numbers, spoken by 14 speakers, we obtain an error rate of 1% for the recognition of the open/close and front/back features and no error is made on candidates given with the highest confidence score. We may therefore conclude that our algorithm is very reliable. We are presently testing the module on a larger corpus and on a greater number of speakers.

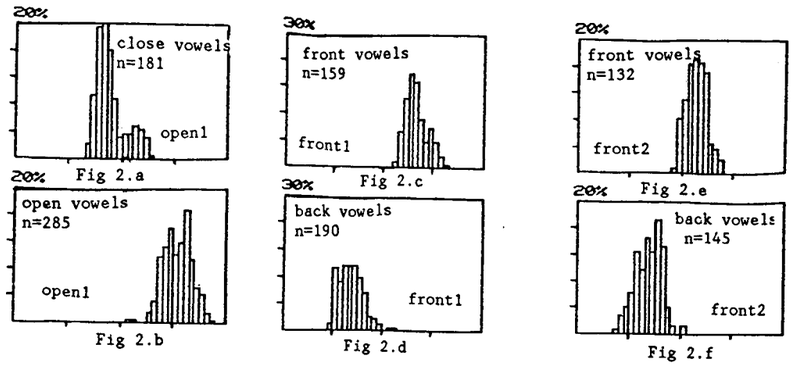


Figure 2: Frequency distribution of acoustic cues values as a function vowel class (results from the training corpus)
n: number of vowels
open1: an acoustic correlate of the open/close feature
front1: an acoustic correlate of the front/back feature
front2: an acoustic correlate of the front/back feature, the same as front1 but with a higher frequency range (adapted for female speaker)

"open1" calculates the difference between EK1 and (EK3 +EK4),
EKi represents the level of the energy in the ith channel.

"front1" calculates the difference between (EK6 + EK7) and (EK4+ EK5)

Fig 2.a, Fig 2.b: two speakers (1 M + 1 F)
Fig 2.c, Fig 2.d: male speaker
Fig 2.e, Fig 2.f: female speaker

REFERENCES

[1] D.W. Shipman, V.W. Zue, "Properties of large lexicons. Implications for advanced isolated word recognition systems", Proc. IEEE ICASSP, Paris, 1982.

[2] G. Adda, M. Eskenazi, P.E Stern, "Reconnaissance de grands vocabulaires: utilisation et evaluation de traits grossiers" Journees d'Etudes sur la Parole, Aix-en-Provence, 1986.

[3] A. Bonneau, M. Mercier, M. Gerard, M. Rossi, "Decodage acoustico-phonetique a l' aide du systeme expert Serac-Iroise" Journees d' Etudes sur la parole, Aix-en-Provence, 1986.

[4] D.H. Klatt, "Models of phonetic recognition I: issues that arise in attempting to specify a feature-based strategy for speech recognition", Proc. Montreal Symposium on speech recognition, 1986.

[5] K. Shirai, T. Kobayashi, J. Yazawa, "Estimation of articulatory parameters by table-look method and its application for speaker independent phoneme recognition", Proc. ICASSP, 1984.

ФОРМИРОВАНИЕ ЭТАЛОНОВ ДЛЯ МОРФЕМНОГО РАСПОЗНАВАНИЯ РЕЧИ

Б.В.ПАНЧЕНКО

г. Минск, 220068, СССР

Рассматривается алгоритм формирования эталонных образов в системе морфемного распознавания слитных фраз, составленных из словоформ заданного словаря. В процессе обучения осуществляется символьное сравнение текстов слов словаря и с использованием морфологических правил формируется каталог морфем. На следующем этапе осуществляется ДП-сравнение речевых образов пар слов, имеющих в своем составе одинаковую морфему, с целью определения границ морфемных фрагментов в речевых образах слов,и образуется массив речевых образов морфем. Каталог эталонов морфем формируется с использованием процедуры кластеризации речевых образов одинаковых морфем.

ВВЕДЕНИЕ

В последнее время благодаря достижениям в области распознавания речи созданы достаточно эффективные системы, пригодные для практического применения. Они способны распознавать сравнительно небольшое количество акустически изолированных речевых команд, произнесенных известным диктором. В алгоритмах этих систем используется так называемый глобальный метод [I] ,основанный на универсальных алгоритмах распознавания сравнением распознаваемого образа с эталонными, как правило, без учета особенностей речевого сигнала. В качестве речевых элементов эти системы используют речевые команды, состоящие из одного или нескольких слов, рассматриваемых как единое целое.
Практическое использование этих систем ограничивается присущими им недостатками: работа с одним диктором, ограниченность словаря команд, необходимость строгого соблюдения дикторской дисциплины, недостаточная комфортность работы с системой, необходимость пословного произнесения при распознавании фраз.
Стремление устранить отдельные недостатки глобального метода привело к разработкам алгоритмов распознавания слитнопроизнесенных команд и ключевых слов в речевом сигнале фраз. Практическое использование этих алгоритмов значительно приблизило системы распознавания к пользователю.Однако распознавание слитнопроизнесенных фраз по эталонам,построенным по отдельным словам, осуществляется с недостаточной надежностью из-за эффекта слияния слов, требует больших вычислительных затрат и поэтому эффективно только на небольших подобранных словарях.
Ограниченность глобального метода сделало необходимым развитие работ по распознаванию речи в направлении разработки другого метода, называемого аналитическим, в котором используется представление речевого сообщения как композиции относительно небольшого числа базовых речевых элементов. Суть метода заключается в принятии на первом этапе решения о минимальных базовых речевых элементах сообщения и реконструкции сообщения по этим элементам на втором этапе с использованием правил грамматики и синтаксиса.
Таким образом, система распознавания,использующая этот метод, состоит из акустического процессора, в котором принимается решение о речевых элементах, и соединенного последовательно с ним лингвистического процессора, осуществляющего реконструкцию сообщения по последовательности элементов.
В разное время разными исследователями разрабатывались и проводились исследования речераспознающих систем с использованием трехуровневых лингвистических иерархий на основе эталонных слогов, полуслогов, фонем [2] , [3],[4]. Для реконструкции слов (лексический уровень) из этих субсловесных единиц и предложений (третий уровень) применяется сложный грамматический и синтаксический анализ. Привлекательным в смысле минимальности алфавита используемых субсловесных единиц является пофонемное распознавание, однако необходимо отметить его чрезвычайную сложность,связанную с надежностью определения фонемных сегментов и вариативностью параметров фонем в речевом потоке.

МОРФЕМНОЕ РАСПОЗНАВАНИЕ

Не ставя вопрос о принципиальной достижимости высоконадежного фонемного распознавания речи, работающего в реальном масштабе времени, мы считаем, что более скорый практический выход дадут системы,использующие более крупные субсловесные единицы, такие, например, как морфемы, являющиеся минимальными значимыми частями слов.
Выбор морфемы в качестве минимальной субсловесной единицы основан на следующих соображениях.
При использовании эталонов морфем нет необходимости в явном моделировании эффектов коартикуляции, поскольку они естественным образом присутствуют в относительно длинном отрезке речевого сигнала, соответствующего морфеме, т.е. акустический образ морфемы более устойчив к аллофоническим вариациям. Использование морфемного уровня позволяет значительно расширить лексические возможности распознающей системы, поскольку многообразие парадигматических форм слов может быть получено из ограниченного каталога морфем. По данным, приведенным в словаре морфем русского языка [5] , для составления приблизительно 52000 слов, содержащихся в словаре, использовано около 5 тыс. морфем. Интересно отметить, что из них 260 наиболее продуктивных корневых морфем дают почти половину слов данного словаря (около 20 тысяч). Показателем эффективности использования морфем на первом уровне лингвистической иерархии является также индекс синтетичности языка, который определяет отношение числа морфем к числу слов в тексте, и для русского языка составляет I,90 (для английского - I,68).
Получаемый при работе распознающей системы на выходе акустического процессора кортеж основных морфем речевого сообщения позволяет эффективно использовать в лингвистическом процессоре хорошо разработанный набор относительно простых морфологических правил (правила словообразования, структурные закономерности между морфемами, фактор ветвления, показывающий число морфем, которые могут следовать за конкретной морфемой и т.д.).
Работа акустического процессора основана на разработанном Н.П.Дегтяревым [6] алгоритме поиска в речевом сигнале фразы ключевых слов заданного словаря. В нашей системе в качестве ключевых слов выступают морфемы и морфемные блоки. К используемому в акустическом процессоре каталогу морфем предъявляются требования лингвистической достаточности. Это требование должно удовлетворять используемые в лингвистическом процессоре морфологические правила, по которым осуществляется реконструкция слов. С другой сторны, к морфемам, включаемым в состав каталога, предъявляются требования достаточной акустической представительности - каждая из этих морфем должна содержать в себе не менее одного полного слога.
Акустический процессор включает в себя предпроцессор обработки речевого сигнала, формирующего вектор-параметры,значения которых регистрируются каждые секунд (отсчеты речевого сигнала), запоминающее устройство для хранения эталонов морфем и классификатор, который определяет вхождение морфем словаря в распознаваемое сообщение. Классификация морфем сообщения осуществляется средствами динамического программирования.
Для уменьшения влияния эффекта границ на надежность распознавания в процедуре классификации используется весовая функция, учитывающая тип терминальных фонем морфемы и расстояние конкретного отсчета морфемного фрагмента от его границ.
На выходе акустического процессора образуются две последовательности: I) морфемы в символьном представлении и 2) номера первых отсчетов речевых образов распознанных морфем и нераспознанных фрагментов речевого сигнала в порядке возрастания.
Надежность работы морфемного классификатора в значительной степени зависит от качества эталонов и именно алгоритму автоматического формирования эталонов морфем уделяется основное внимание в данной работе.

АЛГОРИТМ ОБУЧЕНИЯ

Процесс получения морфем ω_j^* ,включаемых в каталог (словарь) W_j^* , связан с определением границ конкретных морфов в речевой реализации слова (здесь и далее знаком астериска обозначаются речевые образы слов и морфем). В процессе обучения решается задача определения временных границ $t_к$, t_m отрезка речевого сигнала при известном фонемном содержании морфемы $\omega_j \in W$, ограниченной этими границами. Общая структура алгоритма приведена на рис. I.
Процедура обучения, целью которой является формирование каталога W^* эталонов морфем, происходит в несколько этапов и представляет собой суперпозицию алгоритмов $A = (A_1 (A_2 (A_3)))$.
На первом этапе, реализуемом алгоритмом A_1 ,осуществляется фонемное транскрибирование введенных текстов слов $v_i \in V$ словаря пользователя и на основе морфемного анализа с использованием морфологических правил формируется каталог $W = \{\omega_j\}$.
При морфемном анализе используются списки префиксальных W_ρ и суффиксальных W_S морфем (а также списки словообразующих морфемных блоков). Алгоритм A_1 осуществляет вычленение в каждом $v_i \in V$ морфем на основе символьного попарного сравнения слов и сравнения с элементами $\omega_\rho \in W_\rho$, $\omega_s \in W_S$, в результате которого v_i представляется как цепочка

$$v_i = \omega(1), \omega(2), \ldots \omega(j), \ldots \omega(\varkappa)$$

морфем, где $\varkappa$ может принимать значения от I до II [7] . Затем производится анализ морфемного состава словаря V и методом исключения формируется каталог W ,в

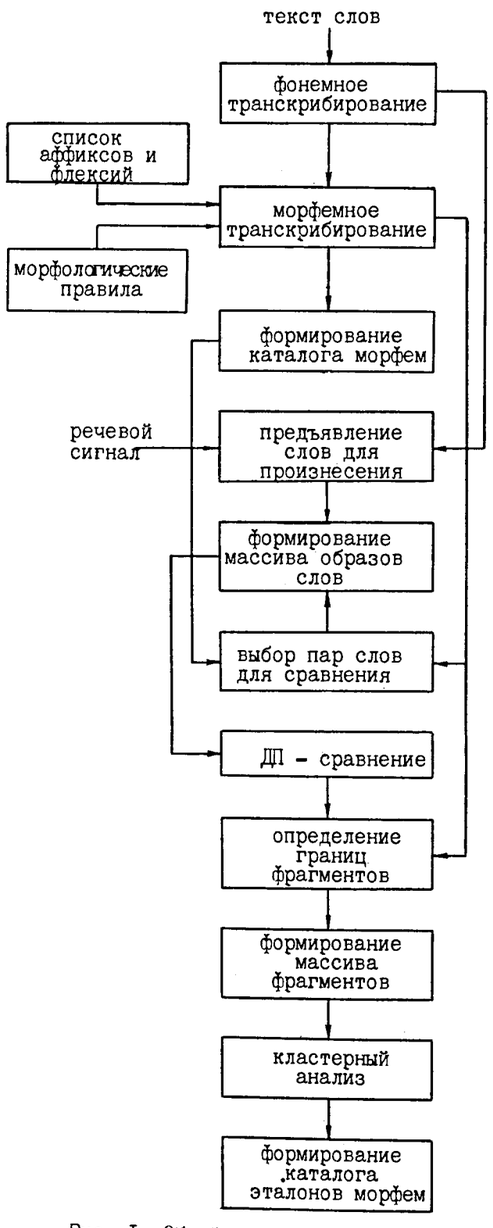


Рис. I. Общий алгоритм формирования эталонов морфем

который включаются только морфемы, удовлетворяющие требованию акустической представительности и лингвистической достаточности.

На втором этапе (алгоритм А2) процедуры обучения в систему заносятся в речевые реализации v_i^* слов словаря пользователя, т.е. формируется словарь V^*. Как обычно, речевой сигнал слова описывается последовательностью

$$v_i^* = (X_{1i}, X_{2i}; \dots X_{ji}, \dots X_{Li})$$

из элементов X_j, представляющих собой вектор в пространстве принятого параметрического описания, L – длина сигнала с некоторым шагом ΔT.

Итак, каждому слову v_i поставлен в соответствие речевой образ v_i^*. Это соответствие используется для поиска в образе v_i^* фрагментов $\omega_j^* = (X_k, \dots X_m)$, соответствующим морфемам. Если слово v_i состоит из одной морфемы, то $v_i = \omega_j$ и его образ v_i^* заносится в массив W^{**} под именем ω_j.

Для определения границ морфем в речевом сигнале многоморфемного слова используется процедура ДП-сравнения, реализуемая теми же средствами, что и в процессе классификации.

Из элементов $\{v_i\}$ выбираются пары слов v_n и v_q, отличающихся только одной морфемой и к соответствующим элементам

$$v_n^*, v_q^* \in V^*$$

применяется процедура ДП-сравнения для определения функционала меры сходства речевых сигналов v_n^* и v_q^*. При выборе пар возможны три варианта:

1) $\omega_n(I) \equiv \omega_q(I)$
2) $\omega_n(x) \equiv \omega_q(x)$
3) $\omega_n(j) \equiv \omega_q(j)$

Для первого случая используется прямое ДП-сравнение, для второго – обратное ДП-сравнение, а в третьем случае для поиска границ нетерминального морфемного фрагмента используется, как и при классификации, алгоритм поиска ключевых элементов в речевом сигнале по эталонам правой и левой терминальных морфем из каталога W^*. Поэтому процедура поиска границ морфемных фрагментов начинается с выбора и обработки пар v_n^* и v_q^*, имеющих одинаковые терминальные морфемы. При сравнении образов

$$v_n^* = (X_{n1}, X_{n2}, \dots X_{ni}, \dots X_{nn}) \quad \text{и}$$

$$v_q^* = (X_{q1}, X_{q2}, \dots X_{qj}, \dots X_{qq})$$

определяется интегральная мера сходства $D(i,q)$ на каждом шаге i и интегральная мера сходства $D(j,n)$, на каждом шаге j. Экстремум каждого из функционалов $D(i,q)$ и $D(j,n)$ меры сходства пар (v_n^*, v_q^*) и (v_q^*, v_n^*), определенный с учетом вероятностной оценки его положения по фонемному содержанию сравниваемых образов, указывает номер граничного отсчета морфемного фрагмента

$\omega_n^*(k)$ и $\omega_q^*(k)$ в речевом сигнале v_n^* и v_q^*, соответственно:

$$i = m = \arg\min_i D(i,q)$$
$$j = r = \arg\min_j D(j,n)$$

Понятно, что фрагменты $\omega_n^*(k)$ и $\omega_q^*(k)$ имеют одинаковые имена в $\{\omega\}$. Полученные при обработке всех возможных пар речевых образов слов морфемные фрагменты образуют массив W^{**}.

Формирование каталога эталонов морфем осуществляется алгоритмом A_3, реализующим кластерный анализ.

В массиве W^{**}, выделяются подмножества образов морфем $\{\omega_k^*\}_1^N$, имеющих одинаковые имена ω_k. Все N элементов подмножества $\{W_k^*\}$ поступают на ДП-сравнение для вычисления матрицы расстояний

$$\{D i,j\} \qquad i = \overline{1,N}, \quad j = \overline{1,N}$$

между образами. По полученной матрице строится обобщенный эталон для всех образов, т.е. кластер представляется одним эталоном, полученным в результате усреднения образов, образующих данное подмножество.

ЗАКЛЮЧЕНИЕ

Алгоритм формирования эталонов морфем разрабатывался для использования в системе распознавания слитнопроизносимых фраз, основанной на алгоритме принятия решений о речевых единицах в текущем речевом сигнале Н.П.Дегтярева [6]. Выбор в качестве базовых элементов морфем, которые являются носителями лексических и грамматических значений значительно расширяет лексические возможности распознающей системы, позволяет строить произносимые фразы с использованием широкого разнообразия словоформ.

Однако необходимо отметить, что использование сложного алгоритма обучения может быть целесообразно и оправдано только при работе с достаточно большими словарями.

Практическое применение систем морфемного распознавания зависит не только от надежности классификации морфем, но и в значительной степени от разработанности правил реконструкции слов и предложений.

ЛИТЕРАТУРА

1. J.M.Pierrel. Use of linguistic constraints for automatic continuous speech understanding: the Myrtille II system, Technology and Science of Informatics, vol. 1, N 5, 1983.
2. S.B.Davis and P.Mermelstain. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, 1980, pp. 357-366.
3. A.E.Rosenberg, L.R.Rabiner, J.G.Wilpon and D.Kahn. Demisyllable based isolated word recognition system. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-31, 1983, pp. 713-726.
4. K.Shicano and M.Kohla. A Linguistic processor in a convergational speech recognition system. Rev. ECL, vol. 26, 1978, pp. 1486-1504.
5. А.Н.Кузнецова, Т.Ф.Ефремова. Словарь морфем русского языка. М., "Русский язык", 1986.
6. Н.П.Дегтярев. Алгоритм распознавания слов в непрерывном сигнале. Доклад на Всесоюзном симпозиуме "Бионика интеллекта". Харьков, 1987.
7. А.И.Кузнецова. Морфемная глубина слов в русском языке. ПСЛ, 1984.

ПОВЫШЕНИЕ НАДЕЖНОСТИ РАСПОЗНАВАНИЯ СЛОВ СЛИТНОЙ РЕЧИ

Дегтярев Н.П., Левков Е.Я.

Минск, 220068, СССР

Надежность распознавания слов слитной речи, при прочих равных условиях, существенно зависит от того, насколько адекватно отображены закономерные вариации слитной речи в используемых эталонах слов и насколько точно определяются границы наилучшего подобия реализации с эталонами распознаваемых слов. В настоящей работе обсуждаются возможные пути приближения к решению названных задач. Рассматривается также одно из возможных решений задачи распознавания слов в условиях, когда к началу и (или) концу реализации слова может примыкать помеха или "чужое" слово, не входящее в заданный словарь.

ВВЕДЕНИЕ

При слитном произношении слов во фразе акустическое взаимовлияние (коартикуляция) граничных фонем может проявляться с той же силой, что и между фонемами внутри слова. Поэтому отсутствие удовлетворительной коартикуляционной модели стыковки эталонов является одним из источников ошибок при распознавании слитной речи[1]. Одним из радикальных подходов к решению этой проблемы представляется использование дифонов в качестве эталонных элементов слитной речи [2,3].
Тем не менее, использование слов в качестве единиц распознавания сегодня также привлекательно по ряду причин. Прежде всего потому, что разработано много эффективных алгоритмов автоматического формирования эталонов раздельных слов, основные принципы которых могут быть использованы и при решении задач формирования эталонов слитной речи. Но принципиально важный аргумент в пользу такого выбора заключается в том, что при использовании в качестве единиц распознавания более крупных частей речи уменьшается относительное число стыковок эталонных элементов и, как следствие, уменьшается количество ошибок, порождаемых несовершенством той или иной коартикуляционной модели стыковки эталонов.
Один из подходов к решению задачи получения эталонных элементов слитной речи состоит в их извлечении из слитно произнесенных обучающих последовательностей [4,5]. Понятно, что полученные таким способом эталоны при распознавании наиболее эффективны в окружении элементов, заданных обучающей последовательностью.
В настоящей работе обсуждаются возможности автоматического получения эталонов слов, которые учитывают краевые эффекты, возникающие при раздельном и слитном произнесении слов, и которые поэтому допускают произвольный порядок слов в контрольной последовательности без снижения эффективности распознавания.
Следующим важным вопросом является выбор задачи распознавания и связанной с ней стратегией ее решения. Известны постановка и решение обобщенной задачи распознавания слитной речи, составленной из слов заданного словаря[6,7]. Решение этой задачи основывается на двухступенчатой процедуре оптимизации сходства реализации заданной длины с эталонной последовательностью по числу, составу и порядку следования составляющих ее слов. Однако, на наш взгляд, ближе к практике такая ситуация, когда входные фразы содержат не только "свои", но и "чужие" слова, не входящие в заданный словарь, а также помехи различного рода.
В этом случае, а он и рассматривается в нашей работе, становится необходимым текущее распознавание слов заданного словаря по мере того, как они реализуются в непрерывном входном сигнале. При этом стратегия распознавания строится также, как и для распознавания раздельных слов, но допускается, что: 1) они могут примыкать друг к другу при слитном произнесении; 2) их могут разделять "чужие" слова или паузы, в том числе и такие, которые заполнены помехами различного рода. В этих условиях основной стратегии распознавания становится определение границ наилучшего подобия эталонов слов с соответствующими им участками непрерывного входного сигнала. По оценке наилучшего подобия может быть принято решение "свой-чужой" или отобраны ранжированные по значениям меры сходства претенденты на окончательное решение с учетом синтаксического анализа.

Таблица I. Характеристики начальных и конечных сегментов граничных фонем эталонов слов, отображающие краевые эффекты при их раздельном и слитном произнесении

Разбиение граничных фонем на группы		Характеристика и требуемое число вариантов начальных сегментов эталонов	Характеристика и требуемое число вариантов конечных сегментов эталонов
Гласные заднего ряда	У, О, А	1. Переходное сонорное начало 2. Стационарное сонорное начало 3. Переходное мягкое сонорное начало 4. Стационарное мягкое сонорное начало	1. Стационарное сонорное окончание
Сонорные	Э, Ы, И, М, Н, Л, Р М′, Н′, Л′, Р′	1. Переходное сонорное начало 2. Стационарное сонорное начало	1. Стационарное сонорное окончание
Дифтонги	Ю, Я, Е, Ё	1. Стационарное звонкое или фрикативное начало	1. Стационарное сонорное окончание
Звонкие фрикативные	З, З′, Ж, Й	1. Стационарное звонкое или фрикативное начало	1. Переходное звонкое окончание 2. Стационарное фрикативное окончание
Глухие фрикативные	С, Ш, Ц, Ч, Ф С′, Ш′, Х′, Ф′	1. Стационарное фрикативное начало	1. Стационарное фрикативное окончание
Взрывные	Б, Д, Г, П, Т, К, В	1. Переходное звонкое начало	1. Переходное звонкое окончание 2. Стационарное аспиративное окончание
Мягкие взрывнфе	Б′, Д′, Г′, П′, Т′, К′, В′	1. Переходное звонкое начало 2. Стационарное фрикативное начало	1. Переходное звонкое окончание 2. Стационарное фрикативное окончание
Аспиративные	Х	1. Стационарное аспиративное начало	1. Стационарное аспиративное окончание

ПРИНЦИПЫ ФОРМИРОВАНИЯ ЭТАЛОНОВ, УЧИТЫВАЮЩИХ СВОЙСТВА ГРАНИЧНЫХ СЕГМЕНТОВ СЛОВ ПРИ ИХ РАЗДЕЛЬНОМ И СЛИТНОМ ПРОИЗНЕСЕНИИ

Как мы уже отметили выше, при слитном произнесении слов имеют место неизбежные изменения параметров пограничных сегментов их реализаций, отражающих взаимовлияние (коартикуляцию) граничных фонем. Отсюда следует принципиальная необходимость учета названных явлений в эталонных описаниях слов слитной речи. Одним из возможных путей решения этой задачи является использование априорных знаний о закономерностях изменений формантно-фонемных связей под влиянием соседних фонем для автоматизации процесса формирования эталонов слитной речи. Очевидно, что наиболее подходящим для этих целей является формантно-параметрическое описание речевого сигнала[8]. В этом случае, например, оказывается возможным автоматическое формирование нужного числа эталонов слова, отображающих краевые эффекты, имеющие место как при слитном, так и при раздельном произнесении слов, по одной реализации отдельно произнесенного слова[9]. Основные элементы работы такого алгоритма состоят в следующем.
Входной информацией является формантно-параметрическое описание реализации слова и его фонемная транскрипция. Алфавит граничных фонем разбивается на группы (см. табл. I), характеризующиеся одинаковым способом образования или общими акустическими свойствами в пространстве параметров описания, связанными с позицией фонемы - начальная/конечная. Соответственно на первом этапе определяется, к каким группам принадлежат начальная и конечная фонемы в транскрипции слова.
Алфавит фонем, сочетаемых с граничными фонемами, разбивается на группы, характе-

ризующиеся единством акустического влияния на граничные фонемы: 1) паузы, взрывные; 2) мягкие согласные, дифтонги; 3) звонкие; 4) глухие. Поэтому на втором этапе в зависимости от групповой принадлежности граничных фонем и возможного влияния на них сочетаемых фонем, определяются варианты начальных и конечных сегментов эталонов (см. табл. I) и соответствующие правила их формирования.
На последнем этапе в соответствии с заданными правилами из принятой реализации слова формируются эталоны, отличающиеся требуемыми вариантами начальных и конечных сегментов. Требуемые модификации граничных сегментов эталонов получаются путем фильтрации и отбрасывания части отсчетов или части параметров на определенном временном интервале описания реализации или путем изменения значений части параметров реализации слова. На одно слово может быть сформировано один, два или четыре эталона. Сформированные таким образом эталоны учитывают наиболее выраженные краевые эффекты реализаций слов как при раздельном, так и при их слитном произнесении и могут поэтому использоваться в едином алгоритме распознавания раздельно и слитно произносимых слов, основные принципы построения которого излагаются в нижеследующем разделе.

АЛГОРИТМ РАСПОЗНАВАНИЯ СЛОВ В ТЕКУЩЕМ РЕЧЕВОМ СИГНАЛЕ

Исходя из обсуждавшихся во введении предпосылок ставится задача распознавания слов из данного словаря N по мере их реализации в неизвестном текущем сигнале

$$X=\{x(i)\};\ i=\overline{1,I}, \qquad (1)$$

где $x(i)=\overline{p}(i)$ - отсчет вектора p-параметров описания сигнала в i-й момент дискретного $i\Delta t$ времени. Каждый из эталонов $n\in N$ описывается последовательностью $y(j)$ отсчетов тех же p-параметров описания

$$Y(n)=\{y(j)\};\ j=\overline{1,L(n)} \qquad (2)$$

где $L(n)$ - длина (число отсчетов) n-го эталона.
В качестве основы для конструирования мер сходства между эталонами и текущим сигналом в данной работе используется симметричный DP-алгоритм, предложенный в [I0] для расчетов частичного расстояния $g(j,i)$ в точке (j,i) на фазовой плоскости сопоставления образов. Основные принципы решения поставленной задачи приведены в [II] и сводятся к следующему.
На первом этапе формируются N функций $R(n,v(i),i)$, отражающих текущие расстояния начальных сегментов длиной $\ell<L(n)$ каждого из $n\in N$ эталонов к сигналу на его i-ом отсчете

$$R(n,v(i),i)=g(n,\ell,v(i),i), \qquad (3)$$

где $g(\cdot)$ - расстояние, определяемое DP-методом для сигналов с неизвестным началом, когда переход к каждому следующему отсчету реализации производится с нулевой мерой сходства; $v(i)$ - отсчет, с которого начинается оптимальная траектория на фазовой плоскости сопоставления эталона и реализации на ее i-ом отсчете [I2]. Для каждого случая, когда

$$R(n,v(i))=\min_{i} R(n,v(i),i)<PR \qquad (4)$$

где PR - пороговое значение, формируется массив эталонов $\forall\{m\}\exists R(n,v(i))<PR$, а также определяется отсчет $i_н$ вероятностного начала реализации слова

$$i_н=v(i)=\arg\min_{n} R(n,v(i)) \qquad (5)$$

Тем самым предполагается, что на i-ом отсчете обнаружен отрезок сигнала $(i_н,i)$ близкий к начальным сегментам отобранных $\{m\}$ эталонов настолько, что можно предположить начало реализации одного из $n\in\{m\}$ слов.
Прежде чем перейти ко второму этапу алгоритма заметим, что используемая нами на первом этапе мера (3) для оценки близости начальных частей эталонов к текущему сигналу не может быть использована для оценки близости к сигналу эталонов слов в целом, т.к. мера (3) оказывается близкой для таких пар слов, когда одно из них является конечной частью другого. На втором этапе алгоритма, начиная с отсчета $i_н$ реализации для $\forall n\in\{m\}$ эталонов определяются текущие интегральная

$$D(y(n),X)=D(n,i)=g(L(n),i) \qquad (6)$$

и локальная [I3]

$$Q(n,i)=\min_{j} g(j,i);\quad j=\overline{1,L(n)} \qquad (7)$$

меры сходства. По найденным $Q(n,i)$ для тех же эталонов вычисляются интегральные меры удаления (различия) эталонов от реализации

$$G(n,i)=G(n,i-1)+[Q(n,i)-P(i)], \qquad (8)$$

где $P(i)$ - адаптивный к числу $(i-i_н)$ порог. Определяются также претенденты на конечный отсчет $i_к$ реализации слова

$$i_к(n)=\arg\min_{i} D(n,i) \qquad (9)$$

и оценки значений интегральной меры близости

$$D(n)=\min_{i} D(n,i). \qquad (10)$$

Второй этап алгоритма заканчивается, когда

$$\forall n\in\{m\}\exists\ G(n,i)<PG,$$

где PG - пороговое значение, или найдена неречевая пауза.
На третьем, последнем, этапе алгоритма по полученным на предыдущем этапе данным принимаются результирующие решения. При выполнении условия

$$\min_{n} D(n)<PD(n), \qquad (11)$$

где $PD(n)$ - пороговое значение, определяется номер эталона

$$\hat{n}=\arg\min_{n} D(n), \qquad (12)$$

который указывает распознанное слово, и по найденным $\hat{n}$ и $i_к(n)$ - отсчет $i_к$, указывающий конец распознанного слова, после чего осуществляется переход на $(i_к+1)$ отсчете реализации к первому этапу следующего цикла распознавания. В случае не выполнения условия (II), принимается решение об отказе, возврат к $(i_н+1)$ отсчету реализации и переход к первому этапу следующего цикла распознавания.

ЗАКЛЮЧЕНИЕ

На основе описанных алгоритмов построена аппаратурно-программная модель распознавания, характеризующаяся следующими свойствами:
I) в одном режиме работы системы реализуется распознавание как раздельно, так и слитно произносимых слов. Это стало возможным, с одной стороны, благодаря тому, что эталоны слов отображают основные характерные свойства граничных сегментов раздельно и слитно произносимых слов, а, с другой, - благодаря определению границ слов непосредственно в процессе распознавания по текущим оценкам близости элементов эталонов к сигналу;
2) распознаваемые слова могут быть разделены "чужими" словами или помехами, которые отвергаются по пороговым критериям;
3) не накладывается принципиальных ограничений на число и порядок следования слов во фразе.
В связи с последним отметим, что при необходимости данная система распознавания может быть дополнена уровнем синтаксического анализа и тогда алгоритм распознавания может быть модифицирован в алгоритм отбора последовательностей эталонов, ранжированным по оценкам меры сходства.
Проведенные экспериментальные исследования в полной мере подтверждают эффективность предложенных решений в рамках настоящей модели распознавания слов слитной речи. Дальнейшее развитие модели связывается, в первую очередь, с совершенствованием методов получения параметров описания речевого сигнала и уточнением и расширением правил учета коартикуляционных явлений на стыках слов слитной речи.

ЛИТЕРАТУРА

I. S.E.Levinson. Structural methods in automatic speech recognition. Proc. of the IEEE, 1984, vol.73, №11, pp.1625-1650.
2. Слуцкер Г.С., Старостина Э.А. Автоматическая выработка эталонов звуковых диад. Труды Акустического института, вып. XII, 1970, с. 3I-42.
3. H.Ney, D.Mergel, S.Macrus. On the automatic training of phonetic units for word recognition. IEEE Trans. Acoust. Speech and Signal Process., 34, N1, 1986, pp. 209-213.
4. L.R.Rabiner, A.Bergh, J.G.Wilpon. An improved training procedure for connected digit recognition. Bell. Syst. Tch. j., vol. 61, 1982, pp. 981-1001.
5. М.А.Абдуллаев, Ю.Н.Жигулевцев, В.И.Спорыш. Алгоритм формирования эталонов для распознавания слитных фраз. В кн.: Автоматическое распознавание слуховых образов (АРСО-I4). Ч. I, Каунас, I986, с. 84-85.
6. H.Sakoe. Two level DP-matching - A dynamic programming based patter matching algorythm for connected word recognition. IEEE Trans. Acoust. Speech, Signal Process., vol.ASSP-27, 1979, pp. 588-595.
7. Т.К.Винцюк. Обобщенная задача распознавания слитной речи. В кн.: Автоматическое распознавание слуховых образов (АРСО-I2). Киев, 1982, с. 345-348.
8. Н.П.Дегтярев. Двухформантная аппроксимация спектров речи. В кн.: Автоматическое распознавание слуховых образов (АРСО-I4). Ч. I, Каунас, I986, с. I2-I3.
9. Н.П.Дегтярев Н.П. Использование формантно-фонемных связей для формирования эталонов слитной речи. Тезисы докладов Всесоюзного симпозиума "Бионика интеллекта". Харьков, I987.
I0. Г.С.Слуцкер. Нелинейный метод анализа речевых сигналов. Труды НИИР, № 2, I968, с. I8-23.
II. Н.П.Дегтярев. Алгоритм распознавания слов в непрерывном сигнале. Тезисы докладов Всесоюзного симпозиума "Бионика интеллекта". Харьков, I987.
I2. Б.М.Лобанов, Г.С.Слуцкер, А.П.Тизик. Автоматическое распознавание звукосочетаний в текущем речевом сигнале. Труды НИИР, № 4, I969, с. 67-75.
I3. Н.П.Дегтярев, Б.М.Лобанов, Г.С.Слуцкер. О двух вариантах построения устройств распознавания речевых команд. - В кн.: Автоматическое распознавание слуховых образов (АРСО-I0). Тбилиси, I978, с. I99-200.

EIN DYNAMIC-PROGRAMMING-ALGORITHMUS ZUR ANWENDUNG IN DER AUTOMATISCHEN SPRACHVERARBEITUNG

FRANK W. A CAMPO

Ruhr-Universität Bochum
Sprachwissenschaftliches Institut
4630 Bochum, P.O.B. 10 21 48, BRD

ABSTRACT

Eine Variante des Dynamic-Programming-Algorithmus' von Sakoe und Chiba wird experimentell untersucht. Es zeigt sich, daß durch sie die Rechenzeit bei der Spracherkennung erheblich gesenkt werden kann.

EINLEITUNG

Gegeben seien die Sprachsignale A und B. Zu gewissen Zeitpunkten sollen von ihnen Meßdaten erhoben werden, die nach Aufbereitung als Vektoren vorliegen. Der Abstand zwischen dem i-ten Meßdatenvektor von A und dem j-ten Meßdatenvektor von B sei durch eine Metrik d gegeben und werde mit d(i,j) bezeichnet. Die Menge der Paare (i,j) von auf A bzw. B definierten Meßdatenvektoren sei mit Ω bezeichnet. Fassen wir Ω als Menge von Punkten in der Ebene auf, bildet Ω ein Rechteck, und jeder Weg über Punkte aus diesem Rechteck läßt sich als Transformation der Zeitachsen von A und B auffassen. Für einen endlichen Weg in Ω läßt sich der mittlere gewichtete Abstand der Punkte (i,j), über die er führt, als Abstand zwischen A und B unter dieser Zeitachsentransformation interpretieren. Ein Durchbruch in der automatischen Spracherkennung bestand darin, daß für gewisse Mengen von Wegen ein rekursiver Optimierungsalgorithmus ("Dynamic Programming", "DP") gefunden wurde, der den minimalen Abstand zwischen den Sprachsignalen bezüglich dieser Mengen ermittelte. In diesem Beitrag möchte ich die von mir in [1] entwikkelte zweite Variante des schon als klassisch zu bezeichnenden DP-Algorithmus' von Sakoe und Chiba [2] untersuchen. Die gemeinsame Theorie der Algorithmen habe ich in [1] dargelegt; dort findet sich auch die Behandlung vieler Fragen, die ich hier nur berühren kann.

ALGORITHMEN

Wir definieren die Mengen von Wegen in Ω mittels Eigenschaften der in ihnen enthaltenen Wege. Diese Eigenschaften müssen einer theoretischen und einer praktischen Anforderung genügen: der praktischen, daß das Minimum bezüglich der durch sie bestimmten Wegmenge sich in annehmbarer Zeit ermitteln läßt, und der theoretischen, daß sich die Eigenschaften der Wege als physikalisch sinnvolle Eigenschaften von Zeitachsentransformationen deuten lassen. Folgende gängige Wegeigenschaften (i. f. We) sollen hier betrachtet werden:

We1) Jeder Weg beginnt mit (1,1) und endet mit (I,J), wobei I bzw. J die Anzahl der auf A bzw. B definierten Meßpunkte sei.

We2) Führt ein Weg hintereinander über die Punkte (i_1,j_1), (i_2,j_2), muß gelten:
- i) $(i_1,j_1) \neq (i_2,j_2)$;
- ii) $0 \leq i_2 - i_1 \leq 1$;
- iii) $0 \leq j_2 - j_1 \leq 1$.

We3) Für ein fest gewähltes $n \in \mathbb{N}$ darf ein Weg höchstens zwischen n+1 aufeinanderfolgenden Punkten parallel zu einer Rechteckseite verlaufen.

We4) Kein Weg darf rechtwinklig abknicken.

We1 und We2 bewirken, daß die Reihenfolge der Meßpunkte unter Zeitachsentransformation erhalten bleibt und keiner der Meßdatenvektoren vernachlässigt wird.
We3 beinhaltet, daß ein einem Meßdatenvektor zugeordneter Signalabschnitt im zeitachsentransformierten Signal um höchstens das (n+1)-fache verlangsamt wird. Durch diese Wegbedingung wird die Menge der Punkte aus Ω, über die ein Weg führen kann, eingeschränkt: Alle Punkte liegen in einem Parallelogramm bzw. auf dessen Rand, wobei die Seiten des Parallelogrammes die Steigungen n+1 bzw. 1/(n+1) haben und seine spitzen Ecken auf (0,0) und (I+1,J+1) liegen. Die Menge dieser Punkte sei mit P_n bezeichnet und als nicht leer angenommen. Dann gibt es für jedes i mit $1 \leq i \leq I$ ein minimales j und ein maximales j' so, daß (i,j) und (i,j') Elemente von P_n sind. Sie seien mit jmin(i) und jmax(i) bezeichnet.
Für We4 eine physikalische Begründung zu finden, war mir nicht möglich und ist meines Wissens bisher auch nicht versucht worden. We4 wurde schon von Sakoe und Chiba einzig unter praktischen Gesichtspunkten eingeführt: "This new constraint reduces the number of paths to be searched" [2]. Die Menge der Punkte aus Ω, über die ein Weg führen kann, ist bis auf höchstens zwei Punkte mit P_n identisch (s. [1], Kapitel 4); die häufig zu lesende Behauptung, die Ecken des Parallelogrammes lägen auf den Punkten (1,1) und (I,J), ist falsch.

Der DP-Algorithmus zur Ermittlung des Minimums bezüglich der durch We1 und We2 definierten Wegmenge ist gut bekannt. Er sei im folgenden V-Algorithmus genannt. Der Algorithmus für die durch We1, We2 und We3 definierte Wegmenge ist der zweite der von mir in [1] entwickelten. Er ist es, dem hier unser besonderes Interesse gilt. Ich bezeichne ihn als F_n-Algorithmus. Den Algorithmus für die durch We1 bis We4 bestimmte Wegmenge nenne ich S_n-Algorithmus. Mit einigen hier belanglosen Abweichungen handelt es sich um den DP-Algorithmus von Sakoe und Chiba.
Im folgenden mache ich von einer bewährten Gewichtsfunktion Gebrauch: d(i,j) wird mit dem Faktor 2 gewichtet, wenn (i,j) in einem Weg diagonal erreicht wird, sonst mit dem Faktor 1. Die Minima lassen sich auf folgende in Fortran-naher Notation dargestellte Arten berechnen:

V-Algorithmus (We1 und We2):

```
    ko(i,j) = ∞ f.a. (i,j) ∉ Ω
    ko(0,0) = 0
    DO 20 i=1,I
    DO 10 j=1,J
    x = d(i,j)
10  ko(i,j) = x + min { ko(i-1,j), ko(i-1,j-1) + x,
                        ko(i,j-1) }
20  CONTINUE
    Minimum = ko(I,J) / (I + J)
```

F_n-Algorithmus (We1, We2 und We3):

```
    ko(i,j) = ∞ f.a. (i,j) ∉ P_n
    ko(0,0) = 0
    g_μ(j) = ∞ f.a. 0 ≤ μ ≤ n und 1 ≤ j ≤ J
    DO 20 i=1,I
    f_μ = ∞ f.a. 0 ≤ μ ≤ n
    DO 10 j=jmin(i),jmax(i)
    x = d(i,j)
    g_μ(j) = g_μ-1(j) + x mit μ = n, n-1, ... , 1
    f_μ = f_μ-1 + x mit μ = n, n-1, ... , 1
    y = ko(i-1,j-1) + x + x
    g_0(j) = min { y, min { f_μ } }
                  1≤μ≤n
    f_0 = min { y, min { g_μ(j) } }
                  1≤μ≤n
10  ko(i,j) = min { g_0(j), f_0 }
20  CONTINUE
    Minimum = ko(I,J) / (I + J)
```

S_n-Algorithmus (We1, We2, We3 und We4):

```
    ko(i,j) = ∞ f.a. (i,j) ∉ P_n
    ko(0,0) = 0
    g_μ(j) = ∞ f.a. 0 ≤ μ ≤ n und f.a. 1 ≤ j ≤ J
    DO 20 i=1,I
    f_μ = ∞ f.a. 0 ≤ μ ≤ n
    DO 10 j=jmin(i),jmax(i)
    x = d(i,j)
    g_μ(j) = g_μ-1(j) + x mit μ = n, n-1, ... , 1
    f_μ = f_μ-1 + x  mit μ = n, n-1, ... , 1
    g_0(j) = ko(i,j) + x + x
    f_0 = g_0(j)
10  ko(i,j) = min { f_0, min { f_μ, g_μ(j) } }
                  1≤μ≤n

20  CONTINUE
    Minimum = ko(I,J) / (I + J)
```

Es sei noch angemerkt, daß bei den hier wiedergegebenen Verfahren zur Berechnung der Minima zur Verwaltung der ko(i,j) eigentlich nur J+2 Speicherplätze erforderlich sind (s. [1], Kapitel 10). Die Darstellung mit (I+1)(J+1) Speicherplätzen ist aber anschaulicher.
Zur Abschätzung der vom S_n- und F_n-Algorithmus benötigten Rechenzeiten machen wir zwei Annahmen: Eine Addition soll ebensoviel Rechenzeit erfordern wie eine Minimierung über eine zweielementige Menge, und eine Minimierung über eine m-elementige Menge ebensoviel wie m-1 Minimierungen über eine zweielementige Menge, $m \in \mathbb{N}$. Dann benötigt der S_n-Algorithmus pro $(i,j) \in P_n$ 4n+2 Operationen mit Dauer einer Addition und der F_n-Algorithmus 4n+3. (Zur Berechnung von Indizes nötige Operationen seien vernachlässigt.) Der prozentuale Mehrbedarf an Rechenzeit pro $(i,j) \in P_n$ bei der Berechnung des F_n-Minimums gegenüber der des S_n-Minimums ist in Tabelle 1 für einige n wiedergegeben.

allgemein	n=1	n=2	n=3	n=4
100/(4n+2)%	17%	10%	7%	6%

Tabelle 1: Prozentualer Mehrbedarf an Rechenzeit pro $(i,j) \in P_n$ bei der Berechnung des F_n-Minimums gegenüber der des S_n-Minimums.

Aus praktischen Gründen ist die Einführung von We4 also nicht zwingend. Folgendes ist aber noch zu beachten: Die geringe Anzahl der pro $(i,j) \in P_n$ benötigten Operationen ist mit der rekursiven Berechnung der Koeffizienten f_μ und $g_\mu(j)$ verbunden. Diese hat aber zur Folge, daß bei der Berechnung des F_n- und S_n-Minimums P_n ausgeschöpft wird. Bei Berechnungsverfahren, bei denen dies vermieden wird [3], steigt die Anzahl der pro errechnetem ko(i,j) benötigten Operationen stark an, und zwar bei Berechnung des F_n-Minimums etwas stärker als bei der des S_n-Minimums. (Dabei muß man an Stelle des originalen Algorithmus' von Sakoe und Chiba die erste in [1] entwickelte Variante verwenden.) Welche Berechnungsstrategie günstiger ist, die hier vorgestellte exhaustive oder die aus [3], muß die Praxis entscheiden.
Im folgenden sollen die Leistungen des V-, F_n- und S_n-Algorithmus' bei der Spracherkennung in einem Experiment untersucht und miteinander verglichen werden.

SPRACHMATERIAL UND DATENAUFBEREITUNG

Als Sprachmaterial wurden die Zahlen von Null bis Sechs

sowie Acht und Neun verwendet. Sie wurden von drei ca. 25-jährigen männlichen Sprechern des Hochdeutschen in einer schallisolierten Aufnahmekabine je zweimal auf Tonband gesprochen. Die Aufnahmen wurden bandpaßgefiltert mit den Abschneidefrequenzen 75 Hz und 5 kHz, mit einem 10Bit-AD-Wandler bei einer Abtastrate von 10kHz digitalisiert und auf Magnetplatte gespeichert. Daraufhin wurde für jedes Wort eine lückenlose Folge von Leistungsspektren mittels einer 256-Punkte-FFT errechnet, wobei die Fensterausschnitte mit einem Blackman-Harris-Fenster [4] gewichtet wurden. Anfang und Ende eines jeden Zahlwortes bestimmte ich manuell mit Hilfe des SONATA-Programmsystems [5]. Die Anzahl der für ein Zahlwort berechneten Spektren schwankte zwischen 15 und 27; im Mittel betrug sie 21,1. Der Gleichanteil der Spektren wurde im folgenden vernachlässigt. Jedes der Spektren normierte ich bezüglich seiner Gesamtenergie. Die normierten Spektren verwendete ich als Meßdatenvektoren und erstellte zu jedem Paar (A,B) von Zahlwörtern die Distanzmatrix $D_{A,B} = (d_{A,B}(i,j))$, wobei ich als Metrik d den mittleren euklidischen Abstand zwischen zwei normierten Spektren wählte.

EXPERIMENTE UND AUSWERTUNG

Ich führte drei Experimente durch, die ich auf zwei verschiedene Arten bezüglich des V-, F_n- und S_n-Minimums mit $1 \le n \le 4$ auswertete. Im ersten Experiment (E1) berechnete ich die Minima für die Distanzmatrizen, im zweiten (E2) verkleinerte ich die Distanzmatrizen, indem ich jede zweite Zeile und Spalte aus ihnen strich und die Minima bezüglich der verkleinerten Matrizen berechnete. Im dritten Experiment (E3) erstellte ich für jedes Spektrum i eines jeden Zahlwortes A eine Pseudo-Zufallszahl $r_A(i)$, die mit der Wahrscheinlichkeit 1/6 den Wert 4/5, mit gleicher Wahrscheinlichkeit den Wert 5/4 und mit der Wahrscheinlichkeit 2/3 den Wert 1 annahm. Für alle Paare (A,B) von Testworten berechnete ich nun die gestörte Distanzmatrix $\Delta_{A,B} = (\delta_{A,B}(i,j))$ mit $\delta_{A,B}(i,j) = d_{A,B}(i,j)\, r_A(i)\, r_B(j)$. Man überzeugt sich leicht, daß in jedem $\Delta_{A,B}$ die Hälfte der Elemente gegenüber $D_{A,B}$ verändert ist.
Bei jedem Experiment erhielt ich pro Testwort 53 Abstände zu von ihm verschiedenen Testwörtern. Um sprecherabhängige Effekte zu vermeiden, schied ich die vom jeweils selben Sprecher stammenden Testworte aus; es verblieben 36 Referenzworte pro Testwort. Die Worterkennung führte ich auf die beiden folgenden Arten durch: Beim MIN-Verfahren wies ich das Testwort dem Zahlwort zu, unter das das Referenzwort mit dem kleinsten Abstand zum Testwort fiel; beim MINQUER-Verfahren diente der minimale mittlere Abstand zu den unter ein Zahlwort fallenden Referenzworten als Kriterium.

INTERPRETATION DER VERSUCHSERGEBNISSE

An den in Abbildung 1 wiedergegebenen Erkennungsraten fällt zunächst auf, daß sie vergleichsweise niedrig sind. Das ist eine Folge der nur rudimentären Datenaufbereitung. Dennoch glaube ich, daß sich aus den Erkennungsraten einige Aussagen über die von mir verwendeten Algorithmen ableiten lassen.

- Bei Verwendung gestörter Distanzmatrizen werden beim MINQUER-Verfahren die Erkennungsraten der S_n- und F_n-Algorithmen nivelliert.
- Die schlechtesten Erkennungsraten liefert der V-Algorithmus. Die vom S_n-Algorithmus her bekannte Tendenz, mit steigendem n schlechtere Ergebnisse zu erbringen, findet sich auch beim F_n-Algorithmus.
- Für $1 \le n \le 2$ gilt: Für die S_n -Minima sinken die Erkennungsraten beim Übergang von vollständigen Distanzmatrizen (E1) zu verkürzten (E2) stark ab, während sich die F_n-Minima nicht wesentlich verändern.

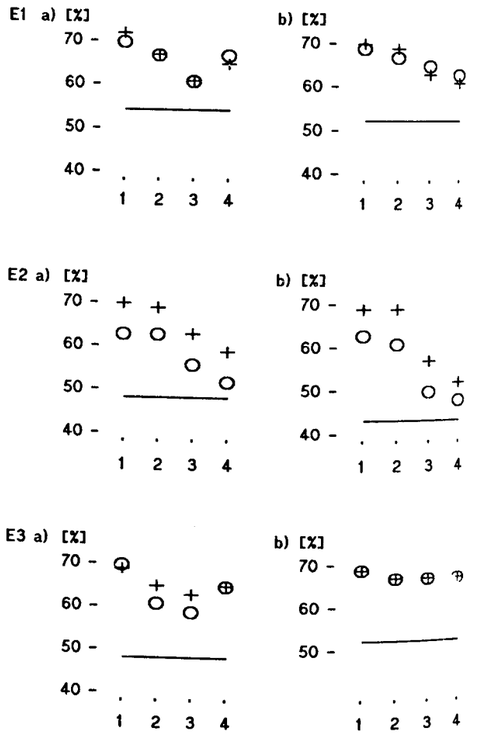


Abbildung 1: Experimentell ermittelte Erkennungsraten.
Abszisse: n, Ordinate: Erkennungsrate in Prozent;
a) MIN-Verfahren, b) MINQUER-Verfahren;
—: V-Algorithmus, +: F_n-Algorithmus,
O: S_n-Algorithmus.

Da bei E2 infolge der Verkürzung der Distanzmatrizen die Anzahl der in ihnen enthaltenen Elemente auf ca. ein Viertel sinkt, erniedrigt sich die Mächtigkeit der jeweiligen P_n auf ungefähr ein Viertel bis die Hälfte, je nach dem Verhältnis von I zu J.
Mit aller wegen der geringen Datenmenge gebotenen Vorsicht sei also folgende Hypothese aufgestellt: Für $1 \le n \le 2$ können wir beim F_n-Algorithmus die Rechenzeit senken, ohne die Erkennungsrate nennenswert zu beeinflussen, indem wir verkürzte Distanzmatrizen an Stelle der vollständigen verwenden. Beim S_n - Algorithmus haben wir diese Möglichkeit nicht. Trotz der etwas größeren Rechenzeit für eine Einzelminimierung (s. Tabelle 1) können wir also für $1 \le n \le 2$ mit dem F_n - Algorithmus bei der für die Worterkennung nötigen Gesamtminimierung Rechenzeit gegenüber dem S_n -Algorithmus sparen. Für die angegebenen Verkleinerungen von P_n ergibt sich aus den oben angegebenen für ein ko(i,j) jeweils benötigten Operationen eine Ersparnis von ca. 42% bis 71% für den F_1-Algorithmus und zwischen ca. 45% bis 73% für den F_2-Algorithmus.

LITERATURNACHWEISE

[1] F. W. a Campo (1987): Varianten des Dynamic-Programming-Algorithmus' von Sakoe und Chiba, IPKöln-Berichte 14, in Vorbereitung.
[2] H. Sakoe, S. Chiba (1978): Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol ASSP-26, 43-49.
[3] M. K. Brown, L. R. Rabiner (1982): An Adaptive, Ordered Graph Search Technique for Dynamic Time Warping for Isolated Word Recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-30, 535-544.
[4] F. J. Harris (1978): On the use of windows for harmonic analysis with the discrete Fourier Transformation, Proc. IEEE 66, 51-83.
[5] J. E. Philipp, F. W. a Campo (1982): SONATA-Handbuch: Bedienungsanleitung zum **SONA**gram and Time Signal **A**nalyser, IPKöln-Berichte Nr. 12, 7-42.

SPEECH RECOGNITION SYSTEM BASED ON WALSH FUNCTION

F.Y. KORKMAZSKY

Computer Design Office
Chernovtsy, USSR 274030

ABSTRACT

The principles of designing a speech recognition system based on Walsh functions are described. For the initial speech signal description the energy spectrum of the Walsh transform with Hadamard ordering is proposed. The advantages of the above signal representation system are its invariance under cyclic shift of signals and a high processing speed of computing the energy spectrum of signals. It is shown that a posteriori informativeness coefficients used in training and recognition procedures give a considerable increase in speech signal recognition rate. A method of reference speech pattern correction at the recognition stage is proposed which should also contribute to a higher speech recognition rate.

INTRODUCTION

Most of the present speech recognition systems using the spectral speech signal representation are designed on the basis of Fourier transform. It has gained its popularity among researchers due to development of fast Fourier transform (FFT) algorithms which had helped to increase the processing speed for computation of signal spectra. Thus, to obtain a complex Fourier spectrum using the FFT, $N\log_2 N$ of complex additions and $N/2\log_2 N$ of complex multiplications are required. The principal advantage of the Fourier transform method is the invariance of Fourier energy spectrum under the cyclic shift of the input signal. This property of the Fourier transform enables one to obtain energy spectra which are independent of the phase of the processed signal. However, the realization of the FFT algorithm would require either special equipment performing the FFT, or the use of integrated circuits for processing of spectra by means of FFT. A microprocessor-based program realization of FFT is difficult because real time performing of the FFT algorithm is not possible, especially when the number of processing points is large. On the other hand, there are some orthogonal transforms which neither require processing in the complex plane, nor the use of multiplication. Therefore, such transforms are performed much faster than FFT, and lend themselves to microprocessor-based realization. An important transformation of this kind is the Walsh transform. A Walsh function is a complete orthonormal set of functions assuming either +1 or -1 values. Thus, the Walsh transform which cosists in a convolution of input signals with Walsh functions, requires only two operations, i.e. addition and subtraction, and does not require multiplication. There are several varieties of Walsh transforms/1/ for which corresponding fast Walsh transform (FWT) algorithms can be used. The FWT algorithms usually require $N/2\log_2 N$ real additions or subtractions which allows real time realization of these algorithms on the basis of modern microcomputers.

Most of the present speech recognition systems using the Walsh functions for describing speech signals give a high speed of obtaining the spectral description of speech signals. However, the vocabulary for recognition amounts at best to several scores of words/2/. The level of recognition errors is also unsatisfactory. We believe that there are two reasons for low quality of such systems. The first is that the energy spectra of speech signals are described using those varieties of Walsh transforms which are not invariant under the cyclic shift. The second reason is that in these systems the problems of speech processing at higher levels, especially training and decision making procedures, have not been given proper consideration. The purpose of the present paper is to fill in these gaps and to demonstrate the possibilities of the Walsh function method in designing of high quality speech recognition systems.

SPECTRAL DESCRIPTION OF SPEECH SIGNALS

To obtain a spectral description of speech signals, the energy spectrum of the Walsh transform with Hadamard ordering is proposed/1/. Such a description possesses two important advantages over other methods of spectral description of speech signals. The first is that the energy spectrum of the Walsh transform with Hadamard ordering is invariant under the cyclic shift. The second is that the realization of the corresponding fast transform algorithm requires $2(N-1)$ operations of real addition (subtraction). Since the processing speed for computations involving the FWT algorithm is determined by $N\log_2 N$, one can consider computations using the proposed algorithm to be $1/2\log_2 N$ as fast as those involving the FWT algorithms. Thus, for example, when the number of data points N is 256, the proposed algorithm gives a 4-fold increase in the processing speed.

As was mentioned earlier, the energy spectrum of the Walsh transform with Hadamard ordering is invariant under the cyclic shift. This means that for any periodic sequence $X(m)$, its energy spectrum of the Walsh transform with Hadamard ordering will coincide with that for a sequence $X(m+n)$, m and n being integers. As for other varieties of the Walsh transform, such as the Walsh transform with Walsh ordering, only the invariance of energy spectrum under dyadic shifts is valid/1/. This means that the corresponding energy spectra of sequences $X(m)$ and $X(m \oplus n)$ will coincide. The $\oplus$ operation means modulo 2 addition. The $E_h(r)$ components of the Walsh transform energy spectrum with Hadamard ordering are calculated according to the formula/1/:

$$E_h(0) = a_h^2(0), \quad E_h(r) = \sum_{k=2^{r-1}}^{2^r-1} a_h^2(k) \qquad (1)$$

$$r = 1,2,\ldots,n; \quad n = \log_2 N$$

Here $a_h(k)$ are the coefficients of the Walsh transform with Hadamard ordering, N is the number of input sequence points. It can be seen from eq.(1) that the number of components in the energy spectrum is $(n+1)$. Each component $E_h(r)$ of the energy spectrum represents the energy content of a group of sequencies rather than a single sequency, as is the case for the energy spectrum of the Walsh transform with Walsh ordering. The set of sequencies contained in each component of the energy spectrum of the Walsh transform with Hadamard ordering is calculated according to the formula/1/:

$$\begin{cases} F[E_h(0)] = 0 \\ F[E_h(1)] = N/2 \\ \vdots \\ F[E_h(k)] = 2^k, 3\cdot 2^k, 5\cdot 2^k, \ldots, N/2 - 2^k \\ \vdots \\ F[E_h(1)] = 2, 6, 10, \ldots N/2 - 2 \\ F[E_h(n)] = 1, 3, 5, \ldots N/2 - 1 \end{cases} \qquad (2)$$

To represent speech signals by means of the energy spectrum of the Walsh transform with Hadamard ordering, N =128 data points have been used for every time slot of 10 milliseconds. Accordingly, the number of components of the energy spectrum with Hadamard ordering amounted to 7 (n =7). The quantity $E_h(0)$ represents the energy of the direct component of the input sequence, and is not considered here.

To calculate the energy spectrum, a 16-bit computing device has been used whose computing speed characteristics were as follows. Execution time of the basic register arithmetic operations is 0.36 μsec., main memory read/write time is about 2 μsec. The device has 16 general registers. To calculate the energy spectrum of a speech segment 10 ms long by means of the above device, required about 2.5 ms (the quadratic coefficients are calculated using the table of squares contained in the main memory). After the energy spectrum is determined, the speech signal is represented by the quantities $\log_2(E_h(r)/E_0)$, r =1, 2, ... n, and $\log_2 E_0$, where $E_0 = \sum_{r=1}^{n} E_h(r)$. Thus, the input speech signal is represented by a 8-dimensional vector for each 10-ms slot. The number of bits representing components of a given vector was equal to 8. Input speech signal points were also represented using 8 bits per point. To store the reference speech patterns, 4 bits per each component of the 8-component vector are used. Experiments on speech pattern recognition have been carried out using the Walsh energy spectrum with Hadamard ordering for spectral description of speech signals. The recognition rate for vocabularies containing 50, 100 and 250 words was 99%, 96% and 92%, respectively.

TRAINING AND RECOGNITION

To increase the recognition rate of the system, a special training and recognition procedure has been developed. It is based on the following assumptions. An important method for improving the recognition rate is to represent each of the speech patterns to be recognized by several reference patterns obtained from corresponding clusterization of training samples. However, the present clusterization methods usually require optimization of some clusterization quality functional which depends on relative cluster position only within one group of patterns, irrespective of clusters formed within other groups. Meanwhile, in the process of recognition one has to relate distances obtained for different groups. In fact, the recognition rate is determined by relative positions of clusters belonging to different groups. Furthermore, one has to consider the dynamics of each cluster informativeness which depends

on the type of input speech signal. This is especially true for speaker-independent speech recognition systems in which speech variations from speaker to speaker necessitates adoptation to a speaker voice at the recognition stage.
We now assume that the clusterization procedure is carried out for each recognition pattern (i.e. a word or a phonem), and for each of the clusters a reference pattern is formed by averaging the training samples wich have fallen into a given cluster. Suppose then that a whole set of reference patterns is represented by J groups, each group containing one reference pattern for each recognition pattern. Then the adoptation to a speaker's voice within a given system will mean an automatic selection of a group of reference patterns for recognition according to "a posteriori" informativeness coefficients which are dynamically calculated for each group of reference patterns. For each of J groups of reference patterns, the following quantities are proposed as the informativeness coefficients:

$$\lambda(j) = \log_2 \frac{\rho_{i_1(j)}(j)}{\rho_{i_2(j)}(j)}, \quad j=\overline{1,J} \qquad (3)$$

where $\rho_{i_1(j)}(j)$ is the distance between the speech pattern being identified and the nearest reference pattern in the j-th group of reference patterns $(i_1(j)=\arg\min_{i=\overline{1,N}} \rho_i(j))$, N being the number of recognition patterns, $\rho_{i_2(j)}(j)$ is the distance to the next nearest reference pattern in the j-th group of reference patterns $(i_2(j)=\arg\min_{i\neq i_1(j)} \rho_i(j))$. The quantity of $\lambda(j)$ is the contrast range of the decision $i_1(j)$ obtained within the j-th group of reference patterns which directly determines the recognition rate of this decision. The decision i^* concerning the identity of the speech pattern being recognized is made on the basis of a set of decisions for all groups of reference patterns according to the formula

$$\begin{cases} j^{(1)} = \arg\max_{j=\overline{1,J}} \lambda(j) \\ i^* = i_1(j^{(1)}) \end{cases} \qquad (4)$$

A decision is not made in two cases. The first is for

$$\lambda(j^{(1)}) < \tilde{\lambda}_1 \qquad (5)$$

where $\tilde{\lambda}_1$ is a positive quantity. The second is for

$$\begin{cases} i_1(j^{(1)}) \neq i_1(j^{(2)}) \\ \lambda(j^{(1)}) - \lambda(j^{(2)}) < \tilde{\lambda}_2 \\ j^{(2)} = \arg\max_{j=\overline{1,J}, j\neq j^{(1)}} \lambda(j), \; i_1(j^{(2)}) = \arg\min_{i=\overline{1,N}} \rho_i(j^{(2)}) \end{cases} \qquad (6)$$

where $\tilde{\lambda}_2$ is a positive quantity.
The above method can be extended to include the training procedure. Suppose that within each of J groups of reference patterns, initial approximations to reference pattens $\tilde{e}_i^{(0)}(j)$ $(i=\overline{1,N}, j=\overline{1,J})$ are formed for each of N recognition patterns as a result of some self-training procedure. For the v-th training sample of the K-th recognition pattern, the following quantities are calculated:

$$\mathcal{L}_K^{(v)}(j) = \log_2 \frac{\rho_K^{(v)}(j)}{\rho_{K_1(j)}(j)}, \quad j=\overline{1,J} \qquad (7)$$

where $\rho_K^{(v)}(j)$ is the distance between the v-th training sample of the K-th recognition pattern and its reference pattern $\tilde{e}_K^{(0)}(j)$ in the j-th group of reference patterns, $\rho_{K_1(j)}^{(v)}(j)$ is the distance between the v-th training sample of the K-th recognition pattern and its nearest reference pattern in the j-th group of reference patterns $(K_1(j)=\arg\min_{i\neq K} \rho_i^{(v)}(j))$, besides the reference pattern of the K-th recognition pattern $e_K^{(0)}(j)$. This sample is involved in the formation of a reference pattern of the K-th recognition pattern only in the $j_K^{(v)}$ group such that

$$j_K^{(v)} = \arg\max_{j=\overline{1,J}} \mathcal{L}_K^{(v)}(j) \qquad (8)$$

Since the quantity of $\mathcal{L}_K^{(v)}(j_K^{(v)})$ represents the contrast range of the v-th training sample representation within some favourable group of reference patterns, the quality of training of the K-th pattern formed by V training samples can be evaluated from the training quality functional

$$\Lambda_K^{(1)} = \sum_{v=1}^{V} \mathcal{L}_K^{(v)}(j_K^{(v)}) \qquad (9)$$

To improve the reference patterns of the K-th recognition pattern, a multiple iteration procedure can be initiated. The r-th iteration will give the reference patterns of the K-th recognition pattern $\tilde{e}_K^{(r)}(j)$ $(j=\overline{1,J})$ by averaging the training samples of the K-th recognition pattern within the corresponding groups of reference patterns, in which they are placed according to eq.(8). The iteration procedure stops when

$$\Lambda_K^{(r)} \leq \Lambda_K^{(r-1)} \qquad (10)$$

where $\Lambda_K^{(r)}$ and $\Lambda_K^{(r-1)}$ are the training quality functionals for the r-th and $(r-1)$-th iteration, respectively. The procedure also stops when the number of iterations reaches its limiting value R_{max}:

$$r = R_{max} \qquad (11)$$

It should be noted that the training quality functional $\Lambda_K^{(r)}$ used in formation of reference speech patterns enables one to correct the vocabulary taking into account the training results, and replace "difficult" words in a vocabulary prior to recognition. The experimental check-up of the above method verified its potential usefulness in increasing the recognition rate. The experiments were carried out using two groups of reference patterns ($J=2$). Significant increase in the recognition rate has been achieved, which amounted to 99.5%, 99% and 98% for vocabularies containing 50, 100 and 250 words, respectively.

CORRECTION OF REFERENCE PATTERNS

As it was mentioned earlier, the present methods of reference pattern formation are not fully adequate since they do not provide optimum separation of speech pattern groups. We propose a new approach to reference pattern formation which cosists in automatic correction of reference patterns formed at the training stage in the process of recognition. Suppose that the training procedure resulted in the formation of reference patterns for N recognition patterns. Let us represent the recognition patterns together with their reference patterns by points in R-dimensional space. To calculate the distance between the recognition patterns and the reference patterns we employ the Chebyshev's metric:

$$\rho_i = \sum_{r=1}^{R} |X_r - \tilde{X}_{ir}| \qquad (12)$$

Here ρ_i is the distance between the speech pattern being recognized $\{X_r\}_{r=\overline{1,R}}$ and the reference pattern $\{\tilde{x}_{ir}\}_{r=\overline{1,R}}$ of the i-th recognition pattern. Let ρ_{i_1} be the distance between the speech pattern being recognized and its nearest reference pattern. Consider the following quantities:

$$\Delta_i = \rho_i - \rho_{i_1}, \; i \neq i_1, \; i=\overline{1,N} \qquad (13)$$

A decision which identifies an input speech pattern as belonging to a certain reference pattern will be the most reliable when the speech pattern is closest to the reference pattern of its group and farthest from reference patterns of those groups, to which it does not belong. Therefore, an optimum placement of reference speech patterns will be such that Δ_i are at their maxima. We propose an heuristic solution of this problem. The procedure consists in a successive correction using the results of reference speech pattern $\{\tilde{x}_{ir}\}_{r=\overline{1,R}}$ recognition. For each recognized speech pattern we select from Δ_i values those values Δ_K which satisfy the relation

$$\Delta^{(1)} < \Delta_K < \Delta^{(2)}, \; \Delta^{(1)} > 0, \; \Delta^{(2)} > 0 \qquad (14)$$

The correction of reference speech patterns will not be made if among Δ_i values there are no values which satisfy condition (14). Otherwise, the correction is applied both to the reference pattern of the i_1 recognition pattern, and the reference patterns of any K recognition patterns for which relation (14) is satisfied. Let $\{\tilde{x}_{ir}(0)\}_{r=\overline{1,R}}$ be the reference pattern of the i-th recognition pattern obtained as a direct result of the training procedure, and let $\{\tilde{x}_{ir}(v)\}_{r=\overline{1,R}}$ be the reference pattern of the i-th recognition pattern obtained as a result of the v-th correction of the initial reference pattern $\{\tilde{x}_{ir}(0)\}_{r=\overline{1,R}}$. Then, if a decision to correct reference patterns is made in the process of the next speech pattern $\{x_r\}_{r=\overline{1,R}}$ recognition, the correction will be carried out as follows. For a reference speech pattern recognized as i_1, the co-ordinates of its new corrected reference pattern $\{\tilde{x}_{i_1r}(v+1)\}_{r=\overline{1,R}}$ are obtained from the formula

$$\tilde{x}_{i_1r}(v+1) = \begin{cases} \tilde{x}_{i_1r}(v) + \Delta\tilde{x}_{i_1r}(v), & \text{if } x_r \geq \tilde{x}_{i_1r}(v) \\ \tilde{x}_{i_1r}(v) - \Delta\tilde{x}_{i_1r}(v), & \text{if } x_r < \tilde{x}_{i_1r}(v) \end{cases} \quad r=\overline{1,R} \qquad (15)$$

The $\Delta\tilde{x}_{i_1r}(v)$ values are calculated from the formula

$$\Delta\tilde{x}_{i_1r}(v) = \beta |x_r - \tilde{x}_{i_1r}(v)|, \; 0<\beta<1, \; r=\overline{1,R} \qquad (16)$$

The correction of co-ordinates for the reference patterns of those K recognition patterns for which relation (14) is satisfied, is made using the formula

$$\tilde{x}_{Kr}(v+1) = \begin{cases} \tilde{x}_{Kr}(v) - \Delta\tilde{x}_{Kr}(v), & \text{if } x_r \geq \tilde{x}_{Kr}(v) \\ \tilde{x}_{Kr}(v) + \Delta\tilde{x}_{Kr}(v), & \text{if } x_r < \tilde{x}_{Kr}(v) \end{cases} \quad r=\overline{1,R} \qquad (17)$$

The $\Delta\tilde{x}_{Kr}(v)$ values are calculated from the formula

$$\Delta\tilde{x}_{Kr}(v) = \gamma |x_r - \tilde{x}_{Kr}(v)|, \; \gamma \geq 0, \; r=\overline{1,R} \qquad (18)$$

The principal advantage of the above method is that training (reference pattern correction) and recognition procedures are performed simultaneously, and it does not require the use of a large number of training samples, which would be the case if these procedures were separated.

SUMMARY

It is shown that a speech recognition system with a large vocabulary and high recognition rate can be developed on the basis of Walsh functions. For vocabularies containing 50, 100 and 250 words, recognition rates up to 99.5%, 99% and 98%, respectively, have been obtained.
The proposed method of reference pattern correction in the process of recognition is expected to further increase the speech pattern recognition rate.

REFERENCES

/1/ N.Ahmed, K.R.Rao, "Orthogonal Transforms for Digital Signal Processing", Berlin, Heidelberg, New York, 1975.

/2/ P.A.Lee, I.Seymour, Evaluation of the Fast Walsh-Hadamard Transform for Speech Recognition by an 8-bit Microprocessor, Acoustica, vol.59, pp.274-278, 1986.

DEVELOPMENT OF METHOD AND DEVICE FOR IMPROVED REAL-TIME SPEECH RECOGNITION RELIABILITY

K.P. MAISTRENKO

V.M. Glushkov Institute of Cybernetics
Academy of Sciences of the Ukrainian SSR
Kiev 207, USSR 252207

ABSTRACT

The method and the device for invariant voice images recognition are suggested aimed at improving the real-time voice images recognition reliability when an arbitrary number of operators are involved.

The experience of numerous investigations dealing with the problem of designing the voice signal recognition systems testifies to the great difficulties involved in providing high reliability for these systems even in the case of one dictating operator. This is explained by great variability of the principal voice signals parameters which considerably grows when a wider circle of operators is involved. The evaluation of variational boundaries of the voice signal parameters variation shows that the interval of its intensity variation makes approximately 60 db, the frequency spectrum ranges from 20 Hz to 20-22 kHz, the continuity of pronouncing the voice images by different operators can be twice as quick or slow varying from operator to operator.
Of importance is also the fact of the great variability of the Russian language phonemes, which is related to the position of a phoneme in a word; the qualitative composition and the form of neighboring phonemes surrounding this one; the fact whether the phoneme is stressed or not; the rhythmic-dynamic structure of word combinations and word-forms, etc.
However, the human aural system possesses the unique capabilities with respect to fruitful voice signals recognition irrespective of the voice, rate and sound intensity of pronunciation. These properties, inherent in the aural system, have recently attracted the great attention of investigators and designers of speech recognition systems.
Working with the voice signals whose boundary parameters and possible variation of intensity, the frequency spectrum, the duration and rate of sounds pronunciation were considered above, one gets convinced that the following procedures are expedient:

efficient signal reception in the environment and formation of its acoustic analog (or establishment of interface between a recognition system and a signal source);
- normalization of voice signal with regard to intensity (amplitude) taking care that the optimal level is constant
- normalization of voice signal with regard to frequency characteristics at the expense of maximum restriction of possible variation of timbre, prosodic and emotional coloration (and/or variation of the principal tone frequency);
- normalization of voice signal with regard to duration of vocal speech units pronunciation adopted to the rate of information inflow.

The solution of these problems affords formation of the invariant description of the voice signal which would be the least influenced by the negative effects of speech variability and obvious redundancy interfering with the recognition system reliability.
Proceeding from this concept, we developed the method of voice images recognition using invariant voice signals processing [1,2] . The essence of this method is better understood on considering the troublespots of the known methods of voice signal recognition.
Thus, there are voice signal recognition methods where isolation of the voice signals' attributes is realized through the use of coding and of articulation attributes [3] .
However, standardization with regard to frequency of the principal tone results in the accuracy reduction when isolating the voice signals attributes.
The well-known method and the device for its implementation imply conversion of the voice images into an electric signal, amplification, phonemes separation, dynamic spectral analysis, quantization, separation of phoneme's attributes, their normalization and comparison with the reference signal [4] .
The reference signal is formed as a sum of power functions possessing the fixed transfer characteristics.

The weak point of this method consists in low accuracy and rate of voice images recognition because the accurate power functions can be obtained only with the help of the ideal multiplier.
The objective of the suggested method resides in improving the accuracy and rate of voice images recognition. To this end and according to the given method, the amplified electric signal is standardized with regard to continuity, after quantization it is normalized in frequency and amplitude and according to the obtained signals the short pulses are formed which are integrated and normalized with regard to continuity and compared with the invariant reference signals of the voice images obtained in the process of teaching the recognition system. The reference functions are formed by choosing the scale factors with respect to the signal of mismatch between the function being compared and the reference one.
The preliminary standardization of the voice image signals with respect to continuity and quantization ensures the consequent and synchronized operation of the entire analysis route. Formation of invariants with respect to frequency by way of simultaneous speech analysis and synthesis eliminates the effects produced by scattering of speech sounds tonality. The normal well-articulated voice of the synthesized sounds will always be heard at the frequency invariantor's output no matter how high the operator's voice tonality. Standardization of voice image signals with respect to amplitude, when uniformly weak or strong voice signals are amplified, expands the dynamic input signal range, preserves the highest formants in the spectrum which are usually lost when the speech is clipped.
Preliminary voice signal processing and normalization in continuity by way of invariants formation with regard to continuity increases the rate of voice image recognition providing high recognition accuracy. What is more, the comparison of voice images converted in the above-mentioned way with the reference signals obtained by functional conversion $X \rightarrow f_i(X)$

where X and $f_i(X)$ –

are independent functions, and by preliminary record of scale factors (during teaching) whose sampling is performed at high speed in the process of recognition, considerably increases the accuracy and rate of voice image recognition.
The investigations carried out with the use of recognition system models and experimental breadboards made it possible to suggest the variants of hardware implementation of invariantors of voice image amplitude, frequency and continuity which sufficiently reduce the redundancy of voice signals, enhance the invariance of recognition systems intended for the real-time operation with an arbitrary circle of operators [5,6] .
The suggested way of invariant voice signal processing can also be applied to processing the sound and acoustic signals for the purpose of their analysis, synthesis and recognition [7,8] .

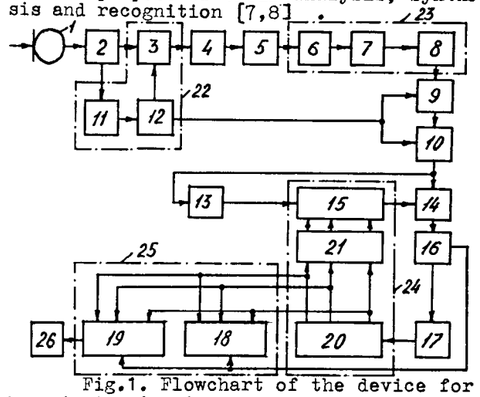


Fig.1. Flowchart of the device for invariant voice images recognition.

Fig.1 shows the device flowchart of the invariant voice image recognition implementing the suggested method. The device contains a microphone 1, an amplifier 2, an electron key 3, a frequency invariantor 4, an amplitude invariantor 5, an amplifier-limiter 6, Schmitt-trigger 7, a pulse shaper 8, an integrator 9, a sound continuity invariantor 10, a sound continuity generator 11, a sound continuity quantization unit 12, a saw-tooth voltage generator 13, a comparator 14, a reference functions generator 15, a zero-organ 16, a pulse generator 17, a printer 18, an electronic digital computer 19, a decoder 20, a code-analog converter 21, a synthesizer 26.
The elements 3, 11, 12 make up a sound continuity standardizer 22. The elements 6, 7, 8 make up a short pulse shaper 23. The elements 15, 20, 21 make up a reference signal unit 24, the elements 18 and 19 make up a register 25. The elements connection is realized as shown in Fig.1.
The device functions in the following way. The operator's voice is transformed by a microphone 1 into electric signals which are intensified by an amplifier 2 and arrive at an electron key 3 and a sound continuity generator 11 of sound continuity standardizer 22. Since the informative part of the elementary sound resides in its initial stage, then these devices ensure the normal passage of the initial sound energy over the interval of approximately 0,1 sec, and then the channel is switched off till the new sound energy pulse appears. The quantized pulses arrive at a frequency invariantor 4. This device carries out the dynamic spectral analysis

of the voice signal and converts the spectrum of the operator's voice signal in such a way that the voice of the synthesized sounds becomes independent of the tone's pitch of the operator's voice. The normal voice will always be heard at the frequency invariantor output, independently of the speaking operator. Artificial voice of frequency invariantor 4 arrives at an amplitude invariantor 5 which converts the voice signal in such a way that the signal at its output becomes no longer dependent on the amplitude but the main characteristics of sound information are completely retained. This is attained through functional transformations in the amplitude invariantor 5 which ensure: sampling of all weak signals, self-sustained or in combination with strong signals within the whole dynamic spectrum, their amplification to the normalized level with regard to amplitude, and comparison with each other followed by summation. As a result, different sound signals turn out to be equal in amplitude, and the output signal reminds of a clipped signal though it is of a higher quality.The amplitude invariantor's 5 output signal is clipped with the aid of an amplifier-limiter 6 and Schmitt-trigger 7, the dependence on amplitude of an output signal of invariantor unit 5 is completely eliminated, then the signal is differentiated and formed as a microsecond pulses package. Having been shaped in the pulse shaper 8, the pulses are integrated in the integrator 9, then the integrated pulses arrive at the continuity invariantor 10 intended for storing integration function and its compression in time for reproducing the integration function with higher frequency. The continuity invariantor 10 makes it possible to record integration function,0,1 sec in continuity at 100 descrete points, to an accuracy of 1% and to reproduce this function periodically repeating it at output, with frequency 200 kHz which ensures high frequency of comparison between the integration function within a comparator 14 and the reference functions generated by a reference functions generator 15 of a reference signal unit 24 and permits of recognitions within short time intervals. Single-argument function converter, preliminary trained to integrated functions of the elementary speech images can be used as the reference functions generator 15. Sampling of the reference functions available in the generator is realized in conformity with the output signal of the sawtooth voltage generator 13. The reference functions are fed in succession and at high speed from the reference functions generator 15 to the comparator 14. In case of disagreement between the reference function and voice image recognition function arriving from the sound continuity invariantor 10 at the comparator 14 output, there appears a signal , passing to zero-organ's 16 input, the zero-organ starts the pulse generator 17 which, in its turn, starts the decoder 20. The spectra of scale factors are correlated with respect to the decoder's 20 codes through the use of the code-analog converter 21. If the selected spectrum of scale factors ensures the similarity of the compared functions in the comparator 14 when the reference functions generator 15 is questioned, then the zero-organ 16 generates the switch off signal for the pulse generator 17 and the signal for fixing the code as an alphabetic record corresponding to the recognized sound in the printer unit 18. It is stored in computer memory and fed into the synthesizer 26. Then the device is cleared and gets ready to recognize the next sound image.

The reference functions shaping in the reference functions generator 15 is realized in the following way. The test voice images are distinctly pronounced before the microphone. Just in this mode the zero-organ 16 controls the system of scale factor adjustment, the latter consists of the pulse generator 17, the decoder 20, the code-analog converter 21. The determined spectra of scale factors are preliminary recorded before dictation of test voice images, and then introduced into the code-analog converter 21. If the scale factors vary smoothly during the reference function generator 15 tuning, so after the data is fed into the code-analog converter, the scale factors instantly assume those values at which the trained curve will be reproduced.The total number of scale factor spectrum variations equals the number of code combinations. For the described device the decoder is designed for ten bit binary code when the number of decoder's combinations amounts to 1024.

The described device is easily tuned, trained and implemented, its high accuracy of voice signals recognition ensures its utilization in the systems of man-machine interaction when robots of "ear-intelligence" type are designed and in other engineering domains.

REFERENCES

[1] K.P. Majstrenko, "The Method and Device for Invariant Voice Images Recognition, Transactions of All-Union School-Seminar "Psychological Bionics", Charkov, 1986, p.20.

[2] B.V. Bolotov, K.P. Majstrenko, G.G. Chub,"The Method of Voice Images Recognition", Certificate of Copyright of the USSR N 621003, BI N 31 of 25.08. 1978.

[3] B.N. Sorokin, Certificate of Copyright of the USSR N 432581, BI N 22 of 15.06. 1974.

[4] V.Ju. Trachtman, "The Device for Voice Signals Analysis", Certificate of Copyright of the USSR N 298943, BI N 11 of 31.03.1971.

[5] B.V. Bolotov, K.P. Majstrenko, "The Device for Frequency Voice Images Normalization", Certificate of Copyright of the USSR N 643959, BI N 3 of 23.08.1979.

[6] B.V. Bolotov, K.P. Majstrenko, G.G. Chub, "The Device for Voice Information Recognition", Certificate of Copyright of the USSR N 758238, BI N 31 of 23.08.1980.

[7] K.P. Majstrenko, A.A. Tyshko, "The Device for Sound Signals Processing", Certificate of Copyright of the USSR N 771709, BI N 38 of 15.10.1980.

[8] K.P. Majstrenko, A.A. Tyshko, "The Device for Acoustic Information Processing", Certificate of Copyright of the USSR N 822248, BI N 14 of 15.01.1981.

AN UNUSUAL EFFECT ON THE PERCEPTION OF STRESS

VINCENT J. VAN HEUVEN

Dept. of Linguistics/Phonetics Laboratory,
Leyden University, P.O. Box 9515,
2300 RA Leiden, The Netherlands

ABSTRACT

In our research we try to understand why (Dutch) listeners tend to perceive the first syllable of a phonated reiterant word as stressed when presented in isolation, but not when the word was whispered or generated in a preceding (but not following) spoken context. We suggested that the listener interpretes the periodicity onset of an isolated word as a(n accent-lending) pitch rise relative to the bottom of the speaker's pitch range. Some consequences of this view are further tested in the present study.

INTRODUCTION

Listeners tend to perceive lexical stress on the first syllable in isolated words. This stress bias is most conveniently demonstrated with so-called reiterant speech, i.e. words made up of repetitions of an identical syllable. The proportion of initial stress perceived in such stimuli was found to range between 59% (binary free choice in English [1]) and 80% (4-alternative forced choice in English [2]), with intermediate values reported for Dutch (65%, ternary free choice [3]), and Polish (79% binary free choice [4]).

The basic question that we set out to answer is: What causes this bias towards perceiving the initial syllable as stressed? Although it would seem obvious to relate this bias to the statistical distribution of lexical stress in the language, which in the above experiments favours onset position, we have reason to believe that the bias is at least partly caused by a perceptual mechanism of a more general nature: Van Heuven & Menert [5] established that:

(i) presenting a target word in a preceding speech context reduces initial position bias;
(ii) replacing the periodic (buzz) source signal in isolated targets by a noise source ("whisper") reduces bias likewise;
(iii) increasing the fundamental frequency (F0) of the periodic source in isolated words from a level 100 Hz to a level 160 Hz increases bias.

These findings were explained as follows. We assume that the listener, on hearing the utterance-initial syllable generates a reference pitch level that is equal to the lowest vocal pitch appropriate for the particular speaker's voice, i.e. his terminal frequency for a declarative sentence. For an average Dutch male this pitch would lie around 75 Hz. The actual pitch of the stimulus onset, which in our experiments was at 100 Hz, is then evaluated against this (lower) reference pitch. The difference between the actual pitch (100 Hz) and the reference pitch (75 Hz) is interpreted as a pitch jump (or: "virtal rise"), and taken as a cue for stress on the first syllable.

When the stimulus is a-periodic, no actual pitch can be determined, and no pitch jump can be inferred. Hence bias disappears in whispered targets. When the target is preceded by a spoken carrier phrase, the reference pitch is provided by the context. Since, in our experiment, the pitch was level throughout the entire stimulus, no pitch jump was heard, and bias disappeared.

How can we show that this - admittedly speculative - account of stress bias is correct? If it is true that the difference between onset pitch and reference pitch is interpreted as a pitch movement, it follows that a higher actual pitch onset, all else being equal, should yield a greater bias. Therefore we shall vary the onset pitch between 70 Hz (coincident with the assumed reference pitch) and 160 Hz. We hypothesize that a 70 Hz onset will generate little or no bias, but that higher onsets (100, 130, 160 Hz) will come out with ever larger bias. In our earlier experiment F0 varied in just two steps (100 versus 160 Hz). Unfortunately the 160 Hz condition comprised only a small proportion of the stimulus material. It was therefore an unusual condition within the experimental context, which by itself may have contributed to the increase in bias. In the present experiment each of the four pitch levels occurs equally often, so that the effect of pitch level can be examined more conclusively.

Secondly, we shall speculate on the mechanism by which the listener generates the reference pitch. Here we hypothesize that the listener assumes the presence of large individual from a speech sample with relatively low formants (i.e. large resonance cavities), whom he associates with a low bottom pitch. Conversely, when the formants in the speech sample are relatively high, the speaker is apparently a small individual, with a correspondingly high-pitched voice. To test this hypothesis we generated stimuli using three different formant settings: starting from a normal male with an average formant setting, we simulated a large individual with lowered formants, as well as a small individual with raised formants. If it is true that lowered formants are associated with a low-pitched voice, the listener will generate a lower reference pitch for this type of voice, so that a larger virtual rise will be perceived relative to a fixed actual pitch onset. Similarly, if raised formants correspond to high-pitched voices, we predict stronger bias for stimuli with up-shifted formants.

METHOD

Seventy-two stimuli were generated by a DEC Micro-VAX-II computer using the LVS speech analysis and (re-) synthesis software developed at IPO-Eindhoven. The stimuli comprised 36 versions of the Dutch word 'kanon' and another 36 of the nonsense word 'saasaas'. Both words are ambiguous for (lexical) stress position: 'kanon' with stress on the first syllable means 'polyphonic hymn' but 'gun' with stress on the second syllable. The Dutch stress rules are compatible with stress on either syllable in the reiterant sequence 'saasaas' (see further [5]). The words were synthesised from diphones which had been excerpted from the accented syllable in nonsense words of the type [CiəCi'VCiə], and stored in parametrised form in computer memory using the AAP-LPC analysis program provided by the LVS-package, using 5 formants and 5 bandwidths in the frequency band up to 4.5 kHz, calculated over a 25-ms time-window that was shifted along the time axis in 10-ms steps. Given their origin, all the sound seg-ments in our diphone synthesis are equally suggestive of strong, primary stress. This is what makes this type of synthesis ideal for our purpose: in the absence of any experimenter-induced parameter changes the distribution of perceived stress should be neatly balanced over the two syllables in our words.

The 36 versions of each word were then obtained through orthogonal combination of three factors:

(i) F0 was varied in 4 steps: 70, 100, 130, and 160 Hz; F0 was level throughout the duration of the stimulus word.
(ii) Formant range was varied in 3 steps. Starting from the formant frequencies F1 through F5 as calculated by the LPC-analysis, a type of voice was synthesised that was typical of a large male (formants lowered to .85 of their original frequencies), and another type that suggested a small male (formants raised to 1.20 of their original values).
(iii) Temporal type was varied in three steps. Next to the original version consisting of temporally unadjusted diphones, a version with a noticeably longer first syllable was obtained by lengthening the steady state portion of the first vowel by 50 ms while shortening the second vowel by the same amount. A complementary version with a longer second syllable was generated by reversing this procedure.

The 72 tokens were recorded onto audio tape in quasi-random order, preceded by 8 practice items. This tape was presented twice to 11 Dutch listeners over a good quality sound reproduction system (Quad ESL-63) in a small, well insulated lecture soom with some soft paneling attached to walls and ceiling so as to limit reverberation. Listeners were instructed to decide for each stimulus whether the stress was on the first or on the second syllable, with binary forced choice. They were to indicate their choice by ticking the appropriate syllable on answer sheets that contained a listing of the 72 stimuli in the order in which they appeared on the tape, typed in ordinary Dutch spelling.

RESULTS

Examining table I, which presents % stress perceived on the first syllable broken down by pitch level, formant setting, and temporal version, we observe the following:

TABLE I: Percent stress perceived on first syllable broken down by F0-level, temporal version, and formant setting.

	F0-level (Hz)			
	70	100	130	160
1st syll long				
formants lowered	93	93	98	98
formants neutral	95	93	93	100
formants raised	98	91	100	98
1st/2nd syll equal				
formants lowered	61	80	86	86
formants neutral	80	84	89	93
formants raised	86	89	86	93
2nd syll longer				
formants lowered	7	9	9	28
formants neutral	14	9	7	28
formants raised	3	23	11	11

1. Manipulating the relative duration of first and second syllable produces 96% stress perceived on the lengthened first sylable, 85% on a temporally neutral first syllable and 13% on a shortened first syllable. This effect was significant by a classical three-way ANOVA with pitch level, formant setting and temporal version as fixed factors, $F(2,141)=1120.1$ ($p < .001$). This effect is well-known from the literature (cf. [6] and references given there) and can therefore serve as a baseline against which the strength of the remaining factors can be evaluated.
2. Changing F0 level has a clear effect on stress perception. Bias for the first syllable increases monotonically with F0 level: 60% stress for 70 Hz, 63% for 100 Hz, 64% for 130 Hz, and 71% for 160 Hz. Although this effect is smaller than that of temporal version, it is still substantial, $F(3,140)=8.5$ ($p < .001$). The effect of F0 is most pronounced in the temporally neutral versions with 76, 84, 87, and 91% stress perceived on the first syllable, $F(3,45)=4.4$ ($p=.020$).

Counter to our prediction, bias does not disappear completely at 70 Hz. Whether a further reduction of bias can be obtained by lowering the F0 level still further remains doubtful: when constructing our stimuli we had to abandon pitches below 70 Hz, as these sounded highly unnatural (rough, creaky voice quality).

3. The predicted effect of formant setting is not borne out by our data. If anything, the results are in the wrong direction, but the effect of formant setting is insignificant, $F(2,141)=1.9$ (ins).

CONCLUSION

Both in this and in our previous experiments we have demonstrated that the perception of stress is not solely dependent on differences between F0, intensity, duration and timbre within the word or utterance, as is generally maintained in the literature (cf. [6]).

We have presented convincing evidence here that the onset frequency of an otherwise perfectly level pitch influences the perception of lexical stress in isolated words: the higher the pitch level, the greater the bias favouring perceived stress on the first syllable. Generally, the results confirm our claim that stress bias is caused by the listener's perceiving the discrepancy between the actual pitch onset and some low reference pitch as a virtual pitch rise cueing stress, i.e. an auditory meachanism, rather than by the listener's knowledge of the statistical distribution of lexical stress in the language.

However, we have not been able to confirm our suspicion that the reference pitch is derived from the average formant setting in the voice of the speaker, which negative finding prompts at least two questions for further research. Firstly, is it really true that listeners associate a particular pitch range with a given formant setting, and secondly, could it be the case that the reference pitch is fixed and speaker independent? These questions will be taken up in our future research.

References:

[1] J. Morton, W. Jassem, "Acoustic correlates of stress", Language and Speech, 8, 1965, 148-158.

[2] A.E. Berinstein, "A cross-linguistic study on the perception and production of stress", UCLA Working Papers in Phonetics, 47, 1979, 1-59.

[3] A.F. van Katwijk, Accentuation in Dutch, an experimental linguistic study, Van Gorcum, 1974.

[4] W. Jassem, J. Morton, M. Steffen-batog, "The perception of stress in synthetic speech-like stimuli by Polish listeners", Speech Analysis and Synthesis, 1, 1968, 289-308.

[5] V.J. van Heuven, L. Menert, "Linguistic and perceptual causes of stress perception bias", ms. submitted to Journal of the Acoustical Society of America, 1968.

[6] I. Lehiste, Suprasegmentals, MIT-Press, 1970.

WORD STRESS OF TRISYLLABICS OF OLD FRENCH ORIGIN IN LATE MIDDLE ENGLISH

TOMONORI MATSUSHITA

Dept. of English Literature
Faculty of Letters
Senshu University
Kawasaki, Kanagawa, JAPAN

ABSTRACT

The unmarked stress pattern obviously differs in the degree of grammaticalness from the marked patterns in line-initial position, caesura, and rhyme and alliteration. Separation of linguistic intuition from poetic license enables us to define the former more narrowly and limit the scope of poetic license. The Main Stress Rule captures unmarked stress patterns of trisyllabic words of Old French origin, while stylistic rules account for marked stress patterns which are only derivatives from the unmarked ones. The stress patterns of Old French words were modified when they were borrowed into Middle English.

INTRODUCTION

Halle and Keyser [1] show that both the Initial Stress Rule inherited from Old English and the Romance Stress Rule were productive in the language of Chaucer and his contemporaries. The former rule correctly determined the place of stress for words like hóli, félawe, while the latter rule assigns stress to the three types of words Týdeus, Satúrnes, and honóur, respectively. They notice that stress doublets like comfórt - cómfort must be described either in terms of a shift into another lexical category or in terms of assumption of two possible pronunciations, one with a lax vowel in the last syllable, the other with a tense vowel.

Nakao [2] argues that the application of the Romance Stress Rule and the Stress Retraction Rule accounts for stress doublets like bargéyn - bárgeyn. Under Nakao's analysis, the stress placement in doublets seems to be highly motivated and the Initial Stress Rule can be eliminated. Closer examination reveals that the Romance Stress Rule, which plays a crucial role in the analyses mentioned above, faces serious difficulties in assigning stress to Old French loanwords [3: 111-12].

Under the two analyses presented by Halle and Keyser and by Nakao, the two stress patterns in doublets would have the same degree of grammaticalness and, therefore, speakers of late Middle English would have possessed a broader and looser linguistic intuition than speakers of any other age.

However, I [3] argue that the unmarked stress pattern obviously differs in the degree of grammaticalness from the marked patterns in line-initial position, caesura, and rhyme and alliteration and that Chaucer and his contemporaries composed their verses utilizing poetic license which was partly deviant from their own linguistic intuition. This approach predicts that difference in judgment of grammaticality of the same stress patterns reflects the difference in derivations. Separation of linguistic intuition from poetic license will enable us to define the former more narrowly and limit the scope of poetic license.

It is argued in Matsushita [3] that the Main Stress Rule captures unmarked stress patterns of bisyllabic words of both native and Old French origin like bósom, bihéest, cítee, and benígne, while stylistic rules account for marked stress patterns like bosóm, bíheest, citée, and bénigne which are only derivatives from the unmarked ones. I assume that the stress patterns of Old French words were modified when they were borrowed into Middle English.

The unmarked stress patterns in late Middle English are subcategorized into three parts. Consider the stress assignment in the following trisyllabic words. Nouns: bácheler (A.Pro1 80, MV bácheler), wýdeweris (PP1A 10.200), sýmonie (PP1B 2.63, MV sýmonye); Adjectives: síngulēer (G.CY 997, MV síngulare), délicāat (E.C1 927, MV délicate), partículēr (E.C1), consérvatīf (HF 847); abhóminable (B.NP 4243); Verbs: óccupīe (F.Sq 64), sácrifīce (TC 5.423, MV sácrifice), mýnistren (PP1B 12.52). To account for the observed stress distribution, I propose to assign primary stress to the antepenultimate vowel except that an unstressed lax vowel is optionally suffixed to the word if the penultimate vowel is nontense and is followed by no more than a single consonant. The last vowel is either lax or tense and may be followed by more than one consonant. Notice that inflectional suffixes are neutral to the Main Stress Rule. Words prefixed with ad-, in-, and dis- like appárayl (GGK 601), apáraunt (Cln 1007), engéndred (E.C1 158), and discóverest (G.CY 696) receive primary stress on the penult by the Main Stress Rule. There are many words that are morphologically analyzable into one of the prefixes aper-, compre-, etc., followed by a stem such as -ceyve, -hende, etc.: apércēyved (RR 6371), cómprehénde (BD 762). The stress placement rule must assign primary stress to the final stem in these words. Using the customary formalism for the statement of phonological rules, I state

the Main Stress Rule as follows [3: 115]:

$$V \longrightarrow [1\ \text{stress}] \;/\; [\underline{X}(=)\ \_\_\_\_ C_0((\begin{bmatrix}-\text{tense}\\ V\end{bmatrix} C_0^1) CV_0)(ę)]$$

where ę stands for a schwa.

Let us turn now to our discussion of the difference between the Romance Stress Rule proposed by Halle and Keyser and Nakao and the Main Stress Rule just presented above since they are formally similar. As I have stated, however, the latter uniquely generates, without requiring any exceptional treatment, the stress patterns which reflect the linguistic intuition of Chaucer and his contemporaries as native speakers, while the former neither applies to words of Old French origin in which the vowel in the last syllable is to be stressed nor discriminates linguistic stress from alliterative and metrical stress. My analysis argues that the Main Stress Rule of Modern English had essentially reached its current form by the late Middle English Period [4]. This view is consistent with the historical fact that the system of English word stress was more greatly influenced by Old French than by other languages.

LINGUISTIC STRESS VS. STYLISTIC STRESS

To clarify the distinction between linguistic stress and alliteration and metrical stress, it may be useful to outline assumptions underlying this paper. Linguistic stress is a concept that belongs to the study of competence, whereas alliteration and metrical stress belong to the study of performance. The former is generated by a context-free rule called the Main Stress Rule. Linguistic stress is perfectly grammatical, independent of any stylistic consideration and has mostly been preserved in Modern English. On the other hand, the latter types of stress are derived by stylistic rules to base-generated stress patterns. Alliteration and metrical stress may be semi-grammatical, dependent on style, alliterative or metrical, and not been preserved in Modern English. It is also important to notice that "Die legitime Accentverschiebung zu Gunsten des Verses hat namentlich im Reim und demnächst in der Cäsur ihre Stelle" and in the line-initial position and that alliterative and metrical verses differ in manner and frequently from legitimate stress shift [5]. I will illustrate three types of stress patterns and discuss stylistic rules relevant to marked patterns. Statistical consideration is made of each type of stress patterns. Examples of metrical verses are drawn from Chaucer's works [5'] (1369-99); alliterative verses, from Patience [6] (?c1380), Cleanness [7] (?c1380), St. Erkenwald [8] (c1386), Sir Gawain and the Green Knight [9] (?1390), and Piers the Plowman (A [10] a1376, B [11] c1378, C [12] ?a1387). The symbols Ⓘ, Ⓕ, and Ⓒ stand for line-initial position, rhyme, caesura, respectively.

Consider first words of Old French origin in which the Main Stress Rule assigns primary stress to the antepenultimate vowel.

Nouns
Ch. ábstinénce 21x (Ⓘ 11x); ábstinence (PP1A 5.220, 6.109, 8.119/PP1C 7.440)
Ch. áppetít(es) 16x (Ⓘ 9x); áppetyt (PP1A 7.251)
Ch. árgumént(s) 20x (Ⓕ 5x); árgumentz (PP1B 15.375/PP1C 20.110)
Ch. áudiénce 18x (Ⓒ 1x, Ⓕ 17x); áudience (PP1C 8.94)
Ch. champíoun 9x (Ⓕ 6x); chá(u)mpion (PP1A 9.41/ PP1C 16.279, 21.104)
Ch. cláryóun 11x (Ⓕ 9x); cláryoun (Cln 1210)
Ch. cóuntenáunce(s) 44x (Ⓕ 33x); cóuntenaunce (PP1B pr.24, 5.183, 11.15, 13.111/PP1C 1.26, 12.164, 16.120), (Cln 792), (GGK 100, 1490, 1539)
Ch. díadéme 5x (Ⓕ 4x)
Ch. élemént(z) 6x (Ⓕ 3x); élement(z) (PP1B 18.235/ PP1C 2.17, 21.247)
Ch. fántasýe 23x (Ⓒ 1x, Ⓕ 20x); fántasye(s) (PP1A pr.36, 11.63/PP1B pr.36/PP1C 1.37)
Ch. fúnerál 4x (Ⓘ 1x, Ⓕ 1x)
Adjectives
Ch. ámoróus 15x (Ⓕ 6x)
Ch. cóntrarie 11x (Ⓕ 4x)/ contrárie RR 2246 (Ⓘ 1x), RR 5312; cóntrarie (PP1C 10.193, 20.325), (Cln 4, 266, 1532)
Ch. dángeróus 15x (Ⓒ 1x, Ⓕ 11x)
Ch. désoláat 9x (Ⓒ 2x, Ⓕ 6x)
Ch. fórtunát 6x (Ⓒ 1x, Ⓕ 3x)
Ch. génerál 8x (Ⓕ 3x)
Ch. hónuráble 18x (Ⓕ 10x); cf. hónourablely (PP1B 12.155)
Ch. náturál 11x (Ⓒ 1x, Ⓕ 3x)
Ch. périlóus 18x (Ⓒ 1x, Ⓕ 5x); périlous (PP1A 7.44/PP1C 7.186n), (GGK 2097)
Ch. príncipál 7x (Ⓕ 5x); prýncipal(e) (Cln 1531, 1781), cf. príncipaliche (PP1B 14.194)
Ch. résonábele 11x (Ⓕ 10x); ré(i)sonable (PP1B pr.158n, 13.286/PP1C 1.176, 4.369, 7.33), (Cln 724)
Ch. sóveráyn 30x (Ⓕ 8x); sóuereyn (PP1B pr.159, 10.210, 11.370, 14.114/PP1C 2.148, 7.27, 16.295, 23.372), (Cln 93, 178, 210, 552, 557, 780, 1152, 1225,1313, 1454, 1643, 1670), (Erk 120), (GGK 1278), (Pat 429)
Verbs
Ch. círcumscríve TC 5.1865 (Ⓕ 1x), cf. MV círcumcise [13]
Ch. éxercíse 3x (Ⓕ 3x)
Ch. éxecúteth A.Kn 1664
Ch. frúctyfýe Scog.48 (Ⓕ 1x)
Ch. glóriffe 4x (Ⓕ 4x)
Ch. mágniffe HF 1.306
Ch. múltiplýe 14x (Ⓒ 2x, Ⓕ 10x); múlteplie (PP1C 19.226), (Cln 278, 522)
Ch. óccupíe(th) 6x (Ⓕ 1x); óccupien (PP1B 16.196/ PP1C 8.18, 19.207)
Ch. sácrifíce 3x (Ⓕ 2x); sácrifised (PP1B 12.118), sákerefyse (Cln 507, 510, 1447, 1497), (Pat 239, 334), (Prl 1064)
Ch. stéllyfýe 3x (Ⓕ 3x)
Ch. vérifíe G.CY 1068
Ch. vérsiffe B.Mk 3168; vérsifie (PP1B 15.367/ PP1C 18.109)

In Chaucer's metrical verses (see Table 1 below), unmarked instances total 396 --- of which two

instances occur in the line-initial position, 9 in the caesura, and 240 in rhyme. 145 unmarked instances occur elsewhere. In the alliterative verses, unmarked stress pattern occurs 85 times while no instances of the marked pattern are identified. In the stylistic component of the LME grammar the Metrical Rule I (MR-I) applies to words of Old French origin like ábstinence, áppetit, and dángerous to derive metrical forms like ábstinénce, áppetít, and dángeróus.

$$V \longrightarrow [1\ \text{stress}] \;/\; C_0\begin{bmatrix} V\\ [1\ \text{stress}]\end{bmatrix} C_0 VC_0 \_\_\_\_ \quad \text{(MR-I)}$$

Only two marked instances remain unanalysed under this approach.

Let us turn next to words of Old French origin in which the Main Stress Rule assigns primary stress to the penult.
Nouns
Ch. álliáunce 10x (Ⓕ 9x), MV alliáunce
Ch. ápparénce 10x (Ⓒ 1x, Ⓕ 7x); cf. apáraunt (Cln 1007), cf. MV appárant
Ch. acquéyntaunce(s) 5x/ ácqueyntáunce(s) 15x (Ⓕ 11x); aquóyntaunce (GGK 975)
Ch. avántage 4x/ ávauntáge 9x (Ⓕ 9x), MV aduántage
Ch. áventúre 80x (Ⓒ 5x, Ⓕ 55x); avénture (GGK 2482)/ áuenture (GGK 29, 250)
Ch. comáundement(z) E.Cl 649, G.CY 1063/cómaundemént(z) 13x (Ⓕ 9x); cómmaundemens (PP1C 12.143), (GGK 1303, 1501)
Ch. covéitise TC 3.261, RR 205/ cóveitíse 24x (Ⓕ 15x); cóue(i)tise (PP1A pr.58, 2.33, 3.158, 5.107, 10.192/PP1B pr.61, 3.68, 9.155, 10.18, 13.391, 14.238/PP1C 1.59, 3.90, 7.39, 11.257, 13.241, 17.223, 20.254, 22.224), (Cln 181), (Erk 237), (GGK 2374, 2380, 2508)
Ch. hábundánce 8x (Ⓕ 7x), cf. abúndant
Ch. obéissance(s) 2x/ óbeissánce(s) 16x (Ⓕ 15x), MV obéysance
Ch. obsérvánce(s) 23x (Ⓕ 17x), cf. MV óbseruance
Ch. philósophre(s) C.Pard 620, G.CY 1394/philósóphre(s) 18x (Ⓕ 7x)
Ch. rémembránce 43x (Ⓕ 37x), MV remémbrance
Ch. répentánce 10x (Ⓕ 9x); répentaunce (PP1A 5.43/PP1B 5.232/PP1C 7.234), MV repéntance
Adjectives
Ch. ácceptáble D.Sum 1913 (Ⓕ 1x)
Ch. ápparáunt 2x (Ⓕ 2x); apáraunt (Cln 1007), MV appárant
Ch. córporéll RR 6757 (Ⓕ 1x)
Ch. enténtif 3x (Ⓘ 1x)/ éntentíf 5x (Ⓕ 5x)
Ch. errátik TC 5.1812
Ch. etérnel 2x/ éternél TC 4.1062 (Ⓕ 1x)
Ch. fántastík A.Kn 1376 (Ⓕ 1x)
Ch. habúndant E.Cl 59/ hábundánt B.NP 4115, MV abúndant
Ch. inférnal A.Kn 2684/ ínfernál 2x (Ⓕ 1x)
Ch. repéntaunt F.Sq 655 (Ⓘ 1x)/ répentáunt A.Prol 228 (Ⓕ 1x), cf. MV repéntance
Verbs
Ch. accómplice RR 2132/ áccomplíce 3x (Ⓕ 1x)
Ch. apárayle 2x/ áparáyle LGW 2473 (Ⓕ 1x); appárayle (PP1A 2.148, 2.190, 7.53/PP1B pr.23, 2.170, 5.523, 6.59/PP1C 8.161), cf. pápailede (PP1C 1.25, 3.224)
Ch. contínu 5x (Ⓕ 2x)/ cóntinú RR 5332
Ch. delývere 26x/ délyvére 8x (Ⓒ 4x, Ⓕ 1x); delýuer (Cln 1084), (GGK 851)/ délyuer (Cln 286, 500)
Ch. détermýne 4x (Ⓕ 4x); cf. termýne PF 530, detérmyned (R. the Redeles 2.97)
Ch. disfigure(d) 3x/ dísfigúre(d) 2x (Ⓕ 2x)
Ch. engéndre 13x/ éngendré 2x (Ⓕ 1x); engéndrede (PP1C 11.215), éngendered (Cln 272)
Ch. enhábite TC 4.443/ énhabíte RR 6355 (Ⓕ 1x); énhabiten (PP1C 10.188)
Ch. enlúmyned 4x (Ⓘ 2x)/ énlumýned RR 5344 (Ⓕ 1x)
Ch. envényme(d) 2x (Ⓘ 2x)/ énvenýme(d) 3x (Ⓕ 2x); énuenyme (PP1B 12.256)
Ch. réconfórte 3x (Ⓕ 3x); réconforted (PP1B 5.287)
Ch. remémbere(th/st/d) 27x/ rémembré RR 4110 (Ⓕ 1x); rémembred (Pat 326)

As shown in the Table 2, Chaucer's verses contain 110 unmarked instances versus 324 marked ones in the line-initial position, in the caesura, and in

Table 1

stress pattern	metrical verses: unmarked: initial (/ _ _ (/))	unmarked: caesura (/ _ _ (/))	unmarked: rhyme (/ _ _ (/))	unmarked: elsewhere (/ _ _ (/))	marked (_ / _)	alliterative verses: unmarked (/ _ _)	alliterative verses: marked (_ / _)	alliterative verses: marked (_ _ /)
nouns	1	2	118	54	0	30	0	0
adjectives	1	5	101	79	2	39	0	0
verbs	0	2	21	12	0	16	0	0
total	2	9	240	145	2	85	0	0
	396						0	

Table 2

stress pattern	metrical verses: unmarked (_ / _)	marked: initial	marked: caesura	marked: rhyme /	marked: elsewhere / _ _	alliterative verse: unmarked (_ / _)	alliterative verse: marked (/ _ _)	alliterative verse: marked (_ _ /)
nouns	17	0	6	207	66	2	30	0
adjectives	9	0	0	13	2	1	0	0
verbs	84	0	4	18	8	11	6	0
total	110	0	10	238	76	14	36	0
		324					36	

Table 3

stress pattern	metrical verses: unmarked: initial (x) _ /́	unmarked: caesura (x) _ /́	unmarked: rhyme (x) _ /́	unmarked: elsewhere (x) _ /́	marked: _ /́ _	alliterative verses: unmarked: _ _ /́	marked: /́ _ _	marked: _ /́ _
nouns	-	-	-	-	-	-	-	-
adjectives	-	-	-	-	-	-	-	-
verbs	1	0	20	21	2	1	3	2
total	1	0	20	21	2	1	3	2
	42						5	

rhyme (0, 10, and 238, respectively). The Stress Movement Rule (SMR) applies prior to the Metrical Rule II (MR-II) to derive metrical forms like álliáunce, hábundánce, and éternél from base-generated forms like allíaunce, habúndance, and etérnel.

$$V \longrightarrow [1\ \text{stress}] \ /\ C_o\left[\underset{1\ \text{stress}}{V}\right]C_o\_\_\_C_o\#\quad \text{(SMR)}$$

$$V \longrightarrow [1\ \text{stress}] \ /\ \_\_\_C_oVC_o\left[\underset{1\ \text{stress}}{V}\right]C_o\quad \text{(MR-II)}$$

78 marked instances occur elsewhere. These cases are captured in terms of a larger scope of poetic composition. In the alliterative verses the unmarked pattern occurs 14 times, while the marked one (/́ _ _) does occur 36 times. The Alliterative Rule (AR) derives from base-generated forms like auénture, commáundemens, and coué(i)tise stylistically-motivated forms like áuenture, cómmaundemens, and cóue(i)tise.

$$V \longrightarrow [1\ \text{stress}] \ /\ \#C_o\_\_\_(C_oV)C_o\left[\underset{1\ \text{stress}}{V}\right]C_o\quad \text{(AR)}$$

Finally, let us consider words of Old French origin to which the Main Stress Rule assigns primary stress to the last vowel. To my knowledge, neither nouns nor adjectives of the LME period belong to this class.

Verbs

Ch. ápercéyv 4x (Ⓣ 4x)/ apérceyv 2x; apérseyuede (PP1B 5.143n/PP1C 20.66n), cf. párceyued (PP1B 5.143/PP1C 20.66)
Ch. cómprehénde 6x (Ⓣ 4x), cf. compréndе 3x
Ch. cóndescénde(d) 2x (Ⓣ 2x)
Ch. éntermédle(d) 2x (Ⓘ 1x)
Ch. éntermétte 11x (Ⓣ 2x); enterméten (PP1B 11.408)/ éntermeten (PP1B 13.291)
Ch. éntrecháunged(en) 2x
Ch. mýsconstrúwe TC 1.346 (Ⓣ 1x)
Ch. mýsdepárteth B.ML 107
Ch. récoménde(th) 9x (Ⓣ 4x); récomaund(iþ) (PP1B 15.228/PP1C 17.355n)
Ch. réprehénde TC 1.510 (Ⓣ 1x)
Ch. répresénte 2x (Ⓣ 2x)

In Chaucer's verses the unmarked stress pattern occurs 42 times whereas the marked pattern occurs only twice (see Table 3 above). The Metrical Rule II applies to base-generated forms like comprehénde and représente to derive metrical forms like cómprehénde and répresénte. In the alliterative lines, however, marked cases exceed unmarked ones with a ratio of five to one. The Alliterative Rule applies to base-generated forms like enterméten and recomáund to yield rhetoric forms like éntermeten and récomaund.

The data illustrated from both of the alliterative and metrical verses of late Middle English strongly support my approach that the Main Stress Rule assigns linguistic stress to the required vowel and later modification by stylistic rules yields alliteration or metrical stress according to rhetorical necessity.

CONCLUSION

The comparison between poetic licenses in metrical and alliterative poems has allowed us to clearly characterize the unmarked stress patterns of late Middle English which reflect the linguistic intuition of native speakers of that period. The formulation of the Main Stress Rule. in turn, sheds light on artistic creativity in alliterative and metrical verses. Reflection on multi-layered linguistic data thus enables us to discriminate what is generated from what is derived even though we are not allowed to obtain judgment immediately from native speakers of the late Middle English period.

REFERENCES

[1] M. Halle and S. Keyser, «English Stress», Harper & Row, 1971.
[2] T. Nakao, «The Prosodic Phonology of Late Middle English», Shinozaki Shorin, 1977.
[3] T. Matsushita, "Word-Stress Assignment in LME" «Studies in Eng Lit», LVII, 1985.
[4] N. Chomsky and M. Halle, «The Sound Pattern of English», Harper & Row, 1968.
[5] B. ten Brink, «Chaucers Sprache und Verskunst», Chr. Herm. Tauchnitz, 1899.
[5'] N. F. Robinson, ed., «The Works of Geoffrey Chaucer», OUP, 1957.
[6] J. Anderson ed. «Patience», ManchesterUP, 1969.
[7] J. Anderson ed.«Cleanness», ManchesterUP, 1977.
[8] H. L. Savage, ed., «St. Erkenwald», The Shoe String Press, 1972.
[9] J. Tolkien and E. Gordon, ed., «Sir Gawain and the Green Knight», Clarendon Press, 1967.
[10] W. W. Skeat, ed., «The Vision of Piers Plowman, Text A», EETS 28, 1867.
[11] W. W. Skeat, ed., «The Vision of Piers Plowman, Text B», EETS 38, 1869.
[12] W. W. Skeat, ed., «The Vision of Piers Plowman, Text C», EEST 54, 1873.
[13] H. B. Wheatley, ed., «Manipulus Vocabulorum , (1570)», EETS 27, 1867.

CORRELATS ACOUSTIQUES DE L'ACCENT DE MOT EN PORTUGAIS BRESILIEN

JOÃO ANTÔNIO DE MORAES

Conselho Nacional de Desenvolvimento
Científico e Tecnológico (CNPq)
Rio de Janeiro, Brésil

RESUME

Les indices prosodiques de l'accent lexical en portugais ont été étudiés en fonction de la position de la syllabe tonique dans le mot, de la position du mot dans la phrase et de la modalité de l'énoncé. Les modulations de durée et d'intensité sont des marques constantes de l'accent. Les variations de la FO indiquent la place de l'accent uniquement en position finale de groupe prosodique, où il y a coïncidence entre accent de mot et accent de phrase.

1. INTRODUCTION

Le portugais est une langue à accent libre, dont la place, phonologiquement pertinente (/'sabia/ sage (fem.),/sa'bia/ il savait,/sabi'a/un nom d'oiseau), se limite toutefois aux trois dernières syllabes du mot.
Traditionnellement décrit comme un accent dynamique, d'intensité /1/, ce n'est que récemment qu'il a eu sa projection acoustique étudiée instrumentalement, ce qui a rendu évidente la participation d'autres paramètres prosodiques dans sa réalisation phonétique.
Ainsi, pour Fernandes /2/,ses corrélats acoustiques sont,dans l'ordre décroissant d'importance,la durée, la fréquence fondamentale et l'intensité. Martins /3/, étudiant l'accentuation en portugais européen, attribue uniquement à la durée et à l'énergie le statut d'indices acoustiques (prosodiques) de l'accent de mot, tandis que Major /4/ estime que la durée est le corrélat primaire de l'accent de mot en portugais brésilien, étant donné que, selon cet auteur, les modulations de la FO et de l'intensité ne se rapportent pas toujours à la place de la syllabe tonique du mot.
Lorsqu'on étudie la projection acoustique de l'accent lexical, il s'impose d'abord d'essayer de contrôler un nombre important de variables qui interfèrent avec le comportement des trois paramètres prosodiques qui manifestent,ou peuvent manifester, la proéminence accentuelle.
Ces variables peuvent être classées en trois catégories, selon qu'elles sont extralinguistiques, paralinguistiques ou linguistiques, comme on peut le voir dans le tableau suivant:

1.facteurs extralinguistiques

- niveau de la syllabe { intrinsèques / co-intrinsèques
- niveau de l'énoncé { ligne de la déclinaison de FO / "decrescendo" d'intensité / allongement final

2.facteurs paralinguistiques

- tempo
- niveau global d'intensité
- registre

3.facteurs linguistiques

- position du mot dans la phrase (interférence avec l'intonation syntaxique)
- modalité de la phrase (interférence avec l'intonation modale)
- position de la syllabe tonique dans le mot (schéma accentuel du mot)

2. METHODES

Cherchant à décrire la réalisation de l'accent de mot en portugais sans négliger les problèmes concernant les interférences avec les variables que nous venons de citer, nous avons élaboré un corpus composé de 36 énoncés, où l'on faisait varier systématiquement a) la position de la syllabe tonique dans le vocable _ mot oxyton, paroxyton, proparoxyton, b) la position du mot dans la phrase _ mot isolé, début d'énoncé, fin d'énoncé, fin de topique, à l'intérieur de topique, à l'intérieur de commentaire, et c) la modalité de la phrase _ assertion et interrogation.
Les vocables originaux qui occuperaient les positions-clés pour l'étude des marques accentuelles ont été substitués dans chaque phrase par la séquence trissyllabique PIPIPI avec ses trois schémas accentuels possibles, les variations microprosodiques (intrinsèques et co-intrinsèques) étant ainsi neutralisées.
Le corpus a été dit d'une manière neutre, en local insonorisé, par huit sujets (quatre hommes et quatre femmes) de niveau universitaire, âgés de 22 à 33 ans,originaires de Rio de Janeiro. L'enregistrement résultant a été soumis au Mingograph 34,relié à un détecteur de mélodie et à un intensimètre.

3. RESULTATS

L'analyse des mingogrammes nous a permis de réunir en deux groupes les corrélats acoustiques de l'accent de mot, selon que le vocable porteur de l'accent occupe : a) une position forte dans la phrase, c'est-à-dire,qu'il est en position finale de groupe prosodique(GP),où il y a coïncidence entre la localisation de l'accent lexical et la perception d'un accent phrasal; b) une position faible, interne par rapport au GP, où la syllabe tonique, bien que perçue comme telle dans le cadre du vocable,n'est plus sentie comme proéminente au niveau de l'énoncé.
Les paramètres prosodiques, fréquence, intensité et durée, se comportent dans ces contextes distincts de la façon suivante:

3.1. Positions Fortes

3.1.1. Fin d'énoncé et mot isolé : FO _ chute ou augmentation sur la tonique par rapport à la syllabe précédente, selon que la modalité de la phrase est assertive ou interrogative. Lorsqu'il y a des syllabes postoniques, la fréquence continue à tomber doucement dans les phrases assertives; dans les interrogatives elle revient au niveau de base.
Intensité _ l'intensité indique la place de la syllabe tonique par une chute substantielle sur la syllabe postonique,niveau qui se maintient bas s'il y a plus d'une syllabe postonique. Dans les mots oxytons, comme il n'y a pas de syllabe postonique, cette chute n'existe pas (Tableau 1).Dans les énoncés interrogatifs, l'intensité ne se manifeste pas seulement de manière négative sur les postoniques, mais également de manière positive (c'est-à-dire, par une légère augmentation) sur les syllabes toniques.
Durée _ on observe une augmentation significative de la durée de la voyelle de la syllabe tonique par rapport à la durée moyenne des voyelles atones,augmentation qui est de 85.3 % dans les assertions et de 107.4 % dans les interrogations.

3.1.2. Fin de GP interne : le fonctionnement des trois paramètres est similaire à celui qui a été observé dans les contextes antérieurs. On remarquera que les modulations de FO sur la syllabe tonique sont toutefois moins amples et dans le sens inverse de celles qu'on trouve en position finale d'énoncé, étant positives (élévation de FO) dans l'assertion et négatives (chute de FO) dans l'interrogation. En ce qui concerne la durée, son augmentation sur la tonique des mots oxytons est ici plus réduite qu'en position finale (Tableau 2).

3.2. Positions Faibles

FO _ dans les mots oxytons il n'y a pas de modification significative de la ligne mélodique sur le vocable; dans les autres schémas accentuels on observe une chute de la FO sur les postoniques.
Intensité _ l'intensité diminue sur les postoniques; il n'y a donc pas de modulation objective d'intensité sur les oxytons.Dans les proparoxytons on peut constater une tendance à la diminution de l'intensité de la première à la deuxième postonique.
Durée _ dans ce contexte, la durée en tant qu'indice accentuel, quoique présente à un niveau significatif, est moins marquée que dans les cas précédents. Dans les mots oxytons,il n'y a pas de marque objective de durée (Tableaux 3 et 4).

(Porque Pedro está assim ? / Pourquoi Pedro est-il comme ça ?)
Ele viu o pipipi. / Il a vu le pipipi.

a. vocable oxyton

paramètres	Fréquence(QT)			Intensité(dB)			Durée(cs)		
syllabes	s1	s2	s3	s1	s2	s3	s1	s2	s3
moyennes	0	-0.25	-8.5	0	-1.75	-3.12	6.06	6.06	13.25
écart type	–	1.75	2.07	–	2.43	6.38	1.35	1.45	2.62
p <0.05		*			*			*	

b. vocable paroxyton

paramètres	Fréquence(QT)			Intensité(dB)			Durée(cs)		
syllabes	s1	s2	s3	s1	s2	s3	s1	s2	s3
moyennes	0	-6.9	-10	0	3.14	-14.5	5.64	10.64	7.37
écart type	–	3.48	4.90	–	10.30	10.17	1.31	2.08	2.87
p <0.05		*	*		*			*	

c. vocable proparoxyton

paramètres	Fréquence(QT)				Intensité(dB)				Durée(cs)			
syllabes	s0	s1	s2	s3	s1	s2	s3		s1	s2	s3	
moyennes	6.14	0	-6	-5	0	-11.71	-19.62		8.87	4	6.2	
écart type	2.54	–	4.42	2.58	–	11.87	11.20		1.75	1.70	2.25	
p <0.05		*	*			*	*			*		

Tableau 1.Comportement de la FO, de l'intensité et de la durée des voyelles de vocable oxyton (a), paroxyton (b) et proparoxyton (c) en position finale d'assertion.Les valeurs de la FO sont en quarts de ton (QT), celles de l'intensité en dB,et celles de la durée en centisecondes. Les valeurs de la FO et de l'intensité représentent des intervalles mesurés par rapport à la première syllabe du vocable, pouvant être positifs ou négatifs. s1, s2 et s3 sont respectivement la première,la deuxième et la troisième syllabe du vocable, s0 étant, dans les mots proparoxytons, la syllabe qui précède immédiatement la première syllabe du vocable. L'astérisque * indique que la différence entre les intervalles (ou les moyennes vocaliques, dans le cas de la durée) est significative au seuil de probabilité p <0.05. Les tests statistiques qui ont été appliqués sont ceux de Wilcoxon-Mann-Whitney, pour la FO et l'intensité, et celui du T de Student-Fisher, pour la durée.

4. DISCUSSION ET CONCLUSIONS

4.1. L'analyse effectuée nous a amené à classer les marques accentuelles en trois catégories, non exclusives: marques positives/négatives, directes/indirectes et unilatérales/bilatérales.
Le fait que la marque soit positive ou négative se rapporte à l'augmentation ou à la chute du paramètre prosodique qui signale la localisation de l'accent. La FO, par exemple,fonctionne,en position finale d'énoncé, comme une marque positive dans l'interrogation et négative dans l'assertion.
Une marque est directe quand elle se manifeste sur la syllabe tonique (les deux exemples antérieurs), et indirecte lorsque c'est la postonique qui est signalée en premier lieu,c'est-à-dire,lorsque c'est sur la syllabe postonique qu'on trouve les modifications prosodiques, comme c'est le cas pour l'intensité en ce qui concerne les mots paroxytons dans n'importe quelle position.
Une marque est classée comme bilatérale ou unilatérale selon que la syllabe tonique contraste avec les deux syllabes atones qui lui sont contiguës, ou avec une seule. La FO fonctionne comme marque bilatérale en position finale d'interrogation, et comme marque unilatérale en position finale de topique assertif.

(Como ele viu o pipipi ? / Comment a-t-il vu le pipipi ?)
Ele viu o pipipi picado. / Il a vu le pipipi déchiré.

a. vocable oxyton

paramètres	Fréquence(QT)			Intensité(dB)			Durée(cs)		
syllabes	s1	s2	s3	s1	s2	s3	s1	s2	s3
moyennes	0	-0.14	5.43	0	0.83	0.83	5.42	7	8.83
écart type	–	2.34	3.60	–	3.87	3.18	0.58	2.45	2.21
p<0.05		*						*	

b. vocable paroxyton

paramètres	Fréquence(QT)			Intensité(dB)			Durée(cs)		
syllabes	s1	s2	s3	s1	s2	s3	s1	s2	s3
moyennes	0	5.43	6.33	0	-0.43	-12	5.86	8.71	3.8
écart type	–	2.57	5.13	–	2.44	9.40	1.07	1.73	1.52
p<0.05		*			*		*	*	

c. vocable proparoxyton

paramètres	Fréquence(QT)				Intensité(dB)				Durée(cs)			
syllabes	s0	s1	s2	s3	s1	s2	s3		s1	s2	s3	
moyennes	-4.29	0	-0.14	0	0	-6.87	-7.87		10.12	5.62	6.31	
écart type	2.14	–	1.77	2.77	–	3.18	6.17		2.36	1.75	1.67	
p<0.05		*				*				*		

Tableau 2. Comportement de la FO, de l'intensité et de la durée des voyelles de vocable oxyton (a), paroxyton (b) et proparoxyton (c) en position finale de topique en assertion (voir légende du tableau 1).

(Quando Pedro viu o pipipi picado ? / Quand est-ce que Pedro a vu le pipipi déchiré ?)
Ele viu o pipipi picado quando saia de casa. / Il a vu le pipipi déchiré quand il sortait de chez lui.

a. vocable oxyton

paramètres	Fréquence(QT)			Intensité(dB)			Durée(cs)		
syllabes	s1	s2	s3	s1	s2	s3	s1	s2	s3
moyennes	0	-0.43	-0.71	0	0.86	1.86	4.79	5.71	5.79
écart type	–	1.27	1.70	–	2.11	3.53	0.64	1.25	1.04
p<0.05	(différences non significatives)								

b. vocable paroxyton

paramètres	Fréquence(QT)			Intensité(dB)			Durée(cs)		
syllabes	s1	s2	s3	s1	s2	s3	s1	s2	s3
moyennes	0	-1.62	-5.75	0	0.87	-13.71	6.06	8.19	4.33
écart type	–	2.92	4.19	–	3.31	8.67	1.24	0.96	2.29
p <0.05		*			*	*		*	*

c. vocable proparoxyton

paramètres	Fréquence(QT)				Intensité(dB)				Durée(cs)			
syllabes	s0	s1	s2	s3	s1	s2	s3		s1	s2	s3	
moyennes	3.71	0	-4.12	-4.67	0	-7.5	-12.29		8.87	5.37	6.08	
écart type	2.63	–	1.36	1.51	–	3.55	11.70		1.69	2.37	1.43	
p <0.05		*	*			*	*			*		

Tableau 3. Comportement de la FO, de l'intensité et de la durée des voyelles de vocable oxyton (a), paroxyton (b) et proparoxyton (c)en position interne par rapport au groupe prosodique (topique) en assertion (voir légende du tableau 1).

4.2. Quant à la participation des trois paramètres suprasegmentaux dans les divers contextes examinés, les données indiquent qu'en position finale de GP on a la conjonction de la FO, de l'intensité et de la durée dans l'extériorisation de l'accent. En position interne par rapport au GP, seules les variations de l'intensité et de la durée se rapportent à la localisation de l'accent.

4.3. La chute de la FO observée sur les syllabes postoniques de vocables en position interne par rapport au GP doit être attribuée à la chute de l'intensité qu'on trouve sur ces syllabes dans ce contexte. Elle ne constitue donc pas une marque négative et indirecte de l'accent, ce qui serait un comportement aberrant de la FO. Ce même phénomène _ chute de la FO sur les postoniques _ apparaît d'ailleurs dans d'autres contextes, comme en position finale d'énoncé assertif, où la modulation de la FO fonctionne, de toute évidence, comme marque directe de l'accent.

4.4. De même, la plus grande participation de l'intensité sur la tonique finale de la phrase interrogative, par rapport à celle de la phrase assertive correspondante, doit être attribuée à la dépendance relative qui existe entre fréquence et intensité, l'élévation de la FO sur la tonique étant responsable ici de l'augmentation de l'intensité avec laquelle la syllabe est émise.

4.5. L'allongement de la tonique en finale d'interrogation par rapport à celle de la phrase assertive correspondante doit être également expliqué par des raisons physiologiques, le temps nécessaire à une montée mélodique étant plus long que le temps nécessaire à une descente de même amplitude /5/, /6/.

4.6. Le fait que l'intensité fonctionne comme marque négative et indirecte de l'accent(chute sur les postoniques) peut expliquer que des études instrumentales antérieures ne l'aient pas considérée comme indice accentuel /3/,/4/. On remarquera que dans les mots oxytons, comme il n'y a pas de syllabes postoniques, la chute caractéristique de l'intensité n'existe pas, ce qui ne signifie pas pour autant que ce paramètre ne signale plus la place de l'accent dans ces cas : sa marque sera alors justement l'absence de modulation objective d'intensité sur les syllabes du mot.

4.7. Le même phénomène se produit avec la durée en position interne par rapport au GP pour les oxytons, d'où il résulte que, dans ce cas spécifique, nous nous trouvons devant une situation qui se caractérise par l'absence de marque objective de l'extériorisation de l'accent, sans que cela affecte pour autant sa perception sur la dernière syllabe du vocable.

4.8. La question souvent soulevée d'une marque prosodique unique, ou du moins plus importante dans l'extériorisation de l'accent de mot,manque de fondement en ce qui concerne l'accentuation en langue

schéma accentuel \ position du mot	Fin d'énoncé	Fin de GP interne	Interne p/ rapport au GP
oxyton	152 %	31 %	12 %
paroxyton	61 %	66 %	65 %
proparoxyton	67 %	70 %	67 %

Tableau 4. Augmentation moyenne de la durée de la voyelle de la syllabe tonique par rapport à la durée des voyelles atones du vocable, selon son schéma accentuel et la position que le mot occupe dans l'énoncé.

portugaise, vu que les trois dimensions suprassegmentales jouent un rôle très important. Un examen plus attentif de la question du poids relatif des différentes marques accentuelles nous amène à établir une distinction entre les marques qui sont plus constantes, et celles qui sont plus fortes, c'est-à-dire, qui dominent en cas de conflit. Les marques plus constantes sont l'intensité et la durée (ce qui ne nous permet pas cependant de les identifier avec la notion d'énergie, puisqu'elles se rapportent à des syllabes distinctes). D'autre part, la marque dominante semble être, du moins dans certains contextes, la fréquence fondamentale, hypothèse qu'il reste à confirmer à l'aide de la synthèse de la parole dans des travaux futurs.

REMERCIEMENTS

Cette recherche a été réalisée sous le patronage du CNPq (procès 302445-84/LA). Les tracés mingographiques ont été obtenus à l'Institut de Phonétique de l'Université de Paris III, auquel nous exprimons notre gratitude. Nous remercions également Mme. Maria D. Cabral pour son aide en ce qui concerne l'application des tests statistiques.

REFERENCES BIBLIOGRAPHIQUES

/1/ J.M. Câmara, "O vocábulo em português". In: J.M. Câmara"Problemas de Lingüística Descritiva",8ème ed, Petrópolis: Vozes, 1976, pp. 34-39.

/2/ N.Fernandes,"Contribuição para uma Análise Instrumental da Acentuação e da Intonação do Português". Dissertation de Maîtrise non publiée,Universidade de São Paulo, 1976.

/3/ M.R.Martins,"Aspects de l'Accent en Portugais". Thèse de Doctorat de 3ème cycle,Université de Strasbourg,1977(publiée par Helmut Buske,Hambourg, 1982).

/4/ R. Major, "Stress and rhythm in brazilian portuguese", Language 61(2): 259-282, 1985.

/5/ J.O'hala, W. Ewan,"Speed of pitch change",Journal of the Acoustical Society of America 53(1):345 A, 1973.

/6/ J. Sundberg, "Maximum speed changes in singers and untrained subjects", Journal of Phonetics 7(2): 71-79, 1979.

REGLES D'ACCENTUATION EN GREC MODERNE

Argyro TSEVA Michel CONTINI

Institut de la Communication Parlée - Institut de Phonétique de Grenoble
Université des Langues et Lettres
B.P. 25 X 38040 GRENOBLE CEDEX - FRANCE

RESUME

L'apport de cette étude réside dans la formulation de règles d'accentuation des mots accentogènes. Elle examine à quel niveau de langue se situent la plupart des règles de prévisibilité et quelles sont leurs performances. Il est demontré que la détermination de la place de l'accent, sur le plan synchronique, ne peut pas toujours être prévue à partir de la structure morphologique du mot, mais, dans une large mesure, en tenant compte du phonétisme en final du mot.

Le grec moderne possède un accent à liberté limitée. Il s'agit d'un accent libre, avec toutefois une restriction : il ne remonte pas au-delà de la troisième syllabe en partant de la fin du mot, quel que soit le nombre des syllabes (1).

Si on veut se baser sur l'état synchronique de la langue (2), la détermination de la place de l'accent à partir de la reconnaissance de la structure morphologique n'apparaît pas satisfaisante dans tous les cas (P. GARDE, 1968). Ainsi, nous référons-nous à certaines catégories de substantifs dérivés dont la place de l'accent ne peut se définir qu'après avoir recours au mot dont ils sont issus.

A titre illustratif, nous donnons comme exemple les substantifs dérivés d'un verbe, de genre féminin à désinence /-a/.

Exemples : ['jɛn-a] "accouchement" dérivé de [jɛ'n-ɔ] "j'accouche"

[prɔs-fɔ'r-a] "l'offre" dérivé de [prɔs-'fɛr-ɔ] ou [prɔs-'fɛrn-ɔ] "j'offre".

Pour cette catégorie de substantifs, dépourvus du morphème dérivatif, l'accent frappe la syllabe pénultième s'il n'y a pas changement de la voyelle thématique du verbe dont le substantif est issu et sur la syllabe finale s'il y a changement. Cependant bien que la place de l'accent soit liée à la structure morphologique du mot d'origine, on ne peut pas soutenir qu'elle dépende des propriétés accentuelles des morphèmes du mot dérivé.

De même, la place de l'accent pour certaines catégories de mots ne peut être expliquée que par référence à des états antérieurs de la langue : Le grec moderne a généralement maintenu l'accent sur la même syllabe qu'en grec ancien, à quelques exceptions près.

A titre illustratif, la distinction des formes verbales (première personne de l'indicatif présent, voix active) en oxytons ou paroxytons s'explique par la loi de la limitation quantitative (3), attestée dès l'antiquité, suivie d'un phénomène de contraction qui s'est produit à l'époque classique pour tous les verbes comportant une voyelle thématique finale (/A/, /E/ ou /O/). Cette dernière s'est agglutinée avec la voyelle de la désinence et a déplacé ainsi l'accent sur la syllabe finale, si la voyelle thématique finale était initialement accentuée.

Exemples :

ἀγαπάω	"j'aime"
παιδεύω	"j'éduque"
παιδεύομαι	"je m'éduque"

Nous avons effectué un test préliminaire qui nous a permis de constater que la délimitation de la place de l'accent est possible dans une large mesure à partir de la structure finale du mot, position la plus riche en information morphologique où l'on retrouve toujours la désinence et les morphèmes dérivatifs, s'ils en existent. Cette procédure apparait également satisfaisante pour les mots qui comportent uniquement un lexème et une désinence, pour lesquels les

trois possibilités accentuelles sont admises :

Exemples : [ura'n-ɔs] "ciel"
[']ðrɔm-ɔs] "rue"
['anθrɔp-ɔs] "homme"

Bien que la désinence (au nominatf singulier), à quelques exceptions près, ne permette pas la délimitation de la place de l'accent, celle-ci est possible en liaison avec la structure du lexème qui la précède.

A notre avis, la méthode présentée ci-dessus, apparait comme la solution la plus satisfaisante pour la délimitation de la place de l'accent. Pour la première fois, les règles d'accentuation, ainsi que les exceptions signalées à l'intérieur de chaque règle, incluent tous les mots existants à l'intérieur d'une catégorie grammaticale du vocabulaire grec.

Les règles formulées ont été testées à partir d'un dictionnaire du grec moderne dont tous les mots, ont été stockés sur mini-ordinateur LSI 11/73 de Digital dans un fichier pour édition et traitement, (au total 45 750 mots).

Nous limitons notre champ d'investigation à deux catégories de mots accentogènes : les substantifs (formes déclinables au nominatif singulier et formes indéclinables) et les verbes (indicatif présent, première personne du singulier) (au total 31 734 mots).

Dans notre démarche, après distinction des mots en verbes ou en substantifs, nous avons reclassé les substantifs en sous-catégories, d'après leur désinence, sans tenir compte du genre. A l'intérieur de chacune de ces dernières, on retrouve les différentes structures syllabiques finales précédant la désinence, le type d'accentuation le plus répandu ainsi que la liste de toutes les exceptions. Parfois, à l'intérieur des cas exceptionnels, d'autres règles d'accentuation (sous-règles) sont formulées.

Ainsi, par exemple, 418 occurrences sur 515 existantes de substantifs se terminant par [-ða] sont accentuées sur la syllabe pénultième (ex. : [ɛfimɛ'riða] "journal"). Parmi les exceptions, 74 occurrences sur 76 se terminant par [-itiða] et tous les mots composés du substantif ['fluða] "écorce" (6 occurrences) sont accentués sur la syllabe antépénultième (ex. : [fle'vitiða] "phlébite", [lemɔ'nɔfluða] "écorce de citron"). Enfin on retrouve le mot "fée" qui présente deux formes; [nɛ'raiða] et [anɛ'raiða] En définitive, la place de l'accent est prévisible dans 497 occurrences (418 + 73 + 5 + 1) sur 515.
La règle annoncée, ci-dessus, sera donc présentée comme suit :

[-ða] (418 occurrences sur 515)
› 497 sur 515

Exemple : [ar'kuð-a] "ours"

Exceptions : Accent sur la syllabe antépénultième

[-iða]
- les substantifs qui se terminent en [-itiða] (76 occur.), ex. : [fle'vitiða] "phlébite", [ðina'mitiða] "dynamite" sauf [ci'tiða], [ri'tiða].

- les substantifs [ayri'ɔjiða], ['vɔtriða], [erasi'texniða], ['eriða], [i'cetiða], ['iriða], [kali'texniða], ['kɔniða] ou [kɔ'niða], [narkɔ'θetiða], [(a)nɛ'raiða], [ksi'lɔviða], [para'statiða], ['zmiriða], ['trɔpiða].

[-uða]
les mots composés avec le substantif ['fluða]/ "écorce" (6 occurrences), ex. : [lemɔ'nɔfluða] "écorce de citron".

Les résultats montrent que la structure finale du mot permet la délimitation de la place de l'accent dans 80 % des cas; 25 433 occurrences sur 31 734 mots étudiés (substantifs et verbes). Avec l'élaboration de sous-règles (complétant les 243 règles principales), appliquées aux cas exceptionnels, on peut atteindre le pourcentage de 89 % (28 329 occurrences). Le 11 % restant est traité sur une liste d'exceptions. Ainsi sur un très vaste lexique, nous arrivons à un système de règles qui fonctionne 9 fois sur 10.

Cette étude se veut une contribution à l'explication du fonctionnement linguistique de l'accent; elle nous semble directement exploitable :

- en linguistique appliquée : apprentissage du grec moderne, langue étrangère,

- en synthèse : exploitation automatique de textes simplifiés (sans marques accentuelles), par exemple : telex...,

- en reconnaissance automatique : récupération de l'accent à partir de la chaîne phonétique.

NOTES :

(1) Pour les problèmes d'ensemble relatifs à l'accent en grec moderne nous renvoyons à A. TSEVA (1987).

(2) Une autre démarche avec notamment des références diachroniques a été présentée par H. TONNET (1984).

(3) L'accent ne peut pas dépasser l'avant dernière syllabe du mot si la voyelle finale est longue.

REFERENCES BIBLIOGRAPHIQUES :

P. GARDE, L'Accent. Presses Universitaires de France, Paris, 1968. 172 p.

A. GEORGOPAPADAKOS, Le Grand dictionnaire de la langue néo-hellénique (en grec). Malliaris - Phaidheia, Athènes, 1984. - 1184 p.

H. TONNET, Manuel d'accentuation grecque moderne (démotique). Klincksieck, Paris, 1984. - 112 p.

A. TSEVA, Contribution à l'étude de l'accent en grec moderne. Règles de prévisibilité et analyse instrumentale. Thèse de Doctorat, Université de Grenoble III, 1987. - 435 p

AN EXPERIMENTAL ANALYSIS OF THE FIVE LEVEL TONES OF THE GAOBA KAM

Shi Feng
Dept.of Chinese
Nankai University
Tianjin, China

Shi Lin
Dept.of Chinese
Nankai University
Tianjin, China

Liao Rong-rong
Dept.of Chinese
People's University
Beijing, China

ABSTRUCT
The analysis of tonemic systems with five level tones is of great sinificance in the study of tone. Languages with five level tones are rarely found in the world. We have not found any language that has more than five level tones. This is the relection of the physiological adjusting process of the local folds. The on-glides may give the informations for us to distinguish different tones possibly.
This experiment shows the pitch range of tone changes in plysyllabic utterances. The general tendency is towards raising of the five level tones, four tend to rise and only the highest one both rises and falls.

I. INTRODUCTION

Languages with five contrastive level tones are rarely found in the World.
It is generally considered that five level tones are the maximum number that a language can possibly have. For this reason, the " tone letters " created by Y-R Chao have only five pitch levels.[1] In order to distinguish five level tones, S-Y. Wang proposed the feature Mid in the system of distinctive features of tones.[2] Maddison considered the maximum number of five pitch levels as the first of the universal features of tone languages in the world. [3]
The analysis of tonemic systems with five level tones is of great significance in the study of tone. What we frequently encounter in our studies are ones systems of languages or dialects which have one, two or three level tones. There is a considerable physiological and psychological space between neighboring level tones, whose acoustic between neighboring level tones, acoustice behavior displays a considerable degree of freedom.And falling or rising tones are often distinguished from each other only interns of pitch contour without being limited to their pitch levels. Therefore, it will enrich our knowledge of tone to study the articulation of the five level tones in the same system, to observe their acoustic behavior and features of recognation, and to perceive their relationship in monosyllables and their variations occurring in polysyllabic words. This is helpful both theoretically and in practice.

II. THE GAOBA KAM

The sound system of Gaoba kam has only 20 initials and 29 finals, (which the southern dialect of kam has 30 initials and 56 finals and Chinese Putonghua has 21 initials and 35 finals.)

The tone system of Gaoba Kam is comparatively more complicated. It has 9 long tones and 5 short tones. In fact, Gaoba Kam has only 9 distinctive tones in terms of pitch, because pitch values of all the short tones are equal to those of the related long tones. In terms of pitch value, Gaoba Kam has 5 level tones, 3 rising tones, and one falling tone, which can be illustrated as follows:
Tone 1 is a high rising tone,the pitch value being 45;
Tone 1' is a low level tone, the pitch value being 11;
Tone 2 is a low-mid level tone, the pitch value being 22;
Tone 2 is a mid level tone, the pitch value being 33;
Tone 3' is a low-rising tone, the pitch value being 13;
Tone 4 is a low falling tone, the pitch value being 31;
Tone 5 is a high level tone, the pitch value being 55;
Tone 5' is a mid rising tone, the pitch value being 24;
Tone 6 is a high mid level tone, the pitch value being 44;
The five level tones above are thus tone 1' (11), tone 2 (22), tone 3 (33), tone 6 (44) and tone 5 (55).
The present paper presents an elementary analysis of the acousitc representation of level tones in the phonological system on the basis of experimental data.

III. THE EXPERIMENT

The speech sample consisted of the following 3 types of read materials.
(1) Monosyllabic words;
(2) Polysyllabic words;
(3) Sentences;

After a native Gaobanese was recorded in the recording room, the anlysis was made Key 7030` sound spectrograph, which produced narrow band spectrograms and amplitude displays for each signal.
The parameters of each syllable's pitch, length and intensity were thus measured and counted:
(1) Length
Syllabic length; pitch length.
(2) Intensity
The greatest syllabic intensity.
(3) Pitch
Five points are selected on the pitch contour for measurement the pitch length is divided into fur sections. Starting, middle and ending point are marked first, and then two subpoint are also marked respectively between the starting and middle,the middle and ending point. Thus we have five points: a,b,c,d and e, of which a is the starting point, c the middle point, and e the ending point. The frequencies of the five points are measured in Hz.

IV. ANALYSIS

1) Tone in monosyllablès
We have examined the variations of the five level tones in different phonological environments according to their behavior in monosyllables. The results show that all of the five level tones behave quite similarly in terms of pitch. They have the same pitch shapes and very limited dynamic ranges as well.
From the average frequencies of the five points (point a,b,c,d,e) of each tone, the related pitch values in the pitch range can be obtained. The patch tern curve of each tone is made on a modified plane logarithm coordinate. For convenience of comparison between the level tones, we take the same length from each tone for consideration. (See Fig. 1)
The speaker's pitch range is from 120 H2, the lowest limit, to 259 H2, the highest limit.

Pitch
Tone 1': the tonal contour lies at the bottom of the pitch range. It begins at a rather high position and then falls slightly, appearing to be a low level tone. This is quite consistent with the behavior of the low level tone of the Chinese Tianjin dialect.[4] It seems that, if a level tone is at the bottom of the pitch range, there must be a falling transition in the initial portion, which is the acoustic representation of the physiological process of the larynx from natural phonation to the stable low pitch.

Tone 2: the contour is at the bottom of the speaker's frequency range, about a " half-step " higher than that of tone 1'. The starting point is close to tone 1'. The initial portion does not fall so low as tone 1'.

Tone 3: this contour lies in the lower middle part of the frequency range, about two " half-steps " higher than that of tone 2. Both the inital and terminal portions are close to level.
Tone 6: this coutour is in the higher middle part of the frequency range, about 4 " half-steps " higher than that of tone 3. Both the initial and terminal portions are close to level. This tone is similar to 1st tone of mandarin in tonal coutour.
Tone 5: the coutour is in the upper part of the range. The starting point is a bit lower. The initial portion has a rising transition, the portion after the middle point becomes stable and the terminal portions is close to level. The pattern is that of a convex tone. The last half of the contour is about five " half-steps " higher than tone 6. It has a contour similar to the high rising tone of Tianjin dialect, i.e. they are both in the upper part of the pitch range, both are convex tones, and both have rising transition [5]. This reflects the process of speech organs from natual phonation to the stable high frequenly.
The contours of the five level tones in Fig. 1 appear to be an interesting radiant symmetry. Tone 3 and 6 are in the middle part the frequency and the initial portions of their contours are close to level. Tones 1' and 2 have a falling transition, while tone 5 has a rising transition. Their initial portions are all toward the middle of the pitch range. This is very helpful for us to understand the appearance of the orglide of the tone.
In connection with the appearance of the onglide, Maddison proposed one of his universals of tone: " phonetically Central tones unmarked, Extreme tones are highly marked." [6]
We have also found that, though the five level tones are distinctive auditorily and basically " equidistant " from each other, the "equidistance" is hard to describe with some objective criterion.
The supposed equidistance does not show up when either on the mel scale or the Herz scale, neither linearly, nor logarithmically

The frequency value from 129 Hz of tone 1' to 259 Hz of tone 5 happens to be that of the musical octave from C2 to C3. For the convenience of explanation, we take the musical pitch as our criterion. There are 8 scales including 12 half-steps from C2 to C3. If we mark five levels "equidistantly ", the distance between each level should be 3 half-steps. But the actual situation turns out contradictory to that assumption. In fact, tone 2 is one "half-step from tone 1', tone 3 is two "half-steps" from tone 2, tone 6 is four "half-steps" from tone 3,and tone 5 is five : half-steps' from tone 6. In terms of musical interval, the frequency between high pitches is larger than that between low pitches, i.e. the frequency of each scale is twice that of the correspond-

ing scale in the next octave. But the five level tones are not "equidistant" even in terms of musical interval. This shows that the human ear's recognition of speech sounds differs from that of musical sounds 1' is upward 11 Hz, which is equal to one " half-step ", but it already enters the limits of its neighbor, tone 2. The frequency span of tone 5 is downward 44 Hz, which is equal to three "half-steps", but it still has two " half-steps " to go to reach the limits of tone 6.
In this study, we have found a way in which the relationship between the five level tones can be approximately explained limit of the pitch range from each level tone respectively, we get 9 Hz of tone 1', 18 Hz of tone 2, 37 Hz of tone 3, 77 Hz of tone 6, and 139 Hz of tone 5. In this case , the frequency difference of each tone is twice that of its lower neighbor. This is a special equal proportional relatonship. Its significance is uncertain until more experimental data are collected.
We can see that the frequency values of the three middle points b, c and d of each level tone are relatively stable and have rather small dynamic ranges, representing the stable values of the tones. For example, the frquency variations of tones 1', 2, 3 and 6 are all within 10 Hz. Since we often can not determine the shape of a tone contour because of the lack of an objective criterion, this method can be taken as a useful reference.

2) Tone in polysyllables
The behavior of the five level tones in polysyllables as well as in sentences is illustrated in Fig 2. As compared with their behavior in monsyllables, the difference is early seen and the variations can thus be understood.

Duration
The average syllabic length and pitch length both show the same feature, i.e. it is longer in monosyllables than in polysyllables, and it is the shortest in sentences. Another feature is that the length varies with the position of the syllable. The syllable of the same tone is short at the beginning or in the middle of polysyllabic utteranes, but is longer at the end. All the five level tones behave in this way.

Intersity
In general, the upper limit of the dynamic range of every tone's intensity does not little in polyllables or in sentences, but the lower limit falls. This reflects the fact that speech sounds in polysyllables and sentences show clear cadence, which enhances the difference between strong syllables and weak ones.

Pitch
The most striking feature in terms of pitch is that the pitch range of every tone is obviously expanded. The pitch range of a level tone combined with a contour tone is larger than that of the pitch range of a level tone in sentences is larger than that in polysyllables. Therefore, the highest From the figures, we see that, in polysyllabic utterances, every tone stubbornly keeps and manifests its own characterestics in order to maintain contrast with the other tones. Therefore, although the starting and end points rise and fall a grent deal and the whole pitch contour also rises and falls in the pitch range, the basic tonal pattern does not change, and neither does the relative position of each tone's variations with-tone sandhi operates in Gaoba Kam only at the phonetic level . Tones do not change tonmic class under the influence of contiguous tones.

V. CONCLUSION AND DISCUSSION

The five level tones of Gaoba Kam attest the existence of five distinctive tone levels. But the distances between the neighboring levels are not equal either in terms of pitch frequency, or musical interval. This observation may be significant in the study of the recognition of tone.
This experiment demonstrates the pitch range of tone changes in polysyllabic utterances. The general tendency is towards raising. Of the five level tones, four tend to rise and only the highest one either rises or falls.
In the present study five points were selected to measure and analyze the pitch contours of tone. We have found that there is a big frequency difference between the starting and end points of the variations of same tone. This because of the occurence of the transitions of both on-glides and off-glides. But the frequency difference between the other three points in the contral region of the variations of the smae tone is small This plays an important role in the stable behavior of level tone. The middle point particularly plays a decisive role. One analyst has described the appointed tone level" . In this case, we may çonsider that this tone level should be defined by the middle point of the pitch contur. The frequencies the three middle points are almost identical and the related stable section occupies more half of the whole pitch contour. It may be considered to be the typical informaton-carrying section of the tone pattern.
Finally the five level tones as pronounced in monosyllables are separated from each other in the frequency range with few overlaps. On the other hand, the overlaps occure quite often in polysyllabice utterance, of which the Tone 1', and 2, and 3, Tone 6 and 5, (or even Tone 1 and 3,and Tone 3 and 5 in a few cases) overlap some portion of frequency range. The influence of this phenomenon is a considerable problem to the recognition of tone and to study within the boundary theory of speech sound.

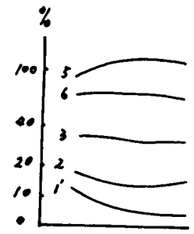


Fig.1a Pattern curves of Five Level Tones in Monosyllables

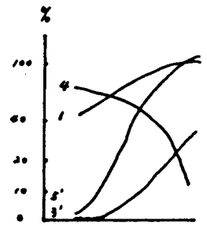


Fig.1b Pattern curves of Four Contour Tones in Monosyllables

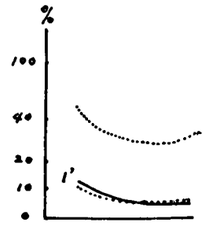


Fig.2a Tone 1' in Polysyllables

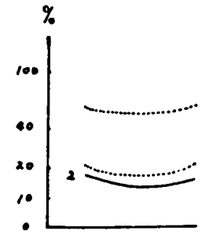


Fig.2b Tone 2 in polysyllables

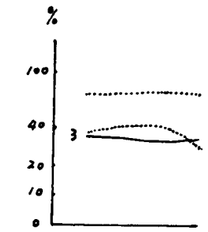


Fig.2c Tone 3 in Polysyllables

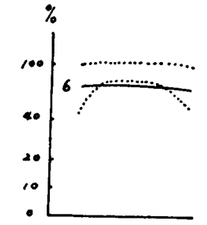


Fig.2d Tone 6 in Polysyllables

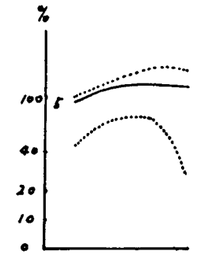


Fig.2e Tone 5 in Polysyllables

Notes:
[1] Yuen-ren chao, a system of tone letters, Le Maitre Phonetique. 45. 24-47 (1930)
[2] William S-Y. Wang, Phonological Features of Tone, International Journal of American Linguistics, Vol. 32, No.2, 93-105 (1967)
[3] I. Maddison, Universals of tone, from Universals of Human language, Vol.2,Stanford Univ. Press, 1978.
[4] Shi Feng, an experimental analysis of the monosyllabic tones in Tianjin dialect. Yu Yan Yan Jiu Lun Cong, Vol. 4 (in print) An experimental analysis of the bisyllabic tones in Tianjin dialect. Yu Ya Yan Jiu vol 10, 1986.
[5] See [4]
[6] See [3]

INTERNAL SPEECH REPRESENTATIONS IN NORMAL AND HEARING-IMPAIRED SUBJECTS

Christopher W. Turner and Lenore A. Holte

Syracuse University

Psychoacoustical input-filter pattens were obtained from individual subjects and were used to calculate the "internal spectra" of acoustic speech and speech-like sounds. The validity of the calculated representations is demonstrated by predictions of the discrimination and recognition performance by normal and hearing-impaired subjects.

INTRODUCTION

Klatt [1] and Chistovich [2] have presented evidence that suggest that the typical listener to a steady-state or slowly varying speech sound derives the phonemic identity of the sound from the detection of local spectral peaks in the overall spectral envelope of the speech sound. The peripheral auditory system has often been modeled as a bank of parallel bandpass filters capable of resolving these spectral characteristics important for speech perception. Specification of the auditory filter characteristics across the frequency range in human subjects allows one to calculate a spectral representation of a peripherally-filtered speech stimulus that the subject's central processing system would then operate upon. This approach has been utilzed by several previous researchers [1,3,4]. Turner and Robb [5] have recently used such representations to predict the audibility of stop consonants.

It is known that persons with sensorineural hearing loss often suffer from, in addition to their elevated sensitivity thresholds, some degree of reduced frequency resolution. This is evidenced by abnormal spread of masking [6,7] and also by broadened psychophysical tuning curves [8,9,10]. More directly relevant to the bandpass-filter models of the auditory system are results showing that sensorineural hearing-loss subjects often display abnormally wide auditory filters [11,12]. Comparisons of vowel-masking patterns from normal and hearing-loss subjects suggest that the formant peaks are less clearly resolved in the "internal representations" of the hearing-loss subjects than in normals [13,14,15].

The present paper describes our measures of auditory filter characteristics in normal and impaired subjects and their relation to the subjects' detection and discrimination of spectral features in speech-like stimuli. The measured filter characteristics are also used to calculate the "internal representations" of speech stimuli, in order to provide an insight to the recognition performance of the subjects. One advantage of using subjects' filter characteristics to calculate "internal representations" of speech signals, in comparison to the vowel-masking pattern approach, is that specification of the filter characteristics in a subject subsequently allows for the calculation of "internal representations" for any arbitrary speech sound. By using both normal-hearing and hearing-loss subjects, whose auditory filtering characteristics vary across a wide range of resolving powers, the dependency of discrimination and recognition performance upon the resolution of the spectral cues can be investigated.

MEASURES OF AUDITORY FILTERING

The masking paradigm chosen for this study has been termed an "input filter pattern" [16]. In this paradigm, the masked threshold for a probe signal at a fixed frequency is measured in the presence of fixed-level maskers across a range of frequencies. The resulting pattern, of probe threshold plotted as a function of masker frequency, is taken to represent a subject's auditory filtering characteristic at the probe frequency location. The level of the (input) masker is held constant, thus the probe threshold reflects the relative output of constant-level input signals following auditory filtering. Thus the input-filter pattern can theoretically be used to yield a measure of the output of each auditory filter. In contrast, the more familiar psychophysical tuning curve reflects the input signal level required to produce a constant output following the filter.

Methods

A forward-masking paradigm was employed for input filter pattern measures in order to reduce potential artifacts such as beats or distortion products. The center frequency for each measured input-filter pattern is determined by the probe frequency; the six probe frequencies used were 250 Hz, 500 Hz, 1000 Hz, 1820 Hz, 3000 Hz and 4000 Hz. Eleven masker frequencies were used to define the shape of the input filter pattern for each probe frequency. The pure-tone maskers were 204-msec in duration including 10-msec rise-fall ramps. All maskers were presented at 95 dB SPL. The probe signal was an pure tone of 25-msec duration including 5-msec rise-fall ramps. Probe onset occured at masker offset.

Results

Figure 1 shows input filter patterns at several probe frequencies from a normal-hearing subject. We interpret these data as depicting a filter shape by viewing the level of the probe threshold at each masker frequency, relative to the peak of each pattern, as the auditory filter attenuation for signals at that particular masker frequency. For the normal-hearing subject's data shown in the first figure, the filters are asymmetrical bandpass in shape. The shallower slope seen on the low-frequency side of the filter is consistent with the well-known phenomenon of upward spread of masking.

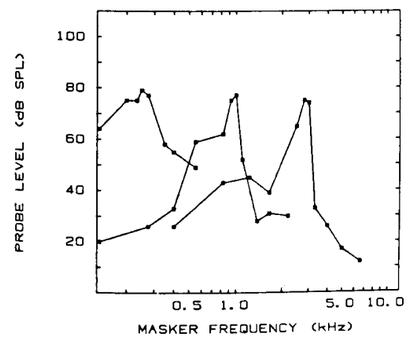


Figure 1. Input filter patterns from a normal-hearing subject

Figure 2 shows input filter patterns at three frequencies from a moderately hearing-impaired subject. The filter patterns are broader in shape than in Figure 1, showing in particular more of a low-pass characteristic, consistent with increased spread of masking in this subject.

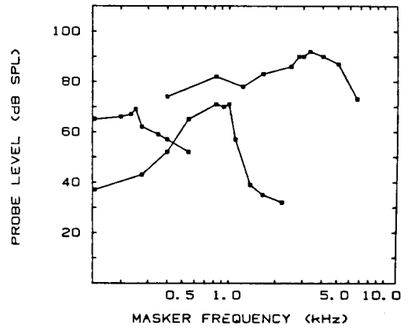


Figure 2. Input filter patterns from a hearing-impaired subject

RELATION OF FILTERING TO FORMANT PERCEPTION

The previous line of reasoning suggests that, for some hearing-impaired patients, the spectral peak characteristics of some normally-presented speech sounds are not distinct enough to allow for normal speech recognition to occur. Turner and Holte [17] demonstrated that a more prominent second-formant spectral peak was required for some hearing-impaired subjects than for the normal-hearing subjects in a discrimination task using speech-like spectral shapes. If auditory filtering characteristics in individual subjects determine the abilities of subjects to resolve the spectral characteristics of speech sounds, then the results of the input filter pattern experiments should offer predictive value for the formant-height detection task described above. Indeed, as described in Turner et al. [12], the bandwidth of subjects' input filter patterns were highly correlated with the minimum detectable formant height for the same subjects (r=-.95). That is, the narrower the measured filter shape for an auditory filter centered at the second-formant frequency, the smaller the spectral peak in the speech-like spectral shape required for equivalent levels of detectability. We therefore concluded that the psychoacoustical input-filter pattern measure is a valid indicator of the perceptual strength of spectral characteristics in individual subjects.

INTERNAL REPRESENTATIONS OF VOWELS

Although the recognition of vowels by both normal and moderately hearing-impaired subjects is generally quite accurate, we determined four vowels for which confusions are obtained. The synthetic vowels /æ, ɛ, ʌ, ɑ/ were constructed using the Klatt synthesis program. Fundamental frequency was varied

slightly throughout the stimulus in order to yield a more natural utterance. All stimuli were presented at 95 dB SPL, at which level the entire stimulus was above each subjects' sensitivity thresholds. The formant frequencies of the vowels were held constant during each vowel and are listed below:

	F1	F2	F3
æ	620(Hz)	1660	2430
ɛ	530	1680	2500
ʌ	620	1220	2550
ɑ	660	1200	2550

Both hearing-loss and normal subjects were tested in a closed-set recognition task using these four tokens. Recognition scores are expressed as the percentage of information transmitted for a 2x2 matrix, which contrasted correct versus incorrect identification for each vowel. Confusions were obtained primarily between the pairs /æ,ɛ/ and /ʌ,ɑ/. For those pairs, the primary distinguishing feature between the vowels was most likely the small differences in F1 frequency, although some small differences at the higher formants may have contributed to recognition. The subjects, both hearing-impaired and normal, rarely confused members of the /æ,ɛ/ pair with members of the /ʌ,ɑ/ pair.

An FFT of each vowel was obtained from the acoustic signal and was then used as the input to a model of the peripheral auditory system based upon input filter pattern measures. The input filter patterns from each subject were then used, along with interpolated filter shapes at intermediate frequencies, to calculate a 200-point "internal spectrum" of each vowel for each subject. This was accomplished via laboratory computer, in which the outputs of each filter were calculated as the sum of the stimulus power passed by each filter.

Figure 3 displays "internal representations" of two vowels /ɑ/ and /ʌ/ for a normal-hearing subject (same subject as in Figure 1). The three formants are quite well represented , due to the narrow auditory filters measured in the subject. The difference in F1 frequency is visible near 600 Hz. This example shows a vowel pair which was sometimes confused by that subject, although his recognition was still quite accurate (96% and 97% respectively). As was the case for all normal-hearing subjects, vowel recognition was near 100% for all tokens. This case demonstrates that fine distinctions in formant frequency location are preserved and available to a subject with normal auditory filtering.

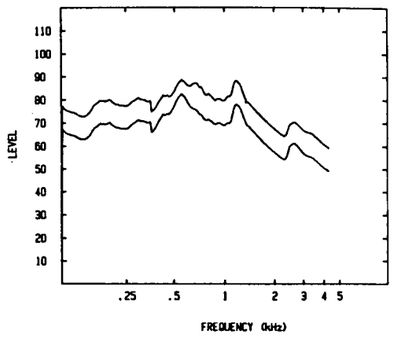


Figure 3. Internal representations for the vowels /ɑ/ (upper) and /ʌ/ (lower) from a normal-hearing subject. The lower curve has been displaced by 10 dB for ease of viewing.

In Figure 4, the calculated "internal representations" of the previous vowel pair, /ɑ ,ʌ/ are displayed for the hearing-impaired subject of Figure 2. These two vowels were often confused by this subject; recognition scores were 69% and 73% respectively. The abnormal frequency resolution of this subject resulted in poorer resolution of the formants in general and in particular, a minimal difference in the internal representations of F1 frequency.

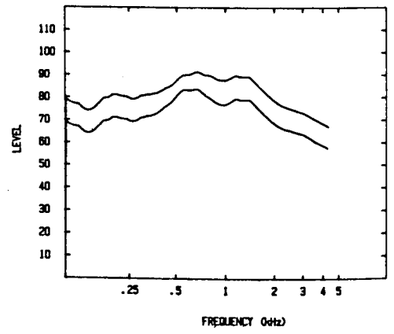


Figure 4. Internal representations for the vowels /ɑ/ (upper) and /ʌ/ (lower) from a hearing-impaired subject.

Figure 5 displays the calculated "internal representations" of the vowel pair /ɑ,æ/ for the same hearing-impaired subject as above. These vowels were never confused by the subject, due to the large difference in F2 frequency, which was preserved even after the abnormal peripheral filtering of this hearing-impaired subject.

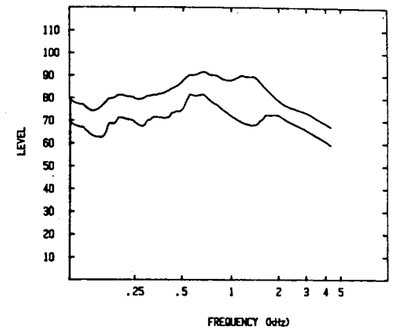


Figure 5. Internal representations for the vowels /ɑ/ (upper) and /æ/ (lower) from a hearing-impaired subject.

Conclusions

The input filter patterns provide a promising approach to expressing the representation of the spectral cues of speech in individual subjects. We have shown that measures of auditory filtering can predict the strength of percept for formant peaks in individual subjects, suggesting that auditory filter measures could be used to dictate the types and degrees of speech-cue enhancement that hearing-impaired subjects may require. The calculated "internal representations" of speech sounds provide a theoretically attractive mechanism for investigating the dependence of speech recognition upon the resolution of spectral detail.

ACKNOWLEDGEMENTS

This research was supported in part by a grant from the Deafness Research Foundation. The authors thank Carol Henn and Swati Lotlikar for assistance in data collection.

REFERENCES

1. Klatt, D.H. Prediction of perceived phonetic distance from crtical-band spectra: A first step. in Proceedings of the International Congress on Acoustics, Speech, Signal Processing. Paris, IEEE pp. 1278-1281 (1982).

2. Chistovich, L., Central auditory processing of peripheral vowel spectra., J. Acoust. Soc. Am., 77, 789-805 (1985).

3. Searle, C.L., Jacobsen, J.F. and Rayment, S.G. (1979) Stop-consonant discrimination based upon human audition., J. Acoust. Soc. Am., 65, 799-809 (1979).

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

5. Turner, C.W. and Robb, M.P. Audibility and recognition of stop consonants in normal and hearing-impaired subjects., J. Acoust. Soc. Am., in press (1987).

6. Jerger, J.F., Tillman, T.W. and Peterson, J.L. Masking by octave bands of noise in normal and impaired ears., J. Acoust. Soc. Am., 32, 385-390 (1960).

7. Trees, D.A. and Turner, C.W. Spread of masking in normal and high-frequency hearing-loss subjects., Audiology, 25, 70-83 (1986).

8. Wightman, F.L., McGee, T. and Kramer, M. Factors influencing frequency selectivity in normal and hearing-impaired listeners., in Psychophysics and Physiology of Hearing, Evans, E.F. and Wilson, J.P. (eds), Academic Press (1977).

9. Zwicker, E. and Schorn, K. Psychoacoustical tuning curves in audiology., Audiology, 17, 120-140 (1978).

10. Nelson, D.A. and Turner, C.W. Decay of masking and frequency resolution in sensorineural and hearing-impaired listeners., in Psychoacoustical, Physiological and Behavioral studies in Hearing., Van den Brink and Bilsen (eds), Delft University Press (1981).

11. Glasberg, B.R. and Moore, B.C.J. Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments., J. Acoust. Soc. Am., 79, 1020-1033 (1986).

12. Turner, C.W., Holte, L.A. and Relkin, E.R. Auditory filtering and the discrimination of spectral shapes by normal and hearing-impaired subjects., J. Rehab. Res. Devel., 24 (1987).

13. Bacon, S.P. and Brandt, J.F. Auditory processing of vowels by normal-hearing and hearing-impaired listeners., J. Speech Hear. Res.,25, 339-347 (1982).

14. Sidwell, A. and Summerfield, Q. The effect of enhanced spectral contrast on the internal representation of vowel-shaped noise., J. Acoust. Soc. Am., 78, 495-506 (1985).

15. Van Tasell, D.J., Fabry, A.A. and Thibodeau Vowel identification and vowel masking patterns of hearing-impaired subjects., J. Acoust. Soc. Am., in press (1987).

16. Vershuure, J. Pulsation patterns and nonlinearity of auditory tuning I. Psychophysical results., Acustica, 49, 288-306 (1981).

17. Turner, C.W. and Holte, L.A. Discrimination of spectral-peak amplitude by normal and hearing-impaired subjects., J. Acoust. Soc. Am., 81, 445-451 (1987).

CRITICAL BANDS
IN THE PERCEPTION OF SPEECH SIGNALS
BY NORMAL AND SENSORINEURAL HEARING LOSS LISTENERS

ROBERT D. CELMER

Acoustics Program and Laboratory
Department of Mechanical Engineering
University of Hartford
West Hartford, CT, USA 06117

GORDON R. BIENVENUE

Communications Department
State University of New York
College at New Paltz
New Paltz, NY, USA 12561

ABSTRACT

The results of a novel speech intelligibility test have been correlated to a tonal complex test, estimating critical bandwidth for normal and sensorineural hearing impaired listeners.

INTRODUCTION

Existing auditory theory suggests a major role for critical bands. Scharf [1] has defined the critical band empirically as "...that bandwidth at which subjective responses rather abruptly change...". In general, two stimuli separated in frequency by less than a critical bandwidth will interact in one of a number of ways, while two stimuli separated by more than a critical bandwidth will not. The critical band phenomenon has been observed in such perceptual phenomena as masking [1,2], loudness [3], and musical consonance [4].
Speech is the most pervasive and significant acoustic stimulus for the human listener, and evidence suggests that the critical band may serve in the analysis of speech [1]. In several studies, bandwidths contributing equally to the perception of speech were approximately equal to critical bands found in pure tone psychoacoustic studies [5,6,7,8,9].
The works of Fletcher [2], Zwicker [3], and Greenwood [10] imply that the critical band serves to band-limit background noise. The narrower the passband of the ear as a filter, the more noise the ear can reject. Thus, a listener may be able to correctly perceive a spoken communication despite background noise simply because much of the energy associated with the noise lies outside the critical bands surrounding the formant frequencies of the speech.
Discrimination of the formant and harmonic content of both speech and non-speech signals requires that these components be separated by at least one critical band [4,11,12]. Synthetic vowels presented to listeners by Remez [13] showed an abrupt changeover from speech-like to non-speech-like sounds as the formant bandwidth increased to greater than a critical bandwidth.

Several researchers have reported evidence of distorted or widened critical bands in subjects with sensorineural hearing loss [14,15,16,17,18,21]. In addition, these data demonstrate that the width of the widened critical band is independent of the magnitude of threshold hearing loss amongst those sensorineurals with critical bandwidth distortion.
The purpose of the present study was to directly test the hypothesis that the critical band is an essential element in auditory speech discrimination.

METHODS AND MATERIALS

A. Subjects

Forty-eight normal hearing listeners, aged 19 to 31 years, and sixty-eight sensorineural hearing impaired listeners, aged 18 to 67 years, participated in the study. Each subject was classified using conventional audiometric techniques. The right ear was the test ear for all subjects.
In addition, subjects' critical bandwidths were independently measured using a loudness of complexes test [1]. The center frequencies of each tonal complex were located at 700 Hz, 1000 Hz, 1600 Hz and 2150 Hz. For each trial a sub-critical tonal complex was presented to the listener and the bandwidth of the tonal complex was increased in time by small amounts. Ten repetitions of each center frequency were performed at a presentation level of 50 dB HL. Subjects were asked to listen to each test signal episode, and indicate the moment they perceived a change in the stimulus. The signal bandwidth required to elicit a perceptual change was recorded for each of the trials for each subject.

B. Taped Stimulus Materials

The stimulus materials used in this study were a set of limited resolution bandwidth speech signals. Prerecorded stimuli (in analog form) Were digitized by a 12-bit A/D converter at 20,000 samples per second, and stored in a computer. The software program separates the signals into 13 ms segments. The average spectrum of each segment is computed by first applying a Hamming window and then performing a Fast Fourier Transformation. The resolution bandwidth of this discrete spectrum is then limited by the software program. The frequency limits used for the normal critical bandwidth condition were those recommended by Scharf [1]. The discrete frequency amplitudes that fell within each bandwidth are averaged; each of the discrete amplitudes of that band are then set equal to this r.m.s. value, limiting the resolution allowed to the preselected bandwidth. See Figure 1. Coarser and narrower filtering schemes were realized by multiplying the bandwidth limits used by a chosen factor (retaining the original center frequency), and averaging the amplitudes contained within these widened limits. Each processed spectrum is then inverse transformed into the time domain, an inverse Hamming window is applied, and the output is converted to analog form via a D/A converter and recorded on audio tape.

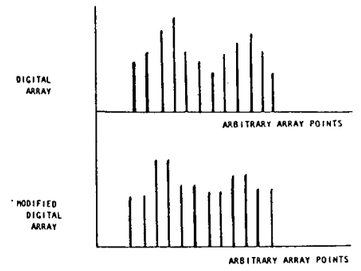


Figure 1. Effect of Digital Filtering

The NU#6 word list, a clinical audiometric word list, was used as the input audio material for the speech processing algorithm, since it includes (consonant-consonant-nucleus-consonant) sounds as opposed to only (consonant-consonant-vowel-consonant) sounds [19].
The tapes generated have seven frequency resolutions: an unprocessed list (UP); a bandwidth equal to one-half the resolution of the normal critical bandwidth (HX) [1]; a bandwidth equal to the resolution of the normal critical bandwidth (1X); two times normal (2X); three times normal (3X); five times normal (5X); and seven times normal (7X).

C. Equipment

The processed signals described above were generated using a hybrid computer system comprised of an EAI (Electronic Associates Incorporated) Model 680 analog computer interfaced with a DEC (Digital Equipment Corporation) digital computer, Model PDP-10. The audio output was recorded via a Crown Model BP824 tape deck.
The discrimination tasks were performed using an Ampex AG-440B tape recorder, a Maico Model MA-18 audiometer calibrated to ANSI 1969 standards, and TDH-39 earphones fitted with MX-41/AR cushions. The tests were performed in a Suttle Corporation Model B1 quiet room.

D. Test Procedure

After the audiometric threshold test and the independent critical bandwidth test, the subjects were presented with the seven fifty-word lists in the following resolution order: 7X, 5X, 3X, 2X, 1X, HX, UP. This sequence was chosen in order to minimize learning effects. The signal reached the earphone at a level of 50 dB HL and a signal-to-noise ratio of +10 dB. Pink noise was utilized as the masking source. Masking was used to minimize ceiling effects. Subjects were provided with answer sheets and were asked to write down the word they felt was said, guessing when necessary. Word lists were scored on a percentage basis.

RESULTS

A. Normal Hearing Subjects

The normal hearing group's mean pure-tone thresholds ranged from 1.5 to 5.6 dB for the frequencies from 500 Hz, to 4000 Hz. On the loudness-of-complexes test for independent critical bandwidth measurement, this group showed a mean critical bandwidth of 0.94 times the normal critical band as reported by Scharf; that is, 0.94X. The range of values was from 0.7X to 1.2X over the four test frequencies measured, confirming that this group did indeed have normal critical bands.
The plotted means for processed speech intelligibility scores are presented in Figure 2. A regression line was computed for the 2X through 7X portion of the data. The bandwidth condition vs the group's speech intelligibility scores exhibited a significant correlation of r = -.75 with a probability of error less than .01.

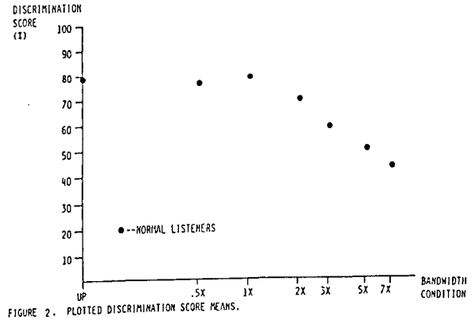


FIGURE 2. PLOTTED DISCRIMINATION SCORE MEANS.

As may be seen from Figure 2, the discrimination score was a direct function of the log of bandwidth resolution, for

those bandwidths wider than one critical band, but was independent of bandwidth for those bandwidth resolution conditions equal to or narrower than the critical band. The average discrimination score for the UP, HX and 1X condition was 77.56%. Thus, the masker did eliminate a ceiling effect.
A single factor analysis of variance indicated a significant main effect of the speech processing. A Newman-Keuls follow-up test [20] demonstrated a significant decreasing trend for discrimination scores as the bandwidths were varied from 2X through 7X. However, no significant decreases in intelligibility scores were observed for the UP (unprocessed) through 1X conditions. In addition, the group scores at the 1X condition were significantly higher than those at the 2X condition.

B. Sensorineural Hearing Impaired Subjects

The sensorineural hearing impaired group's mean pure-tone threshold audiogram ranged from 22.4 dB at 250 Hz to 71.2 dB at 8000 Hz (See Table 1). The mean speech reception threshold was 37.6 dB and the mean speech intelligibility in quiet at 50 dB HL was 74.7 %. On the loudness of complexes test for independent critical bandwidth measurement, this group showed a range of widened critical bands, from 1.1X to 4.23X, with a mean bandwidth of 2.43X. The subjects fell into four critical bandwidth groups: a 1X group (range of bandwidths 1.1X to 1.5X), a 2X group (1.5X to 2.5X range), a 3X group (2.5X to 3.5X range), and a 4X group (3.5X to 4.23X range).

Table 1.

MEAN AUDIOMETRIC DATA FOR HEARING IMPAIRED LISTENERS

f (Hz)	250	500	1000	2000	4000	8000
TL (dB)	22.4	29.3	38.3	60.2	65.6	71.2

The relationship between the critical bandwidth rating determined from the loudness-of-complexes test and the "knee" of the processed speech intelligibility data was of particular interest in the results of the study. This "knee" continued to represent the processing bandwidth at which the intelligibility scores dropped significantly from the unprocessed score. See Figure 3. These observations were made by performing a similar statistical analysis as used on the normal data.
It is interesting to note that the mean speech intelligibility functions for the subjects in the present study show a "knee" that coincides with their critical bandwidth test. This is especially easy to observe when the data are grouped according to the result of the loudness-of complexes test. The correlation coefficient between the loudness-of-complexes bandwidth measurement and the "knee" of the speech intelligibility function was 0.8749. This is a high correlation and indicates a very strong relationship between tonal estimates of the critical band and speech intelligibility of sensorineural, hearing impaired listeners.

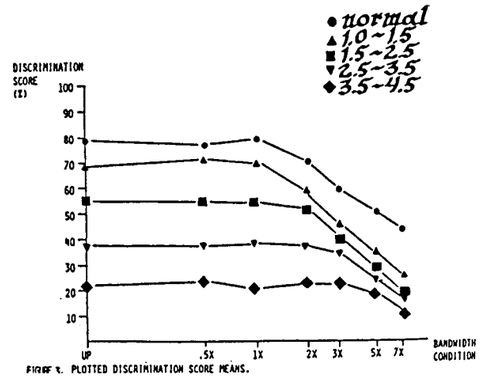


FIGURE 3. PLOTTED DISCRIMINATION SCORE MEANS.

DISCUSSION AND CONCLUSIONS

As noted above, the peripheral auditory system has been described to perform a preliminary frequency analysis of incoming acoustic signals. The limit to which frequency information may be gated is called the critical band and has been observed in a variety of psychoacoustic contexts. In particular, the critical band mechanism performs noise-band limiting and harmonic discrimination, both of which are crucial for the correct perception of such complex acoustic stimuli as speech. Thus, it was hypothesized that the critical band is a contributing factor to normal auditory speech intelligibility.
The purpose of the present study was to test the hypothesis that the critical band is an essential element in the process of speech listening. The performance of normal and sensorineural hearing impaired listeners with the processed speech test indicated the same three distinct trends:

1) A plateau effect was noted for the processing conditions UP through the condition equal to their independent critical band rating. For example, the group whose critical band was measured independently as 2X had a plateau effect occur from UP to 2X.

2) The score of the highest processing condition in the plateau was significantly higher than the next higher processing condition score.

3) A monotonic, decreasing trend was observed as the allowed bandwidth resolution was widened beyond the "knee" in the curve.

The scores above the observed "knee" in the curve demonstrated a close approximation to a logarithmic curve with a negative slope. Also significant is the correlation between the independent critical bandwidth rating and the location of the "knee" in the curve. In other words, the independent critical bandwidth rating corresponded to the observed point of inflection demonstrated by the speech intelligibility scores. This high correlation indicates a very strong relationship between tonal estimates of the critical band and speech intelligibility of normal and sensorineural hearing impaired listeners. It is reasonable to conclude that the integrity of the critical band is an important factor in the understanding of speech signals.
In summary, a correlation has been demonstrated between the performance of listeners on an auditory speech intelligibility test and a test with tonal complexes as the stimuli. The tonal complex test results yielded a bandwidth resolution value that was correlated to the probable point of inflection of the auditory speech intelligibility test results. Existing literature has implied, but has not clearly demonstrated the presence of such a correlation. The authors suggest that any consideration of auditory, speech intelligibility among normal or sensorineural, hearing impaired listeners must include an examination of the integrity of the critical band phenomenon in the subject population.

REFERENCES

[16] Bennett, T., and others. "Procedures for Characterizing Certain Effects of Prolonged Noise Exposure." Journal of the Acoustical Society of America (JASA), 63, Supplement 1, 1978.

[15] Bienvenue, G., and P. Michael. "The Temporary Effects of Short Term Noise Exposure on Masking Phenomenon." Unpublished EAL Research Project, 1977.

Bienvenue, G., and others. "Digital Processing of Speech Materials in the Study of Sensorineural Hearing Impairment." JASA, 67, Supplement 1, 1980.

[22] Bonding, P. "Critical Bandwidth in Presbycusis." Scandinavian Audiology, 8, 1979.

[8] Castle, W. "Effects of Selective Narrow Band Filtering on Perception of PB Word Lists." JASA, 36, 1964.

[9] Chari, N. "Perception of 1/3 Octave-filtered Speech." JASA, 61, 1977.

[21] deBoer, E., and Bouwmeester, J., "Critical Bands and Sensorineural Hearing Loss," Audiology, 13, 236-259, 1974.

[2] Fletcher, H. "Auditory Patterns." Review of Medical Physiology, 12, 1940.
[5] French, N., and J. Steinberg. "Factors Governing the Intelligibility of Speech Sounds." JASA, 19, 1947.

[10] Greenwood, D. "Auditory Masking and the Critical Band." JASA, 33, 1961.

[12] Haggard, M. "Feasibility of Rapid Critical Bandwidth Measurements." JASA, 55, 1974.

[7] Kryter, K. "Speech Bandwidth Compression through Spectrum Selection." JASA, 32, 1960.

[14] Michael, P., and G. Bienvenue. "A Procedure for the Early Detection of Noise Susceptible Individuals." American Industrial Hygiene Association Journal, 1976.

[11] Morton, J. and A. Carpenter. "Experiments Relating to the Perception of Formants." JASA, 35, 1963.

[4] Plomp, R. "The Ear as a Frequency Analyzer." JASA, 36, 1964.

Rabiner, L. and R. Schafer., Digital Processing of Speech Signals, ed. A. Oppenheim, (Englewood Cliffs: Prentice-Hall, 1978).

[13] Remez, R. "Adaption of the Category Boundary Between Speech and Nonspeech: A Case Against Feature Detectors." Haskins Laboratories: Status Report on Speech Research, SR-50, 1977.

[6] Richards, D. and R. Archbald. "A Development of the Collard Principle of Articulation Calculation." Proceedings of the IEE, 103B, 1956.

[1] Scharf, B. "Critical Bands." Foundations of Modern Auditory Theory, ed. J. Tobias, (New York: Academic Press, 1970).

[19] Tillman, T., and R. Carhart. "An Expanded Test for Speech Discrimination Utilizing CNC Monosyllabic Words (Northwestern University Auditory Test No. 6)." Technical Report, SAM-TR-66-55, United States Air Force School of Aerospace Medicine, Aerospace Medical Division (AFSC), Brooks Air Force Base, Texas, 1966.

[3] Zwicker, E. "Uber Psychologische und Methodische Grundlagen der Lautheit." Acustica, 8, 1958.

SPEECH CUE ENHANCEMENT FOR THE HEARING IMPAIRED:
III. AMPLIFICATION OF FRICATION FOR IMPROVED PERCEPTION OF FINAL FRICATIVE VOICING

REVOILE, S.G., HOLDEN-PITT, L., EDWARD, D., and PICKETT, J.M.

Center for Auditory and Speech Sciences
Gallaudet Research Institute
Gallaudet University

ABSTRACT

For 20 severely/profoundly hearing-impaired listeners, voicing perception for final fricatives was tested using spoken syllables with and without enhanced frication. The enhancements involved filtering and amplification of /f/, /s/, and, of iterations of pitch periods in the vowel offset to replace /v/, /z/. Nearly 3/4 of the listeners showed considerable improvement in perception for the fricatives with enhancement. Discrimination training for voicing cognate syllables contributed to the improvement in perception found for the enhanced fricatives. For syllables with natural fricatives, lower performance and less improvement was seen as a result of training. Inaudibility of the enhanced fricatives seemed to explain the poor performances of listeners who showed no improvement in voicing perception for the enhanced fricatives.

INTRODUCTION

Previously, we examined the effects of cue degradation on final consonant voicing perception by moderately/severely hearing-impaired listeners (Revoile et al., 1982[1] and 1985[2]). Elimination of the vowel duration cue and/or the consonant constriction cues (i.e., frication, stop bursts, presence of voiced murmur) degraded perception for most of the hearing-impaired listeners studied. We then began to investigate whether enhancement or exaggeration of cues to final consonant voicing could yield improved perception for hearing-impaired persons who typically manifest reduced ability to distinguish consonant voicing (Revoile et al., 1987[3], 1986[4] & 1986[5]). This paper describes an experiment on amplification and filtering of final fricative consonants as a means of facilitating fricative voicing distinctions by hearing-impaired listeners.

METHOD

The listeners were 20 hearing-impaired Gallaudet undergraduates who had predominantly profound losses. Their threshold averages for .5, 1, and 2 kHz (3FA) ranged from 80 to 108 dB HL, with a mean ($\bar{X}$) of 94 dB HL. The listeners were selected to have 3FA of at least 80 dB HL because our previous findings suggested that reduced consonant voicing perception would be concomitant with such impairments (Revoile et al., 1982[1], 1985[2]).

Stimuli. The syllables /bæ s, bæ z, bæ f, bæ v/ served as the stimuli to test fricative voicing perception. For most of the experiment, 4 different syllable utterances (female talker) representing each fricative were used in a test block (16 different utterances). These tokens had been selected from a larger pool of utterances to differ minimally in vowel duration between voicing cognate syllables. Some mean acoustical characteristics of the test utterances are shown in Table 1.

The phoneme segment durations were measured visually on waveforms of the utterances; rms intensity was measured for the duration of each segment. The vowel/frication boundary was identified on the waveform as the point at which periodic vowel oscillation ceased. The utterances had been digitized (16.67 kHz sampling rate) for these measurements and for processing of the frication enhancements. Further details of the recording and measurement procedures can be found in Revoile et al. (1986[4]).

The frication segments of each utterance were altered to generate the enhanced stimuli. The /f/ and /s/ fricatives were low-pass filtered (5 kHz) and then amplified by 21 dB. The /v/ and /z/ fricatives were deleted and replaced by 4 iterations of 2 to 4 pitch periods copied from the end of the vowel in each utterance. These pitch periods contained some f_o as well as consonant friction noise. The pitch periods were iterated to yield frication durations that nominally matched those of the deleted fricatives. The segments were band-pass filtered (.25 to 1 kHz) to reduce the presence of the vowel f_o and intensified by 18 dB. In the enhanced syllables (Table 1), note that the fricatives had been amplified to approximate the level of the vowels.

Procedures. The testing and training of fricative voicing perception was carried out in an experiment on cue enhancement for final consonant voicing that lasted about 2-1/2 months. During that period, the listeners participated in twice weekly sessions of 50 minutes each. The initial 3 to 4 sessions of the experiment were devoted to screening tests for perception of final consonant voicing, to insure that the students chosen as listeners manifested reduced voicing perception of final fricatives. The training segment of the experiment began with a

TABLE 1. A summary of some acoustical characteristics from 4 utterances each of /bæ f/, /bæ s/, /bæ v/, /bæ z/.

		Duration, ms		Intensity, dB[1]		
					Frication	
		Vowel	Frication	Vowel	Unaltered	Enhanced[2]
/bæ f/ & /bæ s/	$\bar{X}$	289.7	211.3	49.1	27.9	47.8
	S.D.	(8.4)	(18.5)	(1.2)	(2.4)	(0.4)
/bæ v/ & /bæ z/	$\bar{X}$	291.6	114.0	49.4	28.2	49.9
	S.D.	(9.9)	(14.2)	(1.1)	(1.7)	(1.5)

[1] Relative to an arbitrary reference.
[2] Intensity measurements for the enhanced voiced fricatives were of iterations of pitch periods from the vowel offset, which were substituted for the natural /v/ and /z/ frications.

baseline assessment of the syllables both unaltered and with frication enhancement. Training for the syllables with enhanced fricatives was usually followed by training for the unaltered syllables. All stimuli were presented to each listener's better ear at a most comfortable level (MCL) established by an adaptive procedure at the beginning of each session.

Syllable identification trials were used to test fricative voicing perception throughout the experiment. The response buttons used by the listeners were labeled: BAFF, BASS, BAV, BAZZ, and BA. No feedback of correct answers was given.

The unaltered and enhanced utterances were tested in separate blocks of trials. In each block of 48 trials, the 4 different utterances per fricative were each presented 3 times in random order. These syllable blocks were tested for the baseline measurements of fricative voicing perception, and before and after training.

The training of fricative voicing perception involved discrimination trials of voicing cognate syllables with feedback following each trial. In separate blocks, two types of discrimination trials were used--an "oddity" procedure and a "paired-comparison" procedure. The "oddity" procedure presented syllable sequences, such as BASS BAZZ BASS or BAZZ BASS BAZZ in separate trials. The listener chose which sequence had been presented. The "paired-comparison" procedure presented syllables pairs such as BASS BASS, BAZZ BASS, BAZZ BAZZ, or BASS BAZZ and the listener selected the pair heard.

The unaltered versus enhanced syllables were used in separate training sessions. A training session began with assessment of syllable identification. Subsequently the "oddity" followed by "paired comparison" discrimination procedures were presented for each of the syllable sets representing /s/-/z/ and /f/-/v/. Finally, syllable identification was retested at the end of the session.

Most of the the listeners were administered 2 to 6 training sessions ($\bar{X}$=3.8) for the syllables with enhanced fricatives. Typically, listeners who received fewer training sessions were those who evidenced chance performance after training for the fricative-enhanced utterances. Because of their poor performances, these listeners were not trained for the unaltered syllables, to limit their frustration with the training. However, all listeners who showed at least some ability to perceive voicing after training for the enhanced fricatives (>60%) received 1 to 3 training sessions ($\bar{X}$=1.7) for the unaltered syllables.

An additional discrimination procedure, "frication presence", was used to determine whether the frications were audible to the listeners. The procedure required discrimination of an utterance-with-frication, versus that same utterance with frication deleted. The trials used were 3IFC, with the fricationless utterance presented twice and the full utterance once. The listener chose which interval contained the full utterance. Feedback of the correct answer was presented after each trial.

RESULTS

The performance of each listener for each fricative was scored according to percent correct voicing perception (errors in place perception ignored) for each block of syllables presented. A total percent correct score was computed for each test condition (i.e., screening, baseline, etc.) by averaging scores across fricatives and repetitions of stimulus blocks. Much of the data represented chance level performance. When such performance was seen among the listeners overall, no statistical analyses were carried out.

The listeners were assigned to one of 2 groups depending upon whether their average performances were above chance level (at least 60%) for voicing perception on the tests administered after training for the enhanced fricatives. The

6 listeners who scored below this criterion formed the poor group; their mean (X) 3FA was 98 dB HL. The other 14 listeners (X 3FA = 91 dB HL) composed the good group. Mean fricative voicing perception for the groups on each test condition is summarized in Table 2.

Performance for unaltered syllables. The good and poor listener groups can be compared for performance on the unaltered syllables when these stimuli were administered for the screening and for the baseline tests. For both of these tests, similar performances occurred between the groups; mean scores of less than 50% were obtained. These chance level performances indicate that the voicing cues for the natural fricatives were generally not utilized by the listeners, at least prior to training. (Recall that the vowel duration cue had been reduced by pre-selection of voicing cognate utterances having minimal value for this cue.)

For the good group, training for the unaltered syllables resulted in some improvement in voicing perception for the fricatives. The group's scores after training for the unaltered syllables were significantly higher than before [t-test for paired samples: $t(13)=3.9$, $p<.01$], although by a mean difference of just 8%. Because only 1 to 3 sessions of training were used for the unaltered syllables, it's possible that the maximum improvement in fricative voicing perception was not achieved. Lack of familiarity with the training procedure was probably not a factor in limiting the improvement seen for the unaltered syllables because their training was subsequent to that for the enhanced syllables. Nevertheless, a longer period of training might have heightened the listeners' awareness of the natural fricative voicing cues in the unaltered syllables.

Performance for syllables with frication enhancement. When the syllables with enhanced fricatives were initially presented, neither group of listeners performed well. The baseline tests for these syllables, administered just before training, yielded performance at chance level for both groups.

The training for the fricative-enhanced syllables resulted in large performance differences between the listener groups, both before and after training. The poor group performed at chance level for the syllables with enhanced fricatives, regardless of training, while the good group scored at least 20% above chance. An analysis of variance was carried out using as factors: group (good versus poor) by before/after (treated as a repeated measure). Highly significant differences were found between the good and poor groups for performance with the fricative-enhanced syllables during training [$F(1,18)=45.9$, $p<.001$].

An interaction appeared between the factors: group and before/after [$F(1,18)=5.6$, $p=.03$], indicating that the amount of performance change occurring before versus after training was different between the listener groups. Further analyses of simple main effects revealed that the scores after training were significantly greater than before training for the good group [$F(1,18)=27.7$, $p<.001$], while the poor group showed no difference in scores before versus after training [$F(1,18)=.39$, $p=.54$].

The results for the discrimination of frication presence were examined to determine whether the groups' performance differences for the enhanced syllables might be associated with the general audibility of the enhanced frications. Between utterances with enhanced frication versus those utterances fricationless, the good group showed 87% mean discrimination and the poor group, 51%. While the poor group scored somewhat above chance (33%), their reduced level of performance indicates that they were unable to discriminate the presence versus absence of the enhanced frications for the majority of the utterances. This reveals that these listeners' inability to perceive voicing for the enhanced fricatives was generally a result of frication inaudibility. For the unaltered syllables, the mean discrimination score for presence-vs-absence of the natural frications was 51% for the good group and 28% for the poor group.

Unaltered versus enhanced training effects. For the good group of listeners, scores could be compared between the unaltered and enhanced syllables for training of fricative voicing perception. Overall, performance with the enhanced syllables throughout training was markedly better than for the unaltered syllables [$F(1,13)=61.2$, $p<.001$]. However, the effects of training were similar for the unaltered and enhanced syllables as indicated by the absence of interaction for scores representing the two types of syllables before and after training [$F(1,13)=.31$, $p=.59$].

It was of interest to examine whether the level of perceptibility differed between /f/, /s/ versus /v/, /z/, especially because of the disparate spectral characteristics of enhanced /v/, /z/. However, this seemed to have no effect on the listeners voicing perception for the fricatives, as scores for /f/, /s/ versus /v/, /z/ were similar [$F(1,13)=1.4$, $p=.25$].

TABLE 2. For unaltered and enhanced final /f/, /s/, /v/, /z/ in /b C/ syllables, mean percent correct voicing perception by two groups of hearing-impaired listeners for various tests during the experiment.

		Unaltered Syllables				Syllables with Enhanced fricatives		
		Screening Tests[a]	Baseline	Training		Baseline	Training	
				Before	After		Before	After
Good Group (n=14)	X (S.D.)	46.0 (6.2)	47.0 (8.5)	54.1 (7.7)	62.3 (8.7)	50.0 (15.1)	73.0 (10.9)	82.6 (8.0)
Poor Group (n=6)	X (S.D.)	45.5 (9.9)	46.0 (11.6)			54.0 (7.9)	50.2 (6.9)	52.0 (5.1)

[a] Scores based on 40-utterance block. Dashes are inserted where no tests were administered.

DISCUSSION

This study revealed that frication filtering and amplification yielded improved perception of voicing for syllable-final fricatives by almost 3/4 of the severely/profoundly hearing-impaired subjects tested. The improvement in perception occurred for enhanced fricatives that were amplified to the level of the preceding vowels in /bæ C/ syllables. When the frications were at natural levels in the syllables, i.e., -21 dB re the vowels, the listeners' performances were somewhat above chance level, but only after discrimination training for voicing cognate syllables. For the fricative-enhanced syllables, the discrimination training facilitated the listeners' improved perception of fricative voicing.

The remaining listeners, about 1/4 of the total group, showed no improvement in voicing perception for the fricative-enhanced syllables, even after training. For these listeners, most of the enhanced frications were probably inaudible. A discrimination test for presence/absence of the frications in the test syllables revealed that these listeners were unable to distinguish most of the test utterances when presented with versus without enhanced frications. Further evidence of the inaudibility of the enhanced fricatives is apparent from the syllable presentation levels used relative to the tone sensitivity of these listeners. Based on the listening levels chosen by these subjects ($\bar{X}$ MCL = 113 dB SPL), the vowels were presented at sensation levels (SL) of less than 15 dB. The SLs of the enhanced frications were probably lower, due to the listeners' greater hearing loss in the frequency regions of the frication spectra compared to the vowel spectra.

The amount of improvement in final consonant voicing perception effected by the enhanced fricatives in this study is somewhat less than that seen for amplified stop bursts (Revoile, et al., 1987[3]) and enhancement of the vowel duration cue (Revoile, et al., 1986[4]). In future studies, we will investigate the relative effects of single versus multiple cue enhancements in search of the maximum improvement for consonant recognition by hearing-impaired listeners.

CONCLUSIONS

Some severely/profoundly hearing-impaired listeners performing at chance level for fricative voicing perception may obtain improved performance as a result of frication filtering and amplification to a level comparable to that of the preceding vowel. Discrimination training for syllables with enhanced voiceless versus voiced fricatives seems important in effecting the improved perception of fricative voicing.

REFERENCES

[1] Revoile, S., Pickett, J. M., Holden, L.D., and Talkin, D. (1982). Acoustic cues to final stop voicing for impaired-and normal-hearing listeners. J. Acoust. Soc. Am. 72, 1145-1154.

[2] Revoile, S., Holden-Pitt, L., and Pickett, J. M. (1985). Perceptual cues to the voiced-voiceless distinction of final fricatives for listeners with impaired or normal hearing. J. Acoust. Soc. Am., 77, 1263-1265.

[3] Revoile, S., Pickett, J., Holden-Pitt, L., Edward, D., Pickett, J. and Brandt, F. (1987, in review). Speech-cue enhancement for the hearing-impaired: II. Amplification of burst/murmur cues for improved perception of final stop voicing. J. Rehab. Res. Dev.

[4] Revoile, S., Holden-Pitt, L., Pickett, J., and Brandt, F. (1986) Speech cue enhancement for the hearing impaired: I. Altered vowel durations for perception of final fricative voicing. J. Speech Hear. Res., 29, 240-255.

[5] Revoile, S., Holden-Pitt, L., Edward, D., and Pickett, J. (1986) Some rehabilitative considerations for future speech-processing hearing aids. J Rehab. Res. Dev., 23, 89-94.

SPEECH PATTERN ACQUISITION IN PROFOUNDLY HEARING IMPAIRED CHILDREN

EVELYN ABBERTON VALERIE HAZAN ADRIAN FOURCIN

Department of Phonetics and Linguistics
University College London
Gower Street, London, WC1E 6BT

It is well known that even profoundly hearing-impaired children can, in particular instances, acquire highly intelligible speech and excellent conversational skills if educated in an appropriate speaking-hearing environment. The detailed course of their speech development has not, however, been studied. In this paper we outline some of our first findings from a long-term study of the perceptual and productive phonological development of an unselected group of 17 severely-profoundly hearing-impaired children, a whole class, who were aged between 7 and 8 at the start of the project in October 1984. All the children were pre-lingually hearing-impaired and had been fitted with hearing-aids at an early age. Their pure tone audiogram average in the better ear at .5, 1 and 2 KHz varied between 83 and 115 dB. A principal result of the work indicates the extent to which, during the first two years, the children follow an essentially normal pattern of development albeit with a delay which, for some speech contrasts may be great.

Normal phonological development

Observation of the phonology of normally hearing children over the first few years of life shows the development of the ability to control the larynx and vocal tract to produce low frequency, high intensity simple acoustic speech patterns before rapidly changing, high frequency complex elements and patterns. Thus we see the development of control of adequate phonation and fundamental frequency changes in the first year of life, enabling the intonational foundations of conversation to be laid. The contrastive vowel system of the child's language is acquired during the second and third years, but a fully contrastive system of consonants is often not present until 5 - 7 years of age. Phonetic accuracy and stability take several more years. From both acoustic/auditory and articulatory points of view there is a progression from simple to complex, and, within the consonant system a stable voiced-voiceless contrast is slow to mature, and sibilant fricatives and velar consonants are acquired relatively late. [1, 2] Two phonological contrasts from the long-term study have been chosen for discussion here. The first is a vowel contrast, that between [ɜ] and [ɑ], phonemically adjacent but acoustically distinct; and the second is the voicing contrast, salient in itself, between the complex velar plosives [k] and [g]. Children would be expected to acquire the vowel contrast before the voicing contrast, and not to produce the contrasts consistently unless they were perceived contrastively.

Speech Perception Tests: Procedure

The two chosen minimal pairs - [ɜ]-[ɑ] and "goat-coat" - were synthesised on a hardware parallel synthesiser to closely match natural tokens spoken by a female speaker known to the subjects. 6-token continua were then built, in which the most salient parameters marking the contrast, F1 and F2 in the vowel contrast, Voice Onset Time and F1 onset in the voicing contrast, were altered in equal steps, all other speech patterns remaining constant.

The vocalic minimal pair consisted of two-formant syntheses, with F1 varying from 500 Hz in the extreme [ɜ] to 900Hz in the extreme [ɑ] and F2 varying between 1666 Hz in [ɜ] and 1400 Hz in [ɑ]. Stimuli were synthesised with a falling intonation (initial and final values: 237 Hz-155 Hz). A "goat-coat" minimal pair was used to investigate the voicing contrast. The two patterns under investigation were: Voice Onset Time, varying from 20 ms in "goat" to 100 ms in "coat", and the onset frequency of the first formant (F1). A more detailed description of the stimuli can be found in Hazan and Fourcin [3].

The stimuli were presented using a forced-choice identification test procedure. A computer-controlled adaptive testing procedure was used so that the length of the test and degree of complexity were dependent on the subject's ability. The outcome of the test was an identification or "labelling" curve in which the percentage of responses of one label were plotted across the stimulus range. Labelling curve configuration enables one to assess not only the ability to identify the extremes of the range, which are well differentiated, but also to consistently group similar stimuli into a same category. It therefore provides a controlled measure of the way in which subjects deal with the inherent variability of speech.

Simon and Fourcin [4] described three main types of labelling curve configuration which are found as a subject's labelling ability progresses. At first, random labelling may be obtained, where random responses occur even to the extremes of the stimulus range. As development proceeds, progressive labelling is seen. Here, extremes are labelled appropriately but inconsistent labelling is given to intermediate stimuli. Finally categorical labelling is obtained, where the intermediate stimuli are consistently labelled into two clear categories.

On average, three assessments per year were obtained, and labelling tests were presented to the subjects at each testing session. Stimuli were presented free-field, from a loudspeaker situated in front of the subject at a distance of about a metre. The subjects were wearing their own hearing aids. This method of presentation was chosen to best match the children's listening experience.

A detailed examination of labelling function configuration over time offers great insights into the children's stage and rate of development, as has been seen in previous studies [4, 3]. In this overview, however, rather than examining the detailed development in labelling ability in the 17 subjects over a two year period, the general trend in development seen in their ability to label both the vowel and voicing contrast will be compared to the progress seen in their ability to appropriately produce these contrasts. The types of labelling curve configuration found at the beginning andend of the testing period for both contrasts are given in Table I. Subjects are ranked in terms of increasing pure tone threshold average (better ear average over .5,1,2 KHz).

Perception Tests - Results

General development is seen in most children over the two year period, most clearly visible in the children's increasing ability to label the vowel contrast. At the beginning of the testing period, 8 children out of the 17 tested were labelling the vowel contrast at random. Two years later, only one of the children was still giving random labelling to these stimuli. The rate of progress from random to categorical labelling was found to vary from child to child.

In terms of the order of development of contrasts, as in a previous study [3], it appears that the children were categorically labelling the vowel contrast before the more acoustically complex voicing contrast. Because of their short attention span, many of the subjects intially having most difficulty in the labelling of the vowel contrast were not tested on the voicing contrast, after initial live training tests had shown their inability to respond to the extremes of the range. Out of the 8 subjects who were tested on both contrasts at the beginning of the testing period, 2/8 give random labelling to the vowel contrast, while 6/8 gave random labelling to the voicing contrast. In March 1987, all children in this group could label the vowel contrast amd 1/8 was labelling the voicing contrast, at random 7/9 of the more impaired subjects not tested earlier were labelling the voicing contrast at random in March 1987.

Speech production assessment

Recordings have been made three times a year using both conventional microphone and laryngograph techniques. The laryngograph allows non-invasive monitoring of vocal fold activity during normal speech production, and information is thus available for subsequent computer analysis of a range of vocal fold parameters including fundamental frequency [5]. Each speech sample includes some spontaneous conversation, citation forms illustrating certain phonological contrasts, and a traditional picture naming articulation test. The [ɜ] - [ɑ] and [g] - [k] contrasts are elicited in citation form by showing pictures of 2 animals whose names are the vowels concerned (as in the perception tests) and pictures of a coat and a goat which the child labels. At the first recording session the vowels were elicited by imitation.

Discussion of Production

Mere tabulation of binary results inevitably simplifies a complex picture, and, to some extent, can obscure detailed findings which are relevant to a more comprehensive view of phonological development. One important factor is variation in the pronunciation of the test items and of other words in the children's vocabularies that contain the same segments. At the start of the study, in vowel production, the children with less hearing cannot produce the contrast by imitation or labelling but do have a few well known words that are phonetic approximations to adult forms, with [ɜ] and [ɑ] appropriately used, in their spontaneous speech, or produced in the picture naming traditional articulation test. These children are at an even earlier stage of consonant development, and [k] and [g] never appear. The children with better hearing, however, on the whole, have no trouble with the vowels in labelling or spontaneous speech, but show a similar picture for the velar consonants as the more impaired group show for the vowel contrast: inability to produce citation minimal pairs but some appropriate use of velars in a few lexical items. This behaviour could be considered pre-phonological as contrastiveness cannot be demonstrated; but the foundation for it is present as the sounds can be articulated and in some cases matched to the equivalent segments in adult pronunciations. Children at this stage of development may or may not show categorical labelling of the contrasts concerned.

In the speech development of normally hearing children this pre-phonological stage occurs at about eighteen months of age when the child has only a very small vocabulary of some 50 words. For the 7 and 8 year old hearing impaired children, at the start of this study the stage had, therefore, developed several years later - for the contrasts under investigation: not all aspects of phonology were delayed at this stage except for the most severely impaired speakers. The less handicapped children showed well-developed syntax and intelligible speech although with some immaturities such as fronting and stopping of fricatives [1].

In general, two years later a similar picture of vowel development preceding voicing is seen: the less hearing-impaired children who still do not have a [k] - [g] contrast nevertheless now use the

velar consonants in some words, and the same is true for the more impaired speakers' vowels. Although progress is thus very slow and delayed, it is, however, following the normal pattern.

Discussion and Conclusion

Evidence of age-related development in the ability to label increasingly complex contrasts is found in these profoundly hearing-impaired children almost without exception. This development seems to be following a normal course, in terms of the order in which the contrasts are acquired and the manner in which they are individually established. Although this development is delayed by at least four to five years compared to normally-hearing children, it is part of an ongoing progression.

Some variability is found in the stage of development reached by these subjects from a homogeneous age-group. This variability is not fully explained by their degree of residual hearing, although there appears to be a gross difference between more and less advanced children, with a division around an average better ear pure tone loss of 100 dB SL. The results of our work are surprising and encouraging in regard to the auditory speech skill development shown possible in a group of unselected profoundly hearing-impaired children. It is likely, however, that with the use of hearing aids designed explicitly to enable the maximum use of their residual hearing, even greater and earlier speech receptive development could occur.

Table 1 : *Results of Perception Tests*

Child	P,T.A.	Ranking PTA - Nov 1986 Vowel March 85	Voicing July 85	Vowel March 87	Voicing March 87
1	83	P	R	C	C
2	90	R	R	C	R
3	93	R	R	C	C
4	93	P	R	P	P
5	95	P	P	C	C
6	97	P	R	C	P
7	102	P	P	C	C
8	102	R	-	P	R
9	108	R	-	C	R
10	108	P	R	P	P
11	108	R	-	R	R
12	108	P	-	C	R
13	110	P	-	C	P
14	110	R	-	P	R
15	110	R	-	C	R
16	112	P	-	P	P
17	115	R	-	P	R

P = Progressive labelling = at least 87% correct at extremes
C = Categorical = at least 4 steps labelled with 100% consistency
R = Random

Table 2 : *Production of vowel and voicing contrasts in citation forms*

Child	P.T.A.	March 85 Vowels	March 85 Voicing	March 87 Vowels	March 87 Voicing
1	83	Y	Y*	Y	Y
2	90	N	N	Y	Y
3	93	N	N	Y*	N
4	93	Y	Y	Y	Y
5	95	Y	Y	Y	Y
6	97	Y	Y*	Y	N
7	102	N	Y*	Y	Y
8	102	N	N	Y	N
9	108	N	N	Y*	N
10	108	Y	N	Y*	N
11	108	N	N	N	N
12	108	Y*	N	Y*	N
13	110	Y	N	Y	N
14	110	N	N	N	N
15	110	N	N	N	N
16	112	N	N	Y*	N
17	115	N	N	Y*	N

Key: Y = Yes: child consistently produces the contrast
N = No: child produces only one segment or different sounds randomly
* = Clear contrast, consistently produced, but immature phonetically: velars are palatalised; [ɑ] centralised sometimes

Notes: Each contrast was produced at least twice.
Child 11 was recorded for the first time in July 85, and the most recent results for Child 3 are from October 86.

References

1. Abberton, E (1986) Diagnostic implications of phonological analysis? *New Zealand Speech-Language Therapists' Journal*, Vol XL, No 1, 2-16

2. Fourcin, A J (1978) Acoustic patterns and speech acquisition, in *The Development of Communication*. N Waterson and C Snow (Eds), John Wiley

3. Hazan, V and Fourcin, A J (1985) Microprocessor-controlled speech pattern audiometry: preliminary results. *Audiology* 24, 325-225

4. Simon, C and Fourcin, A J (1978) Cross-language study of speech pattern learning. *JASA*, 63, 925-235

5. Abberton, E (in press) Voice development in deaf children. *Proceedings of the 1st Asian Pacific Regional Conference on Deafness*, Hong Kong, December 1986

Acknowledgement
We thank the staff and children of the Birkdale School for Hearing Impaired Children for their willing cooperation at all times.

FROM SYLLABLES TO SENTENCES

YAEL FRANK

School of Education
Tel-Aviv University
Tel-Aviv, ISRAEL

Abstract

In this study lexical stress and intonation patterns of 11 normal hearing and 23 hearing-impaired 9-12 year old children were compared. The comparison was done by spectrographic analysis and Visipitch. It appears that the Hearing Impaired in this group succeeded in producing stress on the word level, but failed to make the necessary changes for intonation on the sentence level.

Prosodic features are of great linguistic importance as well as for the perception of naturalness in speech. The term 'naturalness' is of special meaning for the speech of the Hearing-Impaired because the lack or the distortion of some of the prosodic features contributes to the listener's difficulty in understanding the Hearing - Impaired's speech.

Teaching profoundly hearing impaired children the segmental aspects of speech requires much effort on behalf of the therapeutic team, and many children do not acquire enough skills in order to produce prosodic features. One could argue that most prosodic features could be learned without conscious effort, if the children were exposed to speech in their environment (e.g. hearing parents, siblings, schoolmates, etc.) because even many profoundly hearing impaired have enough hearing left in order to detect prosodic features, especially since most prosodic features are found in the frequencies lower than 500 Hz. However, this does not seem to be the case and logopedic and educational personnel have great difficulty in teaching deaf children to acquire those features.

In Israel speech therapists and teachers usually work without exact measurements and are thereby hampered in teaching systematically. And so every therapist "does his best" according to his own inclinations, and the knowledge that he has acquired from the literature in the field. Therefore therapists cannot be sure that their applied strategies are appropriate for the Hebrew-speaking population. This does little to further a consistent methodology in speech therapy.

The following research was done in order to compare some of the physical properties of stress and intonation between hearing and hearing-impaired Hebrew speaking children so we can be able, in the future, to supplement clinical intuition.

The research group consisted of 23 prelingually profoundly hearing impaired children, 9-12 years of age, who had the same oral education from preschool age and were mainstreamed in two classes in a regular school. Their mean hearing loss in the speech frequencies was 99.5 dB in the better ear. In the control group were 11 age-matched normal-hearing children.

The children read aloud 21 sentences (after previously rehearsing them) of varied length, consisting of declaratives, yes-no questions, wh-questions and imperatives. The sentences were composed of words known to be familiar to the children and their general gist was taken from everyday examples of spoken speech between mother and child, or of children between themselves. Before the reading we made sure that the children had learned punctuation. The sentences were hand-printed in bold letters. They were ordered on 3 tables, each new sentence at the beginning of a line.

Recording was done in a quiet room and a pleasant atmosphere (with two children always together in the room to minimalize the effect of new surroundings). Duration of vowels and their fundamental frequency were measured with the help of a Kay-Spectrograph and Visipitch. In order to gather enough information we produced broad- and narrow-band spectrograms and on top of the broad-band a relative amplitude curve. The duration of the vowels was measured between the recognized end or beginning of a consonant (the Hebrew syllable structure is CV or CVC). Fundamental frequency was measured at the vowel midpoint on the Visipitch-trace. The measured Fo was compared with the analog point on the narrow-band spectrogram. In this way, we obtained the principal physical properties of lexical stress [4,5,6].

Duration (measured in msec) of 192 vowels out of 204 was significantly different (according to a t' test) between Hearing and Hearing-Impaired. This means that nearly every word for the Hearing-Impaired needed more time. These results were predictable [2,7,8]. Fo varied significantly in the production of 105 vowels out of the 204 (with higher frequencies for the Hearing-Impaired). This result was also predictable [3].

For lexical stress we computed the relation between stressed and unstressed vowels in duration and in Fo. We took 33 two- and three- syllabic words from the beginning, the middle and the end of the 21 sentences. In contrast to the absolute duration which was longer in nearly all the vowels produced by the Hearing-Impaired, we did not find a statistical difference in the temporal relation unstressed/stressed vowel between the Hearing and the Hearing-Impaired in the words from the beginning and the middle of the sentences. Only in 5 words from the end of sentences did we find a significant difference. It seems that even the profoundly Hearing-Impaired perceive temporal relations.

But the results of the Fo relations between unstressed and stressed vowels showed (according to the t-test) a significant difference ($p<0.01$) in 17 out of the 33 words. This result means that there is a much greater gap in the ratio of Fo than in the ratio of duration.

More evidence of the learning of stress ratios by the Hearing-Impaired was found in a rather curious way. As it happens there were 6 words in the whole set that are mostly found in mother-child discourse in a special stress form [1]. The hearing children changed this special stress form while performing the formal task of reading. Five of the 23 Hearing-Impaired did the same while the others continued with the once-learned stress. This is an interesting phenomenon because it shows that the Hearing-Impaired have acquired a special stress for certain words through their mothers and through frequent use (these were words like gli'da (ice cream), u'ga (cake), shoko'lad (chocolate) and first names like U'ri, Mo'she, Mi'ki). However, only five of them also had the flexibility to change once-acquired patterns.

Continuing our study from syllables to word and then to the sentence level, we arrived at an explanation for the aforementioned facts. When connecting the mean of the measured values of duration and of Fo we see that the hearing children explicitly express intonation of a sentence as a linguistic unit by the variations of Fo, while the Hearing-Impaired string the individual words. They express the learned lexical stress by lengthening and by higher Fo of the stressed vowel. The following declarative sentence will demonstrate this observation.

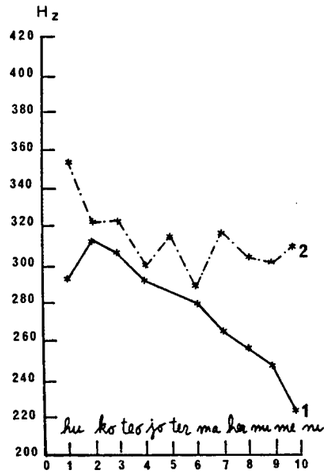


Fig. 1: "Hu ko'tev jo'ter ma'her mi'meni (He writes quicker than me). Group 1, the Hearing, show a declination line for this declarative sentence while group 2, the Hearing Impaired, say each individual word with higher Fo for the stressed vowel.

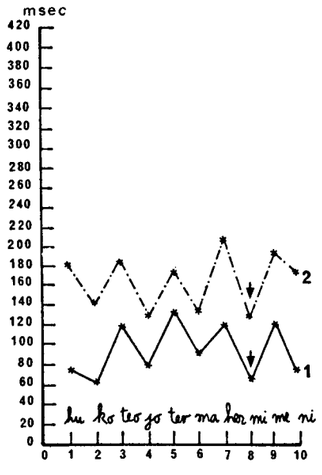


Fig. 2: The measured and connected mean duration values for the same sentence. There is an almost exact parallel between the Hearing (Group 1) and the Hearing-Impaired (Group 2). The Hearing-Impaired have longer absolute values but both groups show the ratio of unstressed/stressed vowel.

Conclusion: In both prosodic features, lexical stress and intonation, we found similar phenemena to the ones mentioned in other languages by other methods of research. It seems that the production of lexical stress is somewhat easier in Hebrew and the hearing-impaired children in this group were able to acquire it. But they are dependent on a learned pattern and therefore not able to divide, like the Hearing, between duration for word stress and variations of Fo for intonation. This could be due to a learning process that did not ask for flexibility but rather emphasized "proper articulation" with the appropriate stress pattern on the word level. Therefore, most of the children were not endowed with the ability to plan a whole sentence and to use pitch according to a plan.

This paper is a portion of a Ph.D. dissertation prepared at the Tel-Aviv University, 1986 and a paper in Tubin J. (ed.). From sign to language; a semiotic view of communication. John Benjamins, Amsterdam, Philadelphia (in press).

References:

1. Bolozky, Sh. (1978) Some aspects of moodern Hebrew phonology in Aronson Berman R., Modern Hebrew Structure. University Publishing Projects, Ltd. Tel-Aviv
2. Bothroyd A., Nickerson R.S., Stevens K.N. (1974). Temporal patterns in the speech of the Deaf. S.A.R.P. #15.
3. Bush, M. (1981). *Vowel Articulation and Laryngeal Control in the Speech of the Deaf.* Doctoral dissertation, M.I.T.
4. Fry, D.B. (1955). Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.* 27, 765-769.
5. Fry, D.B. (1958). Experiments in the perception of stress. *Lang. Speech* 1, 126.
6. Fry, D.B. (1970). Prosodic Phenomena in Malmberg B. (ed.) *Mannual of Phonetics*, North-Holland Publishing Company, Amsterdam, London.
7. Markides, A. (1983) *The Speech of Hearing-Impaired Chioldren.* Manchester University Press.
8. Osberger, M.J., McGarr, N.S. (1982). Speech Production Characteristics of the Hearing Impaired, in Lass, N.J. (ed.) *Speech and Language*, Advances in Basic Research and Practice, Vol. 8. Academic Press Inc., New York.

INSTRUMENTAL QUANTIFICATION OF THE 'OVER-ALL AMPLITUDE' FEATURE

JEAN-PIERRE ANGENOT
Dept. of Linguistics
Federal University of Santa Catarina
88000 Florianópolis (SC) Brazil

ALBERT LANDERCY
Laboratory of Phonetics
University of Mons
7000 Mons, Belgium

ULF HERMANN MONDL
Brazilian Society of Acoustics
c/o Eletrosul
88000 Florianópolis (SC) Brazil

ABSTRACT

The authors present an analytical model which permits a numerical quantification of the energy of linguistic sounds. This method will make possible the identification of a universal scale of over-all amplitude, in reference to which will be tested hypotheses, suggested by Guile (1974) and reformulated by Istre (1981), which explain a restricted natural class of phonological deletions, according to the definition of Natural Phonology (cf. Stampe & Donegan 1979, Dressler 1985).

01. GUILE'S AND ISTRE'S HYPOTHESES

Timothy Guile (1974) brought attention to the existence of a badly explained natural lenition process, characterized by the deletion of one of the obstruents present in a cluster formed by two or three obstruents. Simplification of aspired obstruents as well as that of the affricate to their fricative counterparts can be considered as a special case of this deletion phenomenon. Considering that the stridency binary feature (Jakobson, Fant & Halle 1967) is not able to provide an account for all the cases of deletion, Guile proposed the following hypotheses to account for this deletion phenomenon: obstruents with a larger over-all amplitude will have the effect of out-shouting their more debil neighbors. With the aid of an oscillograph, Guile took amplitude measurements of obstruents in various languages and established language-specific scales of over-all amplitude.

But, as Giles L. Istre (1981) pointed out, the experiment lends itself to some questions, some of which were raised by Guile himself. First, Guile admits that the scales are gross approximations based on comparisons made in only one or in some cases two phonetic environments. Since he furnishes no data in terms of physical measurements, we have no way of knowing how gross are the approximations or wether the measurements were made visually or instrumentally. The value of the use of other phonetic environments in future research is obvious. Second, the fact that the environments were restricted makes one wonder if, as Guile puts it, "ranking of certain segments with respect to over-all amplitude differ as a function of phonetic environment, and that using a scale obtained in one phonetic environment to predict deletions in another phonetic environment is misguided". Third, there is definitively a need for more informants per language since we have no way of proving that the informants utilized represent the mean. Regardless of the above questions, over-all amplitude does seem to account for deletion in the above experiment. Amplitude seems to play a role in assimilation (Ohala 1974). It also has an important role in certain types of lenition processes, including the spirantization cases which Hooper (1976) would like to set apart of the weakeningprocesses suffered by segments in other environments. It also could very well form the basis for many of the processes which phonologists attribute to labels such as 'strength' and 'sonority'. In summary, segments do seem to have some kind of relative property which sets one apart from another acoustically in a type of scale. Phonologists usually come to the conclusion that the scales are language-specific, i.e. each language will place its [s], for instance, at a point on its scale relative to the other segments of the system. A point for [s]in one language may not necessarily coincide with a point for [s] in another language. The first Istre's hypothesis is that scale is a universal one. The scale need not be language-specific if we approach it from a purely phonetic point-of-view. If we accept the premise that over-all amplitude is the factor which governs the perception of consonants, we may make a scale which can accomodate all of the speech segments of all languages. Thus, an English [s] could well have a different over-all amplitude than, say a French [s].

When we speak of amplitude, we are talking about the increase or decrease in air pressure at a given point during a sound. When Guile speaks of sound segments outshouting others, he is using the psychological term of 'loudness', the correlate of intensity. The physical formula $I \propto A^2 x\ F^2$ tells us that intensity, i.e. the kinetic energy in a sound wave, is proportional to the square of the amplitude times the square of the frequency. This relationship is such that if we reduce air pressure, we automatically reduce amplitude and, to a point, intensity. These things are important for the second Istre's hypothesis: sound mutations will originate more often toward the end of the breath-group, more especifically, after the drop in alveolar air pressure. If, for example, the alveolar air pressure drops below the 2 cm H_2O level needed for voicing at the end of the breath-group, we can expect that both voiced consonants and vowels will become voiceless. The condition of least articulatory control will manifest itself as the physiological basis for lenition. It is at the end of the breath-group that segments with more over-all amplitude will have the chance to out-shout their more debil neighbors. Once a mutation has occured in one position of the breath group, its results will have the tendency to spread to other positions of the group. The results of an original mutation will thus provoke a kind of chain effect. The Guile's and Istre's hypotheses presented here differ of other theories in that they are capable of being tested instrumentally.

02. PHYSICO-MATHEMATICAL BASIS

The method uses a Hewlett Packard computer (HP Fourier Analyser System 5451C) and a Brüel & Kjaer transductor (BK-Two Channel Microphone Power Supply Type 2807) which, by means of a condenser microphone (BK-Type 4165) and a pre-amplifier (BK-Type 2619), transform the acoustic energy of an utterance into electrical energy possible of cybernetic treatment. The physico-mathematical theory starts from the equation of the acoustic pressure of a sound wave expressed by the following formula:

$$p(t) = A(t).e^{jwt}$$

where p(t) = acoustic pressure in the time domain p = F.[2].

The acoustic wave incident on the microphone passes through the transductor and arrives in the memory block of the computer in the following form:

$$p(t) = k.v.e^{jwt}$$

where k = constant which takes into consideration the characteristics of the instrument and the physical units used [v]; e = base of the adimensional natural logarithms; t = time; v = voltage indicated by the computer.

The constant k in this case equals to 1 and can be modified by means of calibration measures of the transductor-computer system. In this case it is not applied in view of the fact that the ratio of energy of the different sounds being important in the linguistic evaluation.

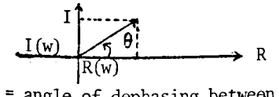
To transform pressure in the time domain to the frequency domain, we applied a Fourier transform:

$$(\mathcal{F} = \int_{-\infty}^{+\infty} p(t)e^{-jwt}dw)$$

which effected a transformation of the physical coordenates of time to frequency, however without altering the energy content:

$$\mathcal{F}\{p(t)\} = p(w) = R(w) + jI(w)$$

A Fourier transform always has a real component R(w) and an imaginary component I(w), whose physical sense expresses the phase for each frequency in relation to a common origin:



where θ = angle of dephasing between the real and the imaginary components; $\theta = \text{arc.tg}\ \frac{I(w)}{R(w)}$

As we wish to express the acoustic energy of the sound captured by the microphone of our instrumental system, we remember from physics that the acoustic energy is proportional to the square of the acoustic pressure, either in the time or the frequency domain:

$$E_{ac} = \mathcal{K} \cdot p^2(t)$$

where $\mathcal{K}$ = coefficient of the proportionality to adequate the pressure to the unit system used.

Applying this concept of acoustic energy, we obtain the acoustic energy of the captured wave by the squaring of the pressure in the time domain, thus furnishing a spectrum of acoustic energy:

$$E(w) = p^2(w) = R^2(w) + I^2(w)$$

It will be observed that the imaginary component disappears, seeing that any energy is a scalar value and not a vectorial, able to be expressed only by real numbers, contrary to pressure which is eminently vectorial.

To obtain the total energy, one only has to sum the components in every frequency, mathematically expressed by the following formula:

$$E_{tot} = \int_{w_1}^{w_2} p^2(w)dw = \int_{w_1}^{w_2} (R^2(w) + I^2(w))dw$$

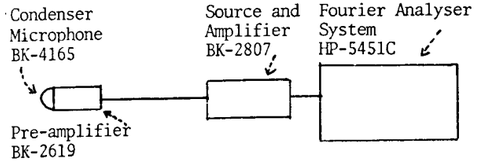
where w1 = frequency of the lower integration limit
w2 = frequency of the upper integration limit
dw = diferential of the frequencies

To compare the total energy of two sounds, one only has to effect its quotient, obtaining the relation in absolute and adimensional terms by the following formula:

$$Q = \frac{E_{tot\ 1}}{E_{tot\ 2}}$$

03. CYBERNETIC PROCEDURES

The instrumental system for the capture of sounds is composed of transductor, which consists of a microphone, a pre-amplifier, a source of voltage, an amplifier, and a computational system:



To capture the signal, one works with 4096 channels, a voltage key between 4 and 8 volts and frequencies up to 25 kHz. The routine used to register the utterance in 4 blocks of memory is the following:

```
RPLAC Ø E
LABEL 3 Ø E
ANALOG-IN 1 E
ANALOG-IN 2 E
ANALOG-IN 3 E
ANALOG-IN 4 E
CLEAR Ø E
END E
TERM E
```

To obtain the signal, we adjust the microphone and enter the utterance to be registered simultaneously with the command: JUMP 3 Ø E. We verify if the signals obtained are in the blocks of memory and we repeat the procedure if necessary. To filter out

background room noise, a good procedure consists of capturing the noise with a microphone, its storing in an adequate memory block and its substraction from other sound signals, annulling in this way any undesirable component.
For the segmentation of various sounds of an utterance, the command CLEAR is used in an adequate manner, deleting the undesirable parts and storing the results in memory blocks for later treatment.
Once a pertinent segment is isolated, a Fourier transform is performed by means of the command:

F Δ ø E

On the result obtained, the command:

M Δ ø E

is applied to obtain the energy spectrum, which is integrated by the command:

∫ Δ ø E

the last channel having to be read by means of the cursor.

04. ILLUSTRATION

For sake of illustration, we use the initial cluster [ps] from the Brazilian Portuguese word psicologia 'psychology', pronounced without [i] epenthesis.

Figure 1: oscillogram of the sequence [p](6.2 msc), [s] (29.4 msc) and the first cycle of the following [i].
Figure 2a: Real part of the Fourier transform of[p]

Figure 2b: Real part of the Fourier transform of[s]

Figure 3a: Imaginary part of the Fourier transform of [p]
Figure 3b: Imaginary part of the Fourier transform of [s]
Figure 4a: Energy spectrum of [p]

Figure 4b: Energy spectrum of [s]

Figure 5a: Integration of the energy spectrum of [p]

Figure 5b: Integration of the energy spectrum of [s]

With the over-all amplitude of [p] ($\int p^2(w)dw$) being 1.3260 E -3.0000 and the over-all amplitude of [s] ($\int s^2(w)dw$) being 36.6263 E -3.0000, the ratio of amplitude between the two obstruents is the following:

$$Q = \frac{36.6263}{1.3260} = 27.6216$$

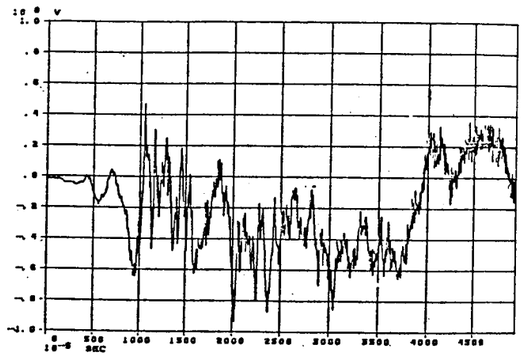


Figure 1

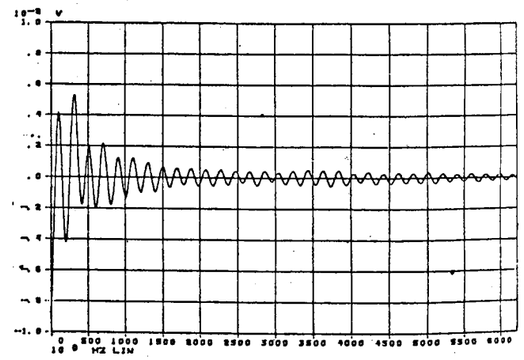


Figure 2a

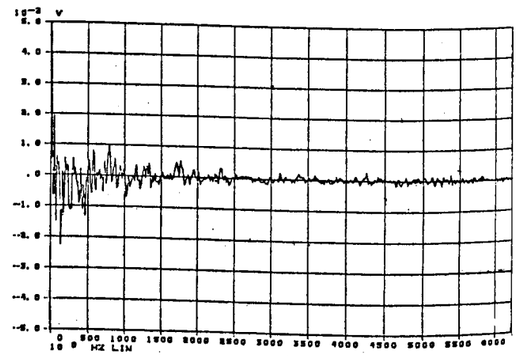


Figure 2b

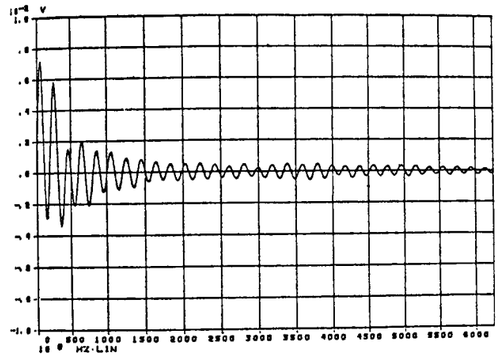


Figure 3a

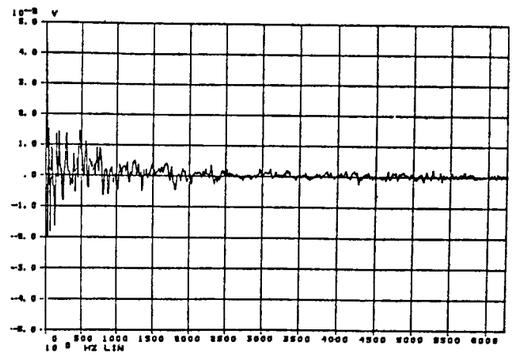


Figure 3b

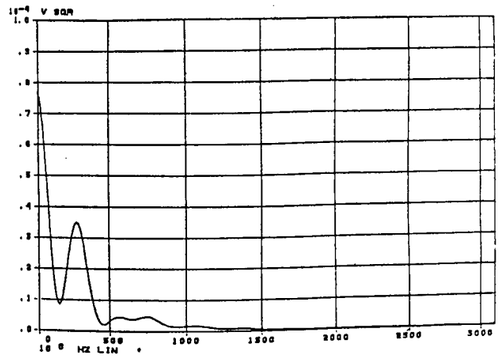


Figure 4a

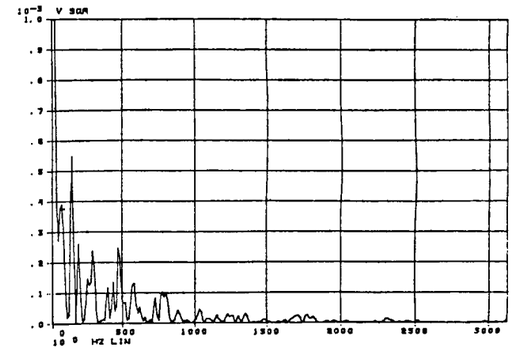


Figure 4b

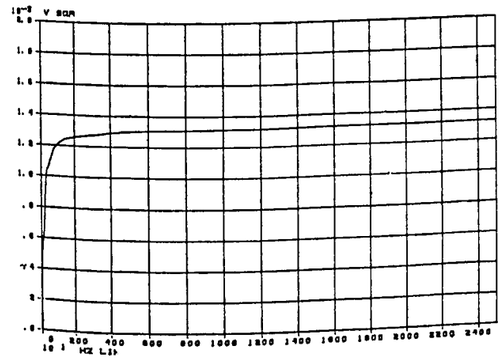


Figure 5a

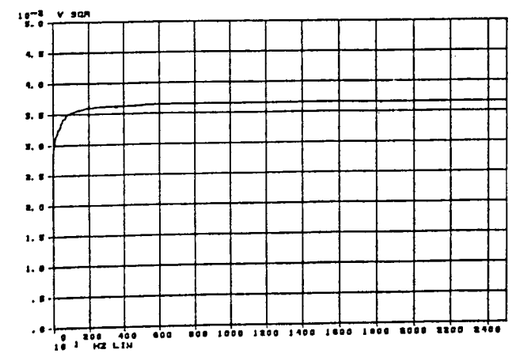


Figure 5b

05. GUILE'S CORPUS (selection)

(1) t --> ∅ / {{f,x}----, ----ʃl, ---- s, ---- θ} (a) Zurich dialect (b) Faroese (c) Danish (d) Colloquial American

Examples:
(a) Rüifft > Rüiff Germ. 'Ranft'
Nachtmaal > Nachmaal Germ. 'Abendmahl'
(b) nytsla [nɯʃla] 'use' cf. nyta[nɯta] 'to use'
(c) Tapetsere[tapesere]Germ. 'Tapezieren'.
cf. Tapet[tapet] Germ. 'Tapete'
Optional in loanword: Notits[notits]/[notis]
(d) hundredths[hʌndɪᵂɛtθs]/[hʌndɪᵂɛθs]

(2) p --> ∅ / {{l,r}----f, # ----f} (a) Middle High German (b) Colloquial Westphalen

Examples:
(a) *helpan > OHG hülpfan > MHG hülfan > hülfen
*werpan > OHG würpfan > MHG würfan > würfen
(b) Pfund [pfunth] / [funth] 'pound'

(3) k --> ∅ / {----st, #----x, ----xw} (a) Faroese (b) Swiss German dialect (c) Colloquial Zulu

Examples:
(a) russiskur /rɯs:iskt/ -->[rɯs:ikst]
--> [rɯs:ist] 'Russian'
(b) chung[xuŋg] 'king' cf. Germ. 'könig'
(c) khwebeza [kʰʷebeza]/[xwebeza] 'shrink'

(4) f --> ∅ / ----s (a) Southern Dutch

Example:
(a) lifst > list 'dearest'

(5)[θ,ʃ] --> ∅ /[s,z](a) Colloquial American English

Example:
(a) depths [dɛpθs] / [dɛps]

(6) s --> ∅ / (a) Colloquial American English (b) Colloquial Standard German (c) Colloquial Westphalen

Examples:
(a) fishsticks [fiʃstiks] /[fiʃtiks]
horseshoe [hɔɪsʃu:] /[hɔɪʃu:]
(b) kindischsten [khindiʃstən]/[khindiʃtən]'most childish'
(c) Du wäschst[vɛʃst] /[vɛʃt] '(you) wash'
ausschneiden [aʊsʃnaidən] / [aʊʃnaidən] 'to cut out'

(7) x --> ∅ / ----s (a) Southern Dutch

Example:
(a) sondes [sondəs] 'sunday' cf. Standard Dutch 'zondags'

(8) h --> ∅ / {x----, f----} (a) Colloquial Dutch (b) Afrikaans

Examples:
(a) heethoofdigheid[he:tho:vdixhɛit] or [he:tho:vdixɛit]hot-headedness
(b) lafhartig [lafaɾtax] 'cowardly'
cf. hart [haɾt] 'heart'

06. REFERENCES

Angenot, J.-P. & G.L.Istre (1986). "The phoneme 'ab ovo' and 'in vitro'", in Angenot et alii, 1-17

Angenot, J.-P. & G.L. Istre (1986), "Phonetic cloning for fast speech perception", paper presented in the Annual Meeting of the Linguistic Society of Europe, Ohrid, Macedonia

Angenot, J.-P., G.L. Istre, A. Landercy, D. Pagel & P. Vandresen, eds. (to appear), "Segmentation temporelle et analyse spectrale par FFT", in Angenot et alii, 1-13

Angenot, J.-P., G.L. Istre, A. Nicolacópulos & D. Pagel, eds. (1986), Miscellaneous Phonology, Florianópolis, Brazil: UFSC Working Papers in Linguistics, 204 pp.

Angenot, J.-P., G.L. Istre, A. Nicolacópulos & D. Pagel, eds. (to appear), Anais do 1º Encontro Nacional de Fonética e Fonologia. Florianópolis, Brazil: UFSC Working Papers in Linguistics, 105.

Angenot, J.-P., G.L. Istre, J.J. Spa & P. Vandresen, eds. (1981), Studies in Pure Natural Phonology and related Topics. Florianópolis, Brazil: UFSC Working Papers in Linguistics, 303 pp.

Bruck, A, R.A. Fox & M.W. Lagaly, eds. (1974),Papers from the Parasession on Natural Phonology. Chicago: Chicago Linguistic Society.

Dinnsen, D.A., ed. (1979), Current Approaches to Phonological Theory. Bloomington: Indiana University Press.

Donegan, P.A. & D. Stampe (1979), "The study of Natural Phonology', in Dinnsen, 126-173.

Dressler, W.U. (1984), "Explaining Natural Phonology" in Phonology Yearbook 1:29-51, Cambridge.

Dressler, W.U. (1985), Morphonology: The Dynamics of Derivation. Ann Arbor: Karoma Publishers, 439 pp.

Hooper, J.B. (1976). An Introduction to Natural Generative Phonology. New York: Academic Press.

Guile, T. (1974), "The amplitude scale and its implication for phonology", in Bruck et alii, 116-126.

Istre, G.L. (1981), "Take a deep breath", in Angenot et alii, 186-212.

Jakobson, R, C.G.M. Fant & M. Halle (1967). Preliminaries to Speech Analysis. Cambridge, Mass.: The MIT Press.

Landercy, A. (1984), "Extraction et analyse par FFT de segments de la chaîne parlée", Universidade Federal de Santa Catarina, Brazil, mimeo, 16 pp.

Ohala, J. (1974), "Phonetic explanation in phonology" in Bruck et alii, 251-274.

NATURAL PHONOLOGY AND SOME COMPETING PARADIGMS, WITH PARTICULAR REFERENCE TO SYLLABIFICATION

HANS BASBØLL

Scandinavian Department
Odense University, 5230 Odense M
Denmark

ABSTRACT
In my written contribution, the focus is on syllabification rules, rather than on competing paradigms. Three different kinds of syllabification principles are considered, and the importance of the syllabic peaks are discussed, particularly in terms of vowel strength. Syllabification structures are illustrated, synchronically with examples from Danish and French, diachronically with the Nordic i-Umlaut.

INTRODUCTORY REMARKS ON COMPETING PARADIGMS

Simon C. Dik, /1/, has classified linguistic paradigms into two main types, viz. formal and functional paradigms, the former being represented notably by Chomskyan linguistics, and the latter e.g. by Dik's own "Functional Grammar". In recent years, the importance of work being done within different kinds of functional paradigms in the broadest sense has been increasing. Some important paradigms are, in addition to Dik's "Functional Grammar" already mentioned, Halliday's "Functional Grammar", Givón's work on functional-typological syntax, and Langacker's "Cognitive Grammar", just to mention a few names and trends. These paradigms all represent a reaction against Chomskyan linguistics, and they use much more space to emphasize in which way they differ from the transformational-generative paradigm than to define their relation to other functional paradigms in the broad sense. Within the Chomskyan paradigm, on the other hand, very little attention is payed to functional paradigms.

The examples of functional paradigms mentioned so far have paid little attention to phonology. The most important functional paradigm within phonology must to-day be considered Natural Phonology, in my view, as represented in Dressler's huge integrative effort, /2/. The term 'functional' is in this context vague, of course, but there seems to me to be important common traits, e.g. in the foundation on basic principles outside linguistics proper, and in their use of plurifunctionality and conflicting goals. An integration of insights from different functional paradigms, which have largely been occupied with different aspects of linguistic structure, seems a promising enterprise.

Within prosody, the distinction between the two basic paradigms has not been as sharp as e.g. within syntax or "morphonology". But this of course does not mean that no differences in paradigm can be distinguished within prosody. In my oral contribution I intend to discuss how the syllabificational problems and structures I present in this paper can be handled within different phonological paradigms.

SYLLABIFICATIONAL PROBLEMS AND EVIDENCE

Theo Vennemann, /3/, sees it as a serious mistake to talk about syllabification rules, i.e. rules which syllabify a string of segments, technically e.g. by introducing syllable boundaries: syllabification cannot be due to universal rules, since different languages may syllabify the same string of segments differently, nor to language specific rules, since "in manchen Sprachen Silbenstruktur kontrastiv sein kann". Instead he talks about "Preferenzgesetze" (preferential laws). I agree with the two premises quoted, but not with the conclusion (but this may be mainly a terminological matter). In the many cases where syllable boundaries cannot be contrastive, I think structure building rules and constraints may be alternative appropriate formalisms, and they should at any rate be combined with markedness or preferential principles etc. The problem with the latter type is, of course, their interaction. The difference here is one of view-point, just as whether one prefers to talk about grammatical boundaries which determine syllable boundaries, or about marked syllabification signalling a grammatical boundary: either the speaker or the hearer defines the perspective.

One specific syllabificational problem in Danish (as in many other languages) is whether the division of an interlude into a possible coda followed by a possible onset is a basic principle, or whether the

division follows from general principles (e.g. relating to sonority, see below) without the necessity to refer to specific word-marginal clusters (which are, of course, "hard linguistic facts" to a higher degree than syllable-marginal clusters which are word-medial). In Danish, the answer is clear, I think, if we want to account for phonetic syllabification and not just for a phonological syllabification, as I have done earlier, /4/. Compare words like blomstre 'flourish' and kunstlet 'artificial'. There is scarecely doubt that, phonetically, the syllable boundary does not occur at similar positions: blom$stre vs. kunst$let. One might say that in kunst$let, the marked syllabification signals that a grammatical boundary cooccurs with $, but why does $ in blomstre not occur with the morpheme boundary? (cf. blomst 'flower'). The answer is, of course, that str is a possible onset, but neither stl nor tl nor dl.

I try to take the strongest possible stand on what will count as evidence for syllabification, viz. not only the phonetic (phenomenological and manifestation) syllable boundaries, but also the choice of "main variants" of the phonemes /p t k d g v r j/ (i.e. whether they are reduced to [b d g ð ɤ̯(/i̯/u̯) u̯ i̯], as they are in "final" position, or not), and whether short /a o/ occur with their particular variants before tautosyllabic consonants or not; in brief, the whole set of evidence which has earlier been used, in addition to the phonetic evidence.

Finally, I try to stay within a coherent general framework as far as prosodic structures are concerned. In the following, I shall consider syllables as composed directly of segments. This is a deliberate simplification, however, since I take there to be a tier with Weight Units, or Morae in between, as proposed by Hyman, /5/. In addition to Hyman's evidence, Danish stød and stress offer good evidence in favour of such a model, in my view, much more than for the now traditional division into onset plus rhyme (notice that I do not accept the argument that every time there is some systematic phonotactic restriction, it must be mirrored in the prosodic tree structure). In agreement with Hyman's model, I take a C to be universally adjoined to an immediately following V. Therefore the intervocalic C in words like bade [bæ:ðə] 'bathe' is taken to be ambisyllabic: the "weak d" seems to belong to the second syllable phonetically, but its manifestation is nevertheless "final" (see above). But notice that if only phonological (as opposed to phonetic) criteria are taken into account, intervocalic consonants before schwa are clearly "final" and not "initial".

PRINCIPLES OF SYLLABIFICATION

It is obvious that there are syllabificational principles of different kinds. These principles in their simplest, most general (completely unrestricted) form may correspond to processes of Natural Phonology, whereas the syllabification rules found in different languages are phonological in nature. The syllabification rules in certain languages are thus restricted in particular ways, both concerning each rule by itself and also the way in which different rules interact.

I shall classify the syllabification principles into three different kinds, forming a sort of hierarchy, as follows:

1) SYLLABIFICATION DEPENDS ON PHONOLOGICAL DOMAINS. This principle is probably not controversial (except for the specific formulation, perhaps). Within a framework of phonological(ly relevant grammatical) boundaries (partly dependent on, but not isomorphic with, the morpho-syntactic boundaries), the principle can be stated as follows: certain phonological(ly relevant grammatical) boundaries are obligatorily syllable boundaries too. Within a framework of phonological domains, /6/, the principle can be stated instead: Syllabification applies within a certain domain. This domain is language-dependent, e.g. syllabification ignores more boundaries in Romance languages than in Germanic. It also depends on formality level etc. And if syllabification is taken to occur at more than one level, its domains will increase during the derivation (the latter part is highly theory-dependent, e.g. in Natural Phonology there will be prelexical syllabification applying to morphemes, and resyllabification, as when morphemes are combined, /2/). The least restricted form of this principle is what you find in low-level phonetic syllabification; i.e. the phonetic syllabification is universally unmarked as against phonological syllabification. Except for a few remarks in the section on Phonological Syllabification in French, I shall not discuss this kind of syllabification principles here.

2) SYLLABIFICATION DEPENDS ON THE SYLLABIC PEAKS. If syllabification is taken to result in a syllable boundary occurring somewhere between any syllabic peak (hereafter abbreviated V, as non-peaks are abbreviated C) and the next one, it may seem quite trivial to point out that syllabification depends on the syllabic peaks. But the claim is, naturally, more interesting, since at least the following three subtypes of this principle can be destinguished:

2a) A following V "attracts" more Cs than a preceding V. This principle is universally accepted, I think: CV-syllables are considered to be the maximally unmarked syllable type, and a string ...VCVCVCV.... will have ...VCVCV$CV... as its unmarked syllabification. But it is going too far, in my view, when many phonologists explicitly or implicitly claim that the natural syllabification of a string VCV containing no grammatical boundaries is always V$CV, regardless of the nature of the vowels in question, since there are at least two competing principles:

2b) A "stronger" V "attracts" more Cs than a "weaker" V. Exactly what counts as a "stronger" or "weaker" V may be different in different languages. I shall argue that the difference between a fully stressed full vowel and an unstressed neutral vowel ("schwa") is prototypical for this distinction, and that a distinction of this sort has consequences for the syllabification in French and Danish. Notice that in many languages, such a distinction does not occur at all (and in these languages principle 2b thus can play no role), whereas the distinction in principle 2a is of course universal. Consequences of principle 2b will be investigated in the following sections on syllabification in Danish and French, respectively.

2c) A short stressed V "attracts" a following C more than a long stressed V, cf. the notions of "close and loose contact", respectively. This principle is related to what Vennemann, /3/ (p. 39), terms "PROKOSCHS GESETZ: Eine (dynamisch) akzentuierte Silbe ist um so stärker bevorzugt, je näher ihr Gewicht bei zwei Moren liegt", cf. the vowel lengthening and shortening in Middle German etc. ne.men>nē.men and dăh.te>dah.te, where the result is (still according to Vennemann) a bimoric stressed syllable, viz. consisting of short V plus C, or long V. The syllabification principle proposed in the present paper also has the consequence of letting the stressed short vowel be followed by a tautosyllabic consonant (although this will not necessarily be a bimoric syllable, according to my view). In the section on Nordic i-Umlaut and Syllabification, I shall illustrate principle 2c. Notice that the formulation of this principle presupposes a distinction between long and short vowels, and thus this principle, like principle 2b, in many languages cannot apply. Perhaps one reason for the recurrent (but by no means general, of course) neglect of something like principles 2b and 2c in the literature is exactly the fact that these principles, in contradistinction to principle 2a on the precedence relation, only apply to certain language types.

It is clear that all three subprinciples of 2 may interact. The situation with most tendency towards syllabification to the right will occur when the V to the left is stressed vowel and the V to the right the weak neutral vowel (schwa) (principle 2b), and where the former V is also short (principle 2c). Since syllabification to the right is not so generally recognized as syllabification to the left (the latter having sequences of open syllables as the prototypical case), I shall in the following sections illustrate exactly this kind of syllabification.

3) SYLLABIFICATION DEPENDS ON THE CONSONANTS. No one would probably deny this proposition, and a wealth of such subprinciples have been discussed in the literature. I shall limit myself to a few general remarks on the subject.

First of all, opinions differ wildly as regards the relation between the phonotactics of the syllable, the morpheme and the word. My own position is the following. The sonority hierarchy is basic; it is derived from the only universal redundancy restrictions for segment types there are: [-cons]≥[+son]≥[+voi] (i.e. non-consonantal segments, = Pike's vocoids, are necessarily sonorant, and sonorants are necessarily voiced), see fig. 1 at the end of the paper. Conceived not as a set of more and more inclusive segment types, but as a linear order (when you draw a diameter through all the circles), it predicts the following sonority hierarchy: vocoids, sonorants, voiced obstruents, voiceless obstruents. Notice that this model is forced to treat e.g. [sp] in onsets and [ps] in codas as non-violations of sonority, whereas it excludes e.g. [bs] in onsets and [sb] in codas (where [b] is voiced and [s] is voiceless); the model cannot be adjusted in any way to treat these clusters differently, and this is the result I would want (for empirical reasons). For further aspects of this model, e.g. its treatment of nasals and laterals, see /7/ with references. Whereas sonority underlies phonotactics, in my view (together with other principles, of course, e.g. concerning heterorganity of certain adjacent segments), the distinction between syllable-onsets and word-onsets, for example, is not crucial, since the prototypical clusters are those which are in words that are syllables, and morphemes too, at the same time, /2/. I take the notion 'possible initial cluster' to be important to the native speaker, and more important than the notion 'possible final cluster', at least in languages like Danish where we have many endings, but no prefixes, consisting of consonants only. I shall therefore use 'possible initial cluster' as a relevant notion in the principles of syllabification in Danish to be presented in the next section. Another corollary of my view of sonority is that intervocalic consonant clusters would be expected never to violate the sonority hierarchy just presented (in the obvious sense of sonority decrease followed by sonority increase).

Among the more specific syllabification principles I have used earlier in my account of Danish, /4/, is that two segments in certain contexts "count" as one with respect to syllabification, viz. /s/ plus plosive and plosive plus liquid. This device is not needed in the present system

of syllabification in Danish. The only consonant to be mentioned specifically in the principles is /g/ (which is manifested as a plosive initially, but as a continuant, a semivowel or zero finally), the weak Danish consonant par excellence (cf. plosive hierarchies proposed with [g] as their weakest member).

SYLLABIFICATION OF DANISH CONSONANTS BETWEEN A SHORT STRESSED VOWEL AND SCHWA

The material for this section is all such clusters registered by Basbøll and Wagner, /8/, departing from Dansk Retrogradordbog and Retskrivningsordbog.
The following syllabification rules are proposed here:

i) The first C is adjoined to the syllable of the preceding V. This is a syllabification rule which is specific to the context before a weak syllable.
ii) The last C is adjoined to the syllable of the following V. This principle, which is probably universal, has the consequence, together with principle i, that a single C between a short stressed V and a schwa is ambisyllabic (cf. the section on Syllabificational Problems and Evidence).
iii) /g/ is adjoined to the syllable of the preceding V. This syllabification rule, like i above, is specific to the context before a weak syllable, and it has the consequence, together with principle ii above, that a postconsonantal /g/ before a schwa will be ambisyllabic. No other consonants in clusters are ambisyllabic, according to the present proposal (recall that ambisyllabic means weak and in practice final as its main manifestation is concerned, see below).
iv) Non-adjoined Cs are adjoined to the syllable of the following V if they form a 'possible initial cluster' together with all immediately following Cs. This principle is by no means specific to the position before a schwa, in contradistinction to principles i and iii. Otherwise to the preceding V.

Recall (from the section on Syllabificational Problems and Evidence) that any C that is final (i.e. is part of the syllable of the preceding V, thus including ambisyllabic Cs under the heading 'final') undergoes consonant gradation if it belongs to the set /p t k d g v r j/ (resulting in [b d g ð ʁ (i̯/u̯) u̯ ʁ i̯], respectively). Notice, however, that non-final /p t k/ in schwa-syllables may be pronounced [b d g] (this is, in fact, the general pronunciation in Advanced Standard Danish); but its classification as non-final is justified by the possibility of a contrast between /p t/ and /b d/ in this position, where non-final /b d g/ are always pronounced as plosives. I thus try to describe all contrasts in the Conservative Standard, which makes the model much easier to falsify (and thus its empirical content higher). Observe also that short /a/ before a final grave C is grave ([ɑ]), and that short /o/ is [ɔ] before a final C, i.e. in closed syllables. Also the phonetic syllable boundaries (both in a phenomenological and a manifestation sense) are claimed to be reconcilable with the structures proposed, i.e. there is no alternative place which is more justified for the phonetic syllable boundaries. Thus the empirical coverage of the principles is much larger than by alternative principles which I know of, including those which I have formulated myself, /4/. I should add that plosives immediately following /s/ are generally not aspirated, but I cannot go into this here.
In fig. 2, at the end of the paper, a number of syllabification structures representative of the whole material are given. Notice that the principles apply correctly to abstract structures like æŋgste /ɛŋgstə/, but that they do not presuppose such abstractness (cf. the alternative /ɛŋstə/). The only case (of the more than 140 clusters tested) where the syllabification rules do not immediately give the desired result is fylgje pronounced [fylgjə], a completely isolated loan word as far as its interlude is concerned. If we would give it a phonological structure with /lkj/ we would suppose that it might be pronounced with [k] in distinct pronunciation in Conservative Standards, at least by speakers not knowing the word. The phonological structure /lgj/ would predict e.g. a pronunciation without /g/ (or with [ɣ̆], in very Conservative Standards). I do not know of any data illuminating the factual pronunciation of this isolated loan word, and I do not consider the problem important. Notice also that words like mugne [mɔgnə] must be posited with the phonological structure /moknə/, but this has nothing to do with syllabification problems, since the isolated root mug is also pronounced [mɔg], and similarly in other cases with written g after short vowels (this is thus, in my analysis, simply an incongruency in the phonology-orthography-relation).

FRENCH E-ADJUSTMENT AS A CONSEQUENCE OF SYLLABIFICATION

Since I have treated this subject elsewhere, /9/, although not within quite the same framework as the one here proposed, I shall just give what I consider to be the relevant syllabification rules and structures.
The syllabification rules apply within the domain delineated by #-boundaries according to my earlier proposals on French word structure. Notice that suffixes are more integrated with the stem than prefixes, and enclitic subjects more than enclitic non-subjects, according to this model. The syllabification here is of course phonological rather than phonetic since phonological rules in certain cases will change some of the conditioning factors, thereby leading to a resyllabification (according to the same principles, however).
The distinction between strong and weak V is simple at this stage of syllabification: final schwas in polysyllables are weak, all other Vs are strong (this distinction is related to the notion of 'accentuability'). At the phonetic surface, all Vs that remain (in Standard French) are strong, which leads to a resyllabification.
The following syllabification rules are proposed here (the rules underdetermine the syllabification structure in complicated clusters, but this problem is irrelevant for E-adjustment, and I do not want to take any strong stand on phonetic syllabification in French here).

A) The last C is adjoined to the syllable of the following V. This principle is probably a universal, as already stated.
B) A plosive immediately followed by a liquid is adjoined to the syllable of the following V. This is a case of "two close-knit-segments count as one with respect to syllabification", already alluded to.
C) If the preceding V is strong and the following weak (see above), the first C is adjoined to the syllable of the former V.
D) If the first C is unadjoined, it is adjoined to the syllable of the preceding V. These principles give rise to syllabification structures like those of fig. 3, at the end of this paper. E-adjustment can thus be formulated simply like this: {e,ə,ɛ} are neutralized in favour of [ɛ] in closed syllables. ({e,ə,ɛ} form a natural class in the technical sense according to my distinctive featureanalysis in /lo/).

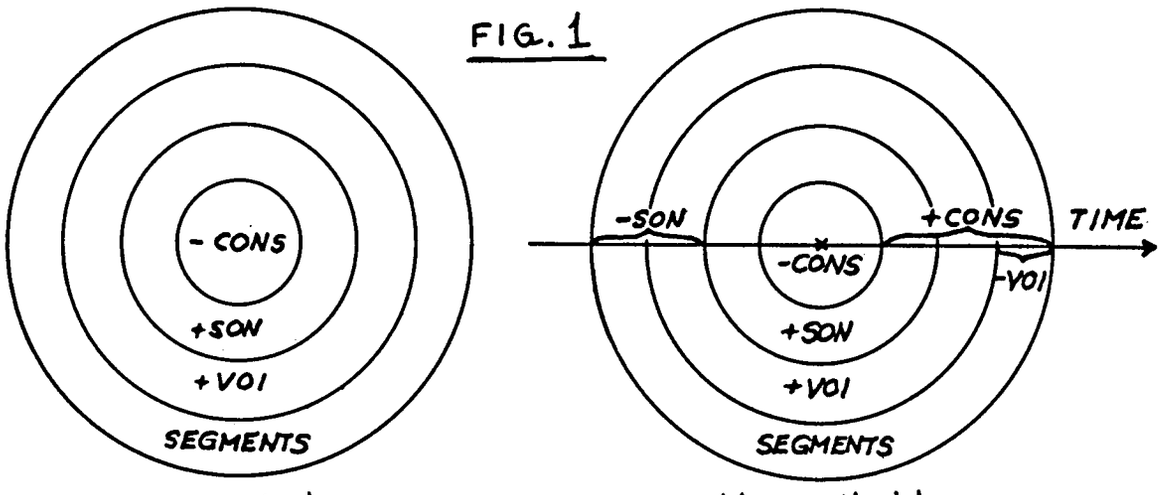
NORDIC I-UMLAUT AND SYLLABIFICATION

Several philologists have discussed the importance of syllabification in relation to the Nordic i-Umlaut, /11/. Here, I can only briefly state my own view, viz. that different syllabification may account for the basic difference (perhaps the crux) between long-syllable stems which as a rule undergo i-Umlaut under the influence of a (syncopated) /i/ in the following syllable, and short-syllable stems which do not, compare gastiR to gæstR and dōmiR to dømR, but staðiR to staðR. The important syllabificational principle in this context is 2c, which has the consequence that a C is adjoined to an immediately preceding short stressed V; thus the first syllable of the three examples would be gas$, dō$, staðˇ$. The account presupposes two phonological rules or processes (see below): a C is palatalized before a tautosyllabic /i/ (perhaps only in weak syllables, which would be natural for such a coarticulation process or rule), and: a segment is palatalized before a palatal(ized) C. The latter process or rule is "stronger" than the former in the sense that it applies across syllable boundaries (this agrees well with the fact that palatal Cs like /j R/ in Nordic generally cause i-Umlaut of preceding short-syllable stems as well as long ones). I have been vague with respect to the processual or phonological nature of the "change"; anyhow, the account does not presuppose that the palatalizations involved are phonological, the i-Umlaut will be phonological when the conditioning factors are lost. The fact that short-syllable stems with secondary stress in general undergo i-Umlaut, agrees well with the syllabification involved: the first V is not so heavy as in the normal case. Notice also that if the intervocalic C in a case like staðiR is taken to be ambisyllabic, it nevertheless cannot be palatalized by the following /i/; thus the crucial phonological distinction goes between "final" and "non-final" C, just as in the Danish and French examples discussed above.

REFERENCES

/1/ Functional Grammar[3], Dordrecht, 1978.
/2/ Morphonology, Ann Arbor, 1985.
/3/ Neuere Entwicklungen in der Phonologie, Berlin, 1986. (Quotation is from p. 25)
/4/ Phonological syllabification in Danish once more: A propos Molbæk's paper, APIPUC 14, 1980, 263-283 (with ref.)
/5/ A theory of Phonological Weight, Dordrecht, 1985.
/6/ H.Basbøll, On the use of Domains in phonology, Proc.Ling. 12, 1978, 763-766; E.Selkirk, On prosodic structure and its relation to syntactic structure, T.Fretheim (ed.), Nordic Prosody II, Trondheim, 1981, 111-140.
/7/ The structure of the syllable and a proposed hierarchy of distinctive features, Phonologica 1976, 1977, 143-148.
/8/ Kontrastive Phonologie des Deutschen und Dänischen, Tübingen, 1985.
/9/ Sur l'identité phonologique du schwa français et son rôle dans l'accentuation et dans la syllabation, P. Verluyten (ed.), Le schwa français, Linguisticae Investigationes, forthcoming (with ref.)
/10/ The feature tenseness in the modern French vowel system: A diachronic Perspective, ARIPUC 8, 1974, 173-200.
/11/ H.Benediktsson, Nordic Umlaut and Breaking: thirty years of research, Nord.J.Ling. 5, 1982, 1-50 (with ref.); H.Basbøll: Nordic i-Umlaut once more: A variational view, Fol.Ling.Hist. III, 1982, 59-86 (with ref.)

FIG. 1



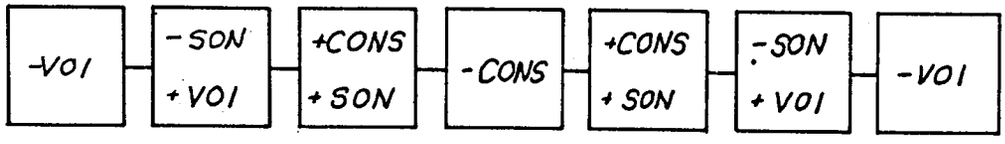
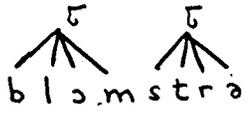
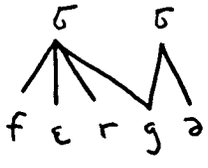


FIG. 2

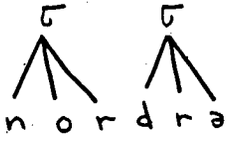
2a) 

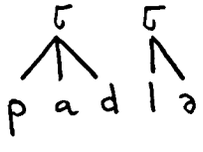
2b) 

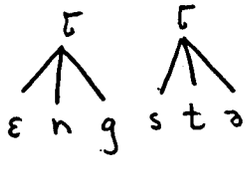
2c) 

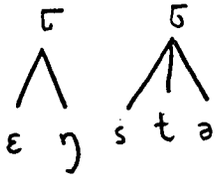
2d) 

2e) 

2f) 

2g) 

2h) 

2h') 

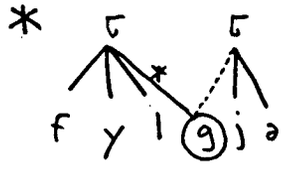
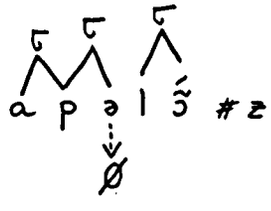
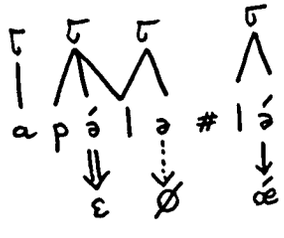
2i) 

FIG. 3

3a) 

3b) 

PHONETICS, PHONOLOGY, AND THE NATURAL OF IT.

PIER MARCO BERTINETTO

Scuola Normale Superiore, Pisa (Italy)

ABSTRACT

This paper discusses the recent debate in phonology between formally-oriented and functionally-oriented approaches (generative phonology and natural phonology). It claims that both views have inspiring insights and drawbacks, pleading for a substantial neutrality of phoneticians with respect to the phonological research. In particular, it is stressed that phonology is a fundamentally abstract discipline: its proper goal is to avoid arbitrariness, rather than attain concreteness, according to the now prevailing (and much too ambiguous) wave. However, it is to be hoped that some recent developments both in phonology and phonetics might provide the ground for a fruitful convergence of these disciplines, within their own domains.

1. The theoretical debate of the last fifteen years or so among phonologists was largely, albeit not exclusively, centered around the problem of naturalness, or 'concreteness' (as I will show below, the two terms should not be confused). Although the debate was held on a multipolar basis, one of its most intriguing aspects was the confrontation between MIT inspired phonology and natural phonology (à la Stampe and Co., but also à la Hooper and Co.). I shall devote my attention to this problem, trying as much as possible to take a neutral viewpoint, namely the viewpoint of a phonetician who keeps an eye on phonology in order to get fruitful inspiration for his own research.

The confrontation between generative phonology and natural phonology (henceforth, GP and NP) on the theme of naturalness/concreteness does not represent a completely new phenomenon. Quite on the contrary, this is just the latest episode of a more general, and much older, contraposition between formally-oriented and (so to say) functionally-oriented phonological theories, which has characterized the whole of phonology since its very beginning. This has been in fact a constant feature in the history of the discipline, much before the advent of GP. Indeed, if viewed from the corner of naturalness, many streams of structuralist phonology, both in Europe and in the USA, look much more abstract than orthodox GP. As an example, think of Bloomfield's /ɔ/ analysis of Menomini, where an underlying /w/ is arbitrarily postulated only to the effect of blocking the expected palatalization of the sequence /tj/ in those words which exceptionally escape this process. Yet, Bloomfield's approach does not represent the most extreme case of abstraction in structuralist phonology: for that matter, just think of such scholars as Hjelmslev or Z. Harris. Needless to say, not all structuralist phonologists show this extreme neglect for the phonetic substance: Jakobson is a very clear example of a phonologist with a sharp interest in the physical support of language. However, in some sense it can be said that GP yielded a substantial change in the orientation of phonological studies. Even the well-known Hallean paradox, concerning the absence of a level of autonomous phonemics in Russian (and in phonological theory), can be understood in this light: the intermediate level of autonomous phonemics is rejected in as much as it does not add any relevant piece of information on the link between the abstract morphophonemic level and the concrete allophonic level.

Curiously enough, then, the kind of objections which NP raises against GP are partly of the same kind as the objections which GP raised against structuralist phonology: namely, the undue neglect of the phonetic substance. There is ground to say, therefore, that GP and NP appear to be very different, or quite similar, to each other, according to the distance from which they are looked at. If compared to the most abstract among the structuralist approaches, they both look quite concrete; if compared to each other, NP looks much more substance-oriented than GP.

In what follows, I shall try to consider the topic of naturalness, and the confrontation of GP and NP, from the viewpoint of phonetics. Tow questions are of particular interest in this context:

1) Who did prevail in the recent theoretical debate (the formally-oriented, or the substance-oriented)?
2) Is there any special lesson to be learned for phoneticians?

The answers to these questions will be tentatively given in sections 3 and 4 below; in section 2, I shall be concerned again with the topic of concreteness/functionality, claiming that too much of an emphasis has been put on this concept in recent works.

2. There is no doubt that the discussions which took place on the matter of naturalness were in the whole very instructive and fruitful. For one thing, it appears to me that these discussions forced the adhe-

rents to GP to reanalyse their own positions, and to pay more attention than ever before to this issue. However, not everything went on as smoothly as that. I would like to provide a couple of examples.

The first example concerns the by now very well-known case of the k-Ø alternation in Turkish, which unpredictably characterizes the declension of a certain class of nouns. The traditional analysis performed by Lees /15/, attributes the alternation to the lexical differentiation between underlying /g/, which regularly devoices in syllable-final position and deletes intervocalically (in the Istanbul dialect), and underlying /k/ which is preserved. This seems to be in accord with the historical facts. However, as was suggested by Zimmer /20/ and Zimmer & Abbott /21/, the historical explanation does not account for the synchronic data, since /k/ deletion tends to operate in monosyllabic roots only, to the exclusion of polysyllabic ones. Actually, the situation is probably more complicated than that, according to Sezer /17/, because /k/ is preserved also in polysyllabic roots when it follows a long vowel (Zimmer & Abbott notice this fact mainly in connection with the behaviour of Arabic loanwords, which often retain /k/). Whatever is the real explanation of this phenomenon, I would like to call attention to a particular aspect of this debate, which has been unduly taken for granted by all participants. Namely, the solution put forth by Zimmer, based on syllabic count, was labelled the 'functional' solution, following an explicit statement by Zimmer himself. This is enhanced even in the title of a notorious paper by Halle /8/, which correctly criticizes Zimmer's claim that his own solution is to be preferred (on the ground of simplicity) to the 'formal' solution based on the underlying alternation between /g/ and /k/. Now, it is difficult to understand just what the term 'functional' means in this case. No doubt, polysyllabic words are longer than monosyllabic ones; consequently, a deletion process which concerns polysyllabic words only, might be regarded as a tendency to equalize the length of the words in the language. However, in the present case this explanation looks quite suspicious: why on earth should Turkish speakers make recourse to this 'functional' tendency only in the case of polysyllabic roots ending in /k/, to the exclusion of those ending in /p,t.../? I believe Halle is perfectly right in claiming that Zimmer's solution is on the same level, from the point of view of simplicity, as the more abstract approach consisting in postulating a lexically idiosyncratic morphophonemic alternation between /g/ and /k/. For that matter, Zimmer's proposal should be accepted rather on the ground of being more 'concrete' than the alternative proposal, although I am not sure that this argument is really compelling in the face of the whole system of Turkish morphophonology, where the alternation between voiced and voiceless stops has to be postulated anyway on independent grounds.

In conclusion, the Turkish case provides a good example of the distorted usage, which is sometimes to be observed, of the word 'functional'. A similar example is provided by Stemberger's /18/ analysis of the so-called empty consonants of French: where the author claims that his analysis is more concrete than the traditional approach (consisting in the postulation of underlying consonants, to be deleted in some particular contexts). The new solution consists in suggesting that the relevant words of French which undergo this phonological process show an extrasyllabic C in their CV-skeleton, which is eventually deleted whenever it is not captured by a process of resyllabification. Here it is very difficult for me to understand what is the meaning of the term 'concrete'. True, a C position in the CV-skeleton does not need to be filled with any phonetic content, whereas an abstract consonant should be provided with a fully specified distinctive feature matrix, at least according to the current view. However, why on earth should a phonological representation look more concrete as a consequence of the insertion of an underlying C slot? If anything which is postulated by a given phonological theory deserved the qualification of 'concrete' just because it has been postulated, then any phonological object would be concrete! Stemberger & Marlett /16/ go so far as to claim that their treatment of empty consonants in Seri (basically similar to the one proposed for French by Stemberger) is even consistent with Venneman's 'Strong Naturalness Condition', which states that underlying representations should coincide with some surface form; but I just fail to see how any French consonant can be said to 'concretely' exhibit a C position in its surface form (apart from the fact that the Strong Naturalness Condition does not seem to be tenable on theoretical grounds, as pointed out e.g. by Kenstowicz & Kisseberth /12/.

It appears to me that what is involved here is a category mistake: Stemberger and coworkers seem to be strongly biased towards concrete analyses, and do not hesitate to call 'concrete' their own analysis of empty consonants, neglecting the simple fact that the CV-skeleton cannot on principled grounds be regarded as a concrete object. Understandably, Klausenburger /14/ argues that Stemberger's analysis of French empty consonants strains the concept of concreteness "beyond recognition".

I chose these two examples (the Turkish and the French one) in order to show two different instances of a quite common fallacy, namely the exasperated need to adhere to the by now prevailing wave of concreteness. This is a characteristically new phenomenon in the history of phonology: fifteen years ago, or so, most phonologists would have regarded as inappropriate any appeal to concreteness in their analyses. Now, things have gone so far, that concreteness is invoked even in the wrong cases. Although it might look funny to do so, it seems to be time for a phonetician (as I mainly consider myself) to remind phonologists of the essentially abstract nature of their own domain of research. As a matter of principle, nothing can be objected against people who incline towards abstract phonological analyses, provided they refrain from taking arbitrary steps: what should be avoided, as a source of confusion, is the undue appeal to concreteness or 'functionality'. Pushing things a little bit further, one might in fact claim that the very idea of concreteness looks quite meaningless as applied to phonology. I bet M. le Prince Trubeckoj turns in his grave every time the word 'concreteness' is uttered in the context of phonology. Alternatively, and choosing a milder formulation, one might claim that too often the polarity 'abstract vs concrete' is mistakingly understood as the dichotomy 'arbitrary vs motivated', where the illegitimate term is of course arbitrariness, not abstraction.

3. Let us now try to consider the first of the two questions announced above. From what I said in the preceding section, it might appear that the functionally-oriented stream of phonology has prevailed over the alternative one during the theoretical debate of the last decade: indeed, as was noted, never before was there such a pervasive acceptence of the concreteness issue. However, things are not as neat as they might appear. The now overwhelming acceptence of the 'concrete' vocation of phonology is just the newest version of the so-called 'naturalness condition', which has always been a topic in most versions of GP. Any analysis which minimizes the number of distinctive features necessary to express a given generalization has always been regarded as the most highly valued, and termed a natural solution. The very search for a set of distinctive features was largely motivated by the goal of capturing as many natural classes of sounds with as few features as possible. Nevertheless, this concern never prevented GP from assuming very abstract positions, nor from taking the step of evaluating its own results on the basis of fairly mechanical features' computations (thus showing a strictly formal, rather than functional attitude). On the whole, the issue of concreteness/naturalness was often paid little more than mere lip service. So, it would be mistaken to say that NP succeeded in imposing a functional perspective on contemporary phonology: the theoretical issues it raised were undoubtedly taken seriously, but (as we saw above) they sometimes ran the risk of being strained beyond recognition.

The point is that the stance of GP on the matter of concreteness has always been somewhat ambiguous. On the one hand, GP has defined its set of distinctive features in quite physical terms (following the Jakobsonian tradition), thus suggesting a fundamental link between phonology and phonetics; on the other hand (as just noted), GP has devoted a great deal of attention to purely formal matters, such as the 'simplicity metrics', based on the mechanical computation of features. In addition to this, GP has very often shown little concern for the distinction between synchronic and diachronic processes. One such example is Kisseberth's /13/ analysis of Yawelmani, which brilliantly reconstructs an underlying vowel system, on the basis of which all the apparently misterious anomalies of the surface phonetic form can be explained. But if one looks at it from the point of view of NP, then one cannot escape the following question: namely, is the postulated underlying /u:/, which never surfaces in any word of contemporary Yawelmani, really present to the phonological competence of the native speaker, or is it nothing more than an historical fact, devoid of any synchronic validity? To put it in other words, is there ground to postulate synchronic rules which derive surface forms such as c'omhun from the underlying /c'u:mhin/ (where the i/u alternation in the final syllable is due to vowel harmony), or should one regard this as a purely morphologized and lexicalized process, largely opaque to the contemporary speaker?

It is clear enough that the answer to these questions cannot be given without taking a strong theoretical position (i.e. a position which far exceeds the available evidence and represents a guess as to the proper basis to develop a framework of investigation and research). Generative phonologists tend to consider the objections raised by NP as irrelevant. They do not deny that morphological alternations exist; however, they believe there is no compelling evidence that these alternations (except for a minor portion of them) are mere reflexes of historical developments, rather than synchronic processes located in the brain of the speaker. After all, who knows how large the human capacity for fairly complex online computations is? The counterobjection of NP is that there are observable differences between processes which are directly reflected in the speakers' phonetic behaviour, and processes which do not surface explicitly. For instance, the speaker of a language with final obstruent devoicing will tend to apply this process also to foreign and nonsense words, whereas it would be quite surprising if the Yawelmani speaker produced (in the appropriate context) /o/ instead of /u:/ when pronouncing a foreign or a nonsense word. Thus, the NP claim that synchronic and diachronic processes must be kept apart from each other seems to be substantiated by observational evidence; and indeed, this is the strongest point in favour of NP. Yet, even this kind of argument does not solve the dispute. The charge which is usually made by generative phonologists against natural phonologists is that the latter reduce the relevance of the phonological component by depriving it of part of its content; and although this deprivation is done to the benefit of morphology, the splitting weakens the predictive power of the theory, since many phono-morphological phenomena are no more considered to be direc

tly linked by a single line of derivation. And one could hardly deny that the theoretical assumptions defended by GP enabled it to provide several inspiring and fully developed pieces of description of the phonological component of many different languages.

As I said, there does not seem to be a neutral or pretheoretical point of view about these matters. And indeed, even if we consider the problem from the point of view of phonetics, it should appear that either answer is perfectly legitimate. There is no reason why a phonetician should be bothered by the possibility that the phonological component is structured in such a way as to contain a fairly abstract set of rules, provided they do not suffer from arbitrariness, and provided of course that the output coincides with what is actually pronounced by the native speaker.

This last consideration might remind us of the position of those phonologists (adhering to GP) who defended the claim of the unnaturalness of phonology, in sharp contraposition to NP (cf., e.g., Anderson /1/ or Hellberg /11/). However, I do not regard those criticisms as especially harmful to NP. Indeed, NP never claimed that every phonological process is natural, for one must take into consideration the often unpredictable development of human languages, subject to the contribution of many diverse factors, such as contact with other languages, or morphological patterning. In fact, a great deal of the sound structure of any language is under the control of morphology, rather than phonology, and this has to be reckoned by any phonological theory. This is not to deny, though, the importance of Anderson's contribution, and of those who made the same move. On the contrary, their warning reminds us that any attempt to squash phonology onto phonetics is bound to failure; and it is a fact that NP, admittedly, has but a weak explanatory power: it does not tell us why a given phonological process occurs, rather it tells us whether that process is to be expected on phonetic grounds. Also, it is a fact that the first attempt to develop a comprehensive framework to motivate morphophonological patternings along the lines of NP is Dressler's /4/, which appeared long after the birth of NP.

4. With this caveat in mind, we can finally address the second question put forth in section 1. Let us recapitulate some of the observations which were advanced so far:

(i) phonology is an abstract domain of research;
(ii) in order to avoid the risk of arbitrariness in the analysis, some restriction must be posited, the most reasonable one being a restriction of phonetic naturalness (i.e. plausibility);
(iii) however, the pursuit of naturalness must not be carried out at the expense of principle (i).

Now, at first glance, one might advance the view that the general trend of contemporary phonology towards the naturalness/concreteness issue goes very much in the same direction (despite the equivocal episodes described above in section 2) towards which phonetics is intrinsically oriented. I would like to claim, though, that this is not entirely true. No doubt, an 'unnatural' phonetics is hardly conceivable; furthermore, it is a fact that a purely abstract (which amounts to saying: arbitrary) phonology would be of no help to phonetic sciences. Nevertheless, I do not think that phoneticians (putting aside personal predilections) should encourage NP any more than GP, or in general functionally-oriented theories at the expense of formally-oriented ones. The basic contribution of phonology to phonetics is to be sought in its propensity to provide theories which can occasionally be tested experimentally, or inspire the conception of new ideas about the production of speech. In this respect, a formally-minded phonology might even provide better material for phonetic speculation, just because of its more provocative character. The search for phonetic motivation for abstract phonological processes is, after all, the fundamental challenge to phoneticians.

From this point of view, non-linear phonologies might easily prove to be more challenging than any version of NP. An interesting example of this can be found in Clements /3/, who explicitly tries to develop a phonetically motivated theory of autosegmental phonology, where each articulatory dimension corresponds to an individual autosegment, all this leading to a hierarchic conception of the feature content of phonemes. It is envisageable that this view of the phoneme will induce a new stream of research in phonetics, just as the traditional view of the phoneme as an internally unstructured matrix of distinctive features inspired important works. And it might be that something of this sort will be eventually triggered by a specific branch of non-linear phonology I alluded to above, namely CV-phonology, although the first attempt at experimental verification carried out by Stemberger & Mc Whinney /19/ is far from successful. A much more successful one is Hayes's/10/, which accumulates empirical evidence (in terms of reactions to a number of phonological processes) for the existence of multiple vs biunivocal linkings between the segmental and the CV tier (thus substantiating with new and convincing arguments the old view that geminates behave differently in the various languages); however, Hayes's approach is not experimental.

It is quite instructive, in any case, to see how often CV-phonologists try to provide a physical basis to the abstract entities they postulate: indeed, this is another instance of the concreteness trend now prevailing in phonological research (but recall the criticisms put forth in section 2 above). CV positions (or x positions, depending on the particular framework) are very often said to be 'timing units' and Marlett & Stemberger /16/ even speak of empty consonant positions as something which produces consequences on the level of motor programming, although they admit that more work has to be done "on the low-level phonetic facts of empty syllable positions" (p. 637). Honestly, I think all this goes a bit too far. I have the impression that most phonologists did not notice what has been going on in phonetics during the last decade, in particular in the area known as 'action theory', where the issue of timing was extensively reexamined in the light of new acquisitions in physiology and psychology (cf. Journal of Phonetics 14,1,1986 for a recent debate on this topic). Of course, some phonologists did notice it, and even raised severe objections: cf. Hammarberg's /9/ polemics against Fowler /5,6/. However, I believe Hammarberg's emphasis on mentalism as opposed to physicalism is ill-founded: the recognition of the articulatory-acoustic correlates of phonemes, and of the intrinsic need for coarticulation (possibly in terms of coproduction, as suggested in Fowler & Smith /7/), does not exclude the fundamentally abstract nature of phonemes as mental entities responsible of the phonological patterns of natural languages. Besides, the physical properties of the speech mechanism are connected to higher-level principles, which themselves depend on the functional properties of the organism. If looked at from this ecological perspective, phonetics is no more a merely concrete domain of research; rather, it becomes a field which incorporates a notable degree of abstraction.

The elucidation of these topics is the matter of future research. Let me just say, to conclude, that I entertain the hope that some recent developments in phonology, as illustrated by Clements /3/, will provide the theoretical basis for a fruitful rendez-vous of phonetics and phonology, and possibly (why not?) for the convergence of NP and GP (or at least the most influential version of the latter) on the common ground of phonetic motivation (and plausibility) of phonological processes.

BIBLIOGRAPHICAL REFERENCES.

/1/ Anderson, S.R. (1981), "Why phonology isn't natural", Linguistic Inquiry 12: 493-540.
/2/ Bloomfield, L. (1939), "Menomini morphophonemics", Travaux du Cercle Linguistique de Prague 8: 105-115 (also in V. Makkai, ed. Phonological Theory, New York, Holt 1972: 58-74).
/3/ Clements, G.N. (1985), "The geometry of phonological features", Phonology Yearbook 2: 225-52.
/4/ Dressler, W.U. (1985), Morphonology. The Dynamics of Derivation, Ann Arbor, Karoma.
/5/ Fowler, C.A. (1980), "Coarticulation and theories of extrinsic timing", J.of Phonetics 8: 113-33.
/6/ Fowler, C.A. (1983), "Realism and unrealism: a reply", J. of Phonetics 11: 303:22.
/7/ Fowler, C.A. & Smith, M.R. (1986), "Speech perception as 'vector analysis': an approach to the problems of invariance and segmentation", in J.S. Perkell & D.H. Klatt, eds., Invariance and Variability in Speech Processes, Hillsdale, N.J.; Erlbaum: 123-39.
/8/ Halle, M. (1979), "Formal vs. functional considerations in phonology", IULC.
/9/ Hammarberg, R. (1982), "On redefining coarticulation", J. of Phonetics 10: 123-37.
/10/ Hayes, B. (1986), "Inalterability in CV phonology", Language 62,2: 321-51.
/11/ Hellberg, S. (1978), "Unnatural phonology", J. of Linguistics 14: 157-78.
/12/ Kenstowicz, M. & Kisseberth, C. (1979), Generative Phonology: Description and Theory, New York etc., Academic Press.
/13/ Kisseberth, C. (1969), "On the abstractness of phonology: the evidence from Yawelmani", Papers in Linguistics 1: 248-82.
/14/Klausenburger, J. (1986), "How concrete is extrasyllabicity?", J. of Linguistics 22: 439-41.
/15/ Lees, R.B. (1961), The Phonology of Modern Standard Turkish, Bloomington, Indiana Univ. Press.
/16/ Marlett, S.A. & Stemberger, J.P. (1983), "Empty consonants in Seri", Linguistic Inquiry 14: 617-39.
/17/ Sezer, E. (1981), "The k-Ø alternation in Turkish", in G.N. Clements, ed., Harvard Studies in Phonology 2.
/18/ Stemberger, J.P. (1985), "CV phonology and French consonants" J. of Linguistics 21: 453-57.
/19/ Stemberger, J.P. & Mc Whinney, B. (1984), "Extrasyllabic consonants in CV phonology: an experimental test", J. of Phonetics 12: 355-66.
/20/ Zimmer, K.E. (1975), "Some thoughts on likely phonologies for non-ideal speakers", in R.E. Grossman et al., eds., Papers from the Parasession on Functionalism, CLS: 556-67.
/21/ Zimmer, K.E. & Abbott, B. (1978), "The k-Ø alternation in Turkish: some experimental evidence for its productivity", J. of Psycholinguistic Research 7: 35-46.

PROTOTYPICAL SPEECH EVENTS AND SPEECH PERCEPTION

Grzegorz Dogil

Fakultät für Linguistik und Literaturwissenschaft,
Universität Bielefeld, D-4800 Bielefeld 1, BRD.

ABSTRACT

Natural Phonology makes a basic distinction between two process types: the processes which foster the production side of speech (lenitions) and the processes which foster speech perception (fortitions). It has been argued that these two process types are not only functionally distinct but that they also apply to distinct sets of structural positions. Thus the fortitions do not generally apply to the syllable final consonants (more generally speaking - VC - syllabic speech events), whereas lenitions usually spare the syllable initial position (the CV syllabic speech event). From this most basic assumption of Natural Phonology it follows that the CV parts of the syllables should form a group of perceptually salient speech events. In a series of experiments on speech parsing by humans and its simulation by machines we will show that this general prediction of Natural Phonology is strongly supported by phonetic facts.

INTRODUCTION

There are two basic mysteries about natural language: the speed and ease with which it is acquired by a child */the acquisition mystery/* and the speed and ease with which it is processed */the processing mystery/*. The speed and ease of language acquisition is so mysterious because it takes place in the environment of highly deficient input data. There must be then some underlying principles which help children override this deficiency of input data. An attempt to find the underlying principles of the acquisition mystery produced the most important innovations in modern linguistics. The solution of the Natural Phonology - acquisition of phonology as the 'unlearning' (increasing inhibition) of process types - has been not only one of the most original, but also the one with the strongest impact in child phonology.[1] The processing mystery, on the other hand, has been much less popular with the general linguistic community,the natural phonological paradigm included. This paper is an attempt to break with this tradition. Following the basic idea that fortitions foster perception we will illustrate an idea of a parses (a general perceptual mechanism) which considers only these parts of the string where fortitions are allowed.

Similarly to language acquisition, language processing faces a strong input-data-deficiency problem. When we speak we alter a great deal in the idealized phonetic and phonological representations. We delete whole phonems, we radically change allophones, we shift stresses, we break up intonational patters, we insert pauses in the most unexpected places, etc. If to such crippled phonological strings we add all background noise which does not help comprehension either, it is difficult to imagine how the parser is supposed to recognize anything at all. However, even in the most difficult circumstances (foreign accent, loud environment, drunkenness, etc.) we do comprehend speech quickly and efficiently. There must be then some signals in the phonetic string which are particularly easy to grasp and to process. We call these signals *PIVOTS* and parsers working with these signals we call *PIVOT PARSERS*. Until now we have thought only of the phonetic and phonological (or to be more exact - prosodic) pivot parsers, but we believe the idea may be transformed to other types of parsing as well.

THE PHONOLOGICAL *PIVOT*

What are then the pivots in the phonetic string? Dogil [2] argues that at each level of prosodic organization there exist prototypical, unmarked structures which not only manifest themselves in patterns of all natural languages but are also clearly visible in the areas of external evidence such as language acquisition, language loss, and language change. Here we will argue that these ideal prosodic types play an important role in language processing.

At the lowest prosodic level - the level of the syllable - such an ideal type is constituted by a *CV syllable*. That is, the prototypical, unmarked syllable consists of a single consonant followed by a vowel. There is plenty of evidence for this prototype.[2] For example:

- there is no language which does not have CV syllables, but there are many languages which have only CV syllables.
- phonological rules which obliterate syllabic structure usually spare CV syllables.
- CV syllables are acquired first in the process of language acquisition.
- CV syllables are preserved even in the most severe forms of motor aphasia.
- historical syllabic restructuring rules tend towards the creation of CV syllables.
- when subjects are asked to synchronize clicks with syllables it turns out that the clicks are aligned at a point, called the *P-CENTRE* (or 'perceptual centre'), which is close to the CV transitions of the syllable.
- listeners can classify stops by place better than chance when they are given only the first 10-46 msec. of CV syllables.
- the parameters of initial and final transition segments of vowels are not symmetrical in symmetrical syllables (pap, bab, etc.). The parameters of initial transitions may be successfully used as features of the adjacent consonant place of articulation, but the parameters of final transitions are useful as features only in few particular cases.
- when place of articulation cues are different at VC and CV transitions, listeners tend to follow the CV cues.
- speakers try to create temporally more well defined, more precise, articulations near the CV as opposed to VC interface.

All this evidence clearly illustrates the prototypical character of this unit. We claim that this unit is also essential for pre-lexical parsing. *What the parser essentially does is recognize CV syllables in the string*. We propose it does it in the following way:[3]

-- The parser searches for the first CV transition (the 'acoustic boundary' between the consonant and the vowel) and once it has found it, it stops. The parser makes a series of overlapping spectra which spread outwards from the transition point. This gives a *diphonic* representation of the CV syllables.

-- The parser recognizes the syllable. Strictly speaking it recognizes only the unmarked, prototypical CV part of the syllable. These prototypical CV's are stored as diphones in the diphone dictionary. If the syllable contains other units, for example if it is CCVCC syllable (like in the name *Planck*) these other units will be disregarded, and only the CV ([la] of [plaŋk]) will be available after the initial parse.

-- Having identified the syllable the parser makes its first hypothesis about the word of which this syllable is a part.

-- The parsing strategy is continued by jumping to the next transition, i.e. the next CV syllable.

Given all the grammatical, contextual and background knowledge that we possess when parsing strings, the syllabic pivot parser might be actually sufficient for comprehension. Even if it is insufficient in the form that we have presented it so far, it is fast enough to incorporate a number of modification strategies that can make it sufficient for comprehension.

EXPERIMENTAL VALIDATION FOR THE PIVOT: Experiment 1

As a first step we simulated the strategy of the pivot parser by doing some simple speech editing. For the first experiment we chose 10 sentences which show the broad range of syntactic and phonological (segmental and syllabic) complexity to be found in German. The comprehensibility of these sentences had been thoroughly checked on large groups of speakers before the test was devised. These sentences also include the whole range of possible German syllables and a substantial number of consonantal clusters. Speech editing gave us two sets of sentences to be played to the informants. The first set contained the sentences with CV pivots only, and the second set contained the same sentences in which the parts of the pivot were removed. As an example consider the two versions of a sentence *Sie ist nicht leicht zufrieden zu stellen* 'She is not easy to please':

[ʒi ɪst nɪçt laeçt tsufridən tsu ʃtɛlən]	-- unedited version
[zi ɪ nɪ lae tsu ridə tsu tɛlə]	-- CV version
[ɪ ɪst ɪçt aeçt uf i ən u ʃ ɛ ən]	-- VC version

We randomised the order of these sentences and played each sentence from one set (CV and VC sets, respectively) to 20 native speakers of German and asked them to report on the comprehensibility of the sentences.

The results of this experiment clearly show that the comprehension of sentences of varying syntactic and phonological complexity is significantly better in the case where these sentences are presented in their pivotal - CV form (87.3% correct scores) - than when these pivots are missing from the string (39.8% correct scores).

Furthermore, the syntactic complexity of a sentence does not influence comprehension. Our results show once again that the derivational theory of complexity may not be maintained.

Experiment 2

The hypothesis underlying the Pivot Parser strongly predicts that the CV transitions, which are the only parts of the phonological pivot, are comprehended more quickly and more precisely than these parts of the string which are outside of the pivot, for example the syllable final consonants. This should obviously be the case even if these syllable final consonants *precede* the CV pivot.

In order to check this prediction we asked the informants (the same group as in experiment 1) to tell us the word which they thought was phonetically most similar to the set of non-words. The idea behind this experimental design was that each of the non-words in the stimuli set corresponded phonetically to a minimal pair of existing words, and where one of the members of the pair was similar within the CV pivot and the other pair member was similar outside of the pivot, but the similarity point al-

1. Cf. Edwards [6], Dressler [4].

2. Cf. Clements & Keyser ([1], 19-23, 28ff.);Ohala & Kawasaki ([10], 115-119); Kelso, Saltzman & Tuller ([9], 5Off.); Dogil & Braun [3] for the most recent treatments of CV's status in phonetics and phonology.

3. The presentation here is a rough outline of the Pivot parsers recognition strategy. More detailed examples will be provided in the section 'The PIVOT as word recognizer'.

ways preceded the CV pivot. Table I gives you some of the examples from this non-word set together with the minimal pair set, which we tried to elicit, and the most frequent responses of the informants.

TABLE I

Non-word stimuli	Words (expected results)	Most frequent results
ELDE	ENDE - ELFE - ENTE	ENDE
WEPTE	WESPE - WESTE	WESTE
ALVEN	ALGEN - ALBEN	MALVEN
MAFTEL	MANTEL - MANDEL	MANTEL
WEKPE	WESPE - WESTE	WESPE
ELPE	ELFE - ELCHE	ELBE
RAUSCHPE	RAUSCHTE - RAUCHTE	RAUPE
RESKE	RESTE - RECHTE	FRESKE
WÄRLER	WÄRTER - WÄCHTER	WÄHLER
GÄNCHE	GÄNSE - GEMSE	KÄNNCHEN
RAULTE	RAUCHTE - RAUSCHTE	RAUTE
MANSEL	MANTEL - MANDEL	MAMSEL

If the informants chose the similarity within the CV pivot - for example, if they reacted with ENDE (end) to the stimulus ELDE - this would mean that they were using the pivotal recognition strategy rather than the left-to-right strategy. If latter were the case they should react with ELFE (elves) to the stimulus ELDE.

A tendency which was noticed was that the informants were not really trying to match the non-word with the words in which one CV pivot was identical, but showed strong preference for choices where both CV's matched. For example, the stimulus RESKE, which we expected to activate such words as RESTE or RECHTE, actually activated a word FRESKE in which both CV pivots match the stimulus. The same was the case with the non-word stimulus WÄRLER, which did not activate either the word WÄRTER (predicted by the left-to-right matching strategy) nor the distant word WÄCHTER, but rather the word WÄHLER which almost matches the non-word at both pivots. The only difference between the non-word WÄRLER as it was pronounced by the instructor and the response WÄHLER was in the length of the vowel, property which is not distinctive in german phonology. A similar case holds for numerous other pairs like: GÄNCHE - KÄNNCHEN, RAULTE - RAUTE, RAUSCHPE - RAUPE, MANSEL - MAMSEL.

THE PIVOT AS WORD RECOGNIZER

The PIVOT PARSER predicts that some parts of the string - the CV-pivots - are perceived more precisely and more exactly than other, non-pivotal, parts of the string. How is this *CV parser* supposed to work? We mentioned some general principles in the third section of this paper (The phonological PIVOT). The details will be illustrated immediately below.

Let us suppose the CV parser is confronted with the word *donkey* [dɔŋkɪ]. As a first step the speech envelope of this word will be stored in the form given in the figure below.

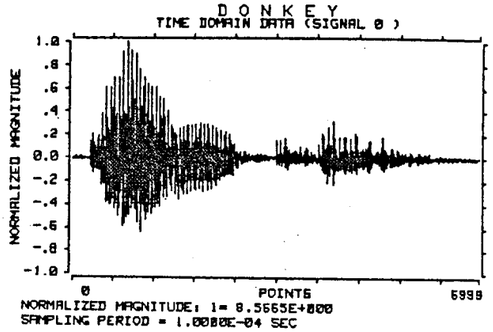


FIGURE 1: Speech envelope of the word *donkey*.

The signal will be transformed into its intensity tracing. This intensity tracing is the input to the segmentation algorithm. As we already mentioned in the section 'The phonological PIVOT' our aim is to sample acoustic information at a point of a transition between a consonant and a vowel within a CV. Stevens, in a series of experiments on acoustic cue recognition (Stevens, [11]), has provided convincing evidence for the perceptual importance of acoustic events in the vicinity of the 'consonant-vowel boundaries'.

> ...these brief time intervals when there is a rapid change in spectrum or amplitude create regions that are rich in information concerning the phonetic features in an utterance. (...) it would appear that a great deal of information is carried by these one-eighth-inch time slots in the spectrogram - much more than one would expect on the basis of the space they occupy in linear time.
>
> (Stevens, [11]: 253)

Moreover, these 'consonant-vowel boundaries' are relatively well marked by the speech producing system. At places where they occur there is usually an abrupt change in the amplitude. This change has been often considered (and used) as a cue to a boundary between individual speech sounds within an utterance. Our approach to these regions of abrupt amplitude change is quite different. We consider them as landmarks of a segmenting algorithm which considers them as *centers* (pivots) of units to be used in speech recognition.

Such an algorithm is being developed at the University of Bielefeld by Dafydd Gibbon (Gibbon, [8]). Given the intensity tracing as in the figure above, it automatically fixes these points where the most abrupt changes occur. Obviously, we are interested only in these changes where the amplitude jumps (characteristic of CV) not where it makes a dip (this is the characterization of the VC transition). Having fixed the first CV transition region we start sampling spectral information in its vicinity. We fix the Hamming window of the length of about 20 msec. and center it around the transition area. We suggest that the mechanism which samples spectral information should never leave the transition region. The only method which guarantees this is to make stepwise growing spectra with the transition point (the bar of the segmenting algorithm) as the center. The figure below illustrates this method.

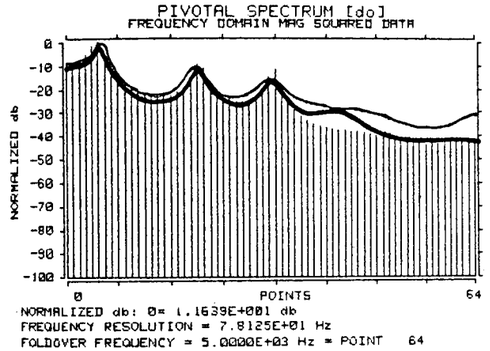
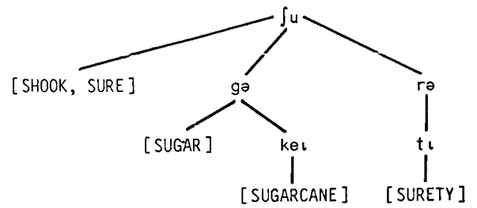


FIGURE 2: Pivotal spectra selected around the transition from [d] to [o] in *donkey*. The bars indicate the central spectrum, the thick line the next largest one, and the thin line the largest one.

These 'pivotal spectra' will be the input to the speech decoder. The difference to the usual procedure is that the acoustic information from the transition area is always present within the spectrum. What changes is only the amount of information which is sampled in the immediate (left and right) neighbourhood of the transition. We believe this change of perspective to be important, particularly from the point of view of 'time normalization' in speech recognition systems. It might turn out that spectra corresponding to different time-size windows around the transition correlate with various speech rates. If this were the case, we would have had a mechanism of encoding various speech rates in the spectral matrix itself.

The [do] of *donkey* has been, thus, decoded, and it is forwarded to the lexical recognition procedure. The lexicon on which this procedure will be stimulated is the 20.000 word phonetic lexicon of American English (Pocket Merriam-Webster, cf. Zue [12]). We have coded it in such a way that each lexeme is stored as the series of CV syllables that it contains. The lexical search itself is taken care of by a programme in PROLOG. This programme searches up lexical trees, which are organized in such a way, that each non-terminal node contains a CV syllable, and each terminal node contains word (or words) which can be made up of this CV syllable, and other CV syllables combinable with it. Such a tree is much easier to illustrate than to describe (after all illustration is what trees are for). The tree in the figure below illustrates the words starting with the CV syllable [ʃu].

FIGURE 3: Lexical subtree for words containing [ʃu] as their first CV syllable.



To return to the recognition of the word *donkey* - we have decoded the first CV [do]. The Pivot parser in PROLOG will give us a tree with this CV at the top, and all possible CV's which may be combined with it will be its daughters on the tree. All in all, they form a cohort of 45 words. However, as soon as the speech decoder decodes the second CV syllable of *donkey* - the [kɪ] - only one word remains: DONKEY.

This recognition procedure is very fast, and the reduction of initially large cohorts is quite optimal. It seems to be the case that the CV syllables do not combine so freely to form words as one would imagine they should. We tried out this recognition procedure on a number of words, and we never got really bad results. Consider the well-ploughed example *trespass*. After decoding the first of its CV syllables [trɛ] we are confronted with 21 word candidates, but having decoded the [pa], we immediately recognize *trespass*. Even in complex cases, where the division of the string into the CV syllables is difficult, and where there are many other consonants between the pivots, recognition is very fast, and, actually, unique. The word *abstract* [æbstrækt] is such an example. Out of its five consonants only one is decoded by the parser - the [tr] of the second CV syllable [træ].[4] Still the blank CV's - [æ] and [træ] - suffice to reduce the cohort into the following words - ABS'TRACK, 'ABSTRACT, ABSTRACTS and ABSTRACTION!

We have shown here that the parsing strategy of the PIVOT, when applied to words as heard in isolation, enables very fast and efficient recognition. This is obviously true mainly of polysyllabic words, the monosyllabics are a problem. Who would, however, want to stop half-way and consider recognition of words spoken in isolation?! The real test for any model of speech parsing is the recognition of connected speech. Let us see what the PIVOT has to offer in this area.

THE PIVOT AS PARSER OF CONNECTED SPEECH

We decided to make use of the apparently limited 'combinability' of CV syllables by giving the parser not just words, but the whole utterances in their PIVOTAL - CV form. Actually, PROLOG's command for this subroutine is - *get_sentence*. Thus, when we fed our parser with the string like the following:

get_sentence ([wɔ, də, ðɪ, pɹ, və, du], X).

4. Note that [tr] is a monosegmental affricate.

one sentence, with variation at two structural positions was our result. Incidentally (accidentally), the first of the 'possible' variants is the sentence that we were aiming at: *WHAT DOES THIS PIVOT DO*. Note, the combination of six CV syllables was analysed into one, single sentence with only slight variation in two positions. In the input we skipped a number of consonants (codas), we did not mark any boundaries between words, and we did not use any repair strategy - neither syntactic nor intonational nor semantic, nor frequency of cooccurrence. It was just the CV PIVOTs which were matched with the lexicon! If you consider the size of the lexicon (20.000 words), this result clearly speaks for the fact that the PIVOT is not the worst of the connected speech recognizers. It is also not one of the slowest! Although PROLOG is not the fastest of the 'intelligent' languages and although the machine on which it is implemented, does not do any MIPS, the simulations described here are all a matter of milliseconds. We tried a couple of other sentences. We will not give you them all, but the consideration of the modest one - *THIS PIVOT SIMULATES HUMAN PERCEPTION* - will give you the idea where the problems lie. The PIVOTAL input form for this sentence is : [ðɪ, pɪ, və, sɪ, mju, leɪ, hju, mə, pə:, sɛ, ʃə]. As an output we got 16 sentence analyses and the last one[5] was as the following:

[[THIS][PIVOT][SIMULATE][HUMAN, humid, humus][PERCEPTION]]

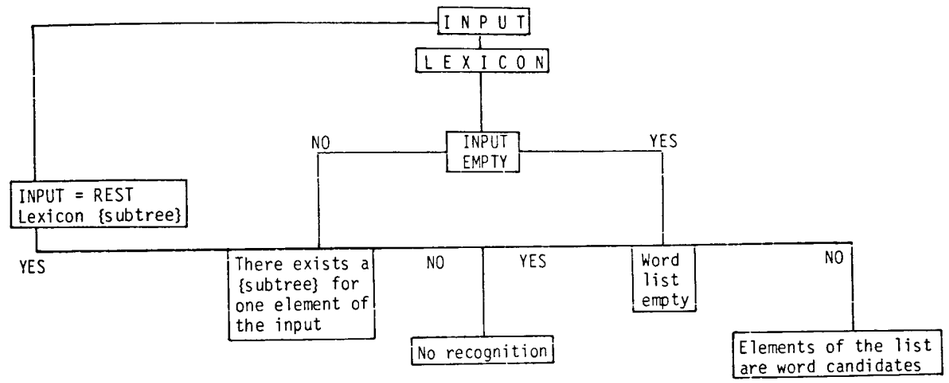
As you see, we do not need some sort of top-end parsing to get the optimal reading out of this output (i.e. to eliminate HUMID and HUMUS as possible word candidates). A quick look at other 15 alternatives makes it also apparent that some syntactic parser would be of great help in eliminating most of these analyses. However, the work on parsers for syntax, semantics and other 'higher' knowledge systems has advanced so much, that we do not doubt that they can help us. What is much more important is, that in this, and in every other case of simulated speech recognition which we have carried out, it was always the case that all the words in an utterance were recognized in at least one of the output analyses.

5. The PIVOT, as it is implemented now, tries to build *shortest* possible words as soon as the allowable CV combination has been found. The rest of the possibilities are found by backtracking. This need not actually be the optimal solution. In our example under discussion the *longest* allowable CV combination leads to the best results. The problem as to which of the combinations should be analysed first, is an empirical problem, which can be satisfactorily answered only after numerous simultations and psycholinguistic experiments have been carried out and analysed. At any rate, the PIVOT belongs to the 'no alignment' class of recognition theories. These theories do not pretend to 'know' where the boundaries are in the signal, but they single out some speech events (e.g. CV or a distinctive feature) and allow it to combine with any other speech event of the same time. The boundaries arise through constraints on the combinability of these units (CV's in our case).

This gives us a guarantee that the bottom-up PIVOT speech decoder and recognizer may be considered to constitute a fast, efficient and *sufficient* input to the top-end parsing strategies. As far as we can see, it is the optimal 'ear' for speech recognition.

In summary, and for those of our readers who appreciate pictures more than text, we give a graphic sketch of the lexical recognition procedure which we tried to describe in words in this paper.

TABLE II



Input : List of CV's
Lexicon : {subtrees}
Subtrees : [CV, WORD LIST, SUBTREES*] OR [ ]
Word list : [(WORD*)]

CONCLUSION

The most general conclusion of the PIVOT PARSER, and the one which makes our research programme distinct from all other approaches to language processing, is, that our parser does not require the exhaustive processing of strings, and that it explicitly claims that all language processing is based firstly and foremostly on the prototypical, unmarked units, which we called PIVOTS. Wheather our choice of the CV as the phonological PIVOT (the prototypical phonetic gesture) is correct or not, is, given the plausibility of this most general conclusion, only of secondary importance. However, the strong support that CV gets from the work done within the theory of Natural Phonology is an additional argument to consider it a prototypical speech event.

Although it is true that the event theory in phonetics is just at its beginnings (cf. Fowler [7]), and the event theory of phonology is only emerging (out of some ideas in the Natural Process Phonology, cf. Dressler [5]), the enterprise of replacing the segment oriented approach with an event oriented approach may prove highly rewarding in the studies of speech.

REFERENCES

[1] Clements, G.N. & S.J. Keyser (1983). *CV phonology*, Cambridge, MA., The MIT Press.

[2] Dogil, G. (1985). *Theory of Markedness in nonlinear phonology*, Habilitationsschrift, University of Bielefeld. (Available from the author)

[3] Dogil, G. & G. Braun (1986). *The Pivot Model of Speech Parsing*, distributed by LAUD (Linguistic Association University of Duisburg)

[4] Dressler, W.U. (1984). Explaining Natural Phonology, *Phonology Yearbook* 1, 29-53.

[5] Dressler, W.U. (1985). *Morphonology: the dynamics of derivation*, Ann Arbor, Karoma Press.

[6] Edwards, M.L. & L.D. Shriberg (1983). *Phonology*, San Diego, College Hill Press.

[7] Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective, *Journal of Phonetics* 14, 2-28.

[8] Gibbon, D. (1986). Prosodic parsing with parallel sequence and hierarchy incrementation (PSI/PHI), ZiF Workshop on Speech Parsing, October 15-17, Center for Interdisciplinary Research (ZiF), Bielefeld.

[9] Kelso, J.A.S., Saltzman, E.L. & B. Tuller (1986). The dynamical perspective on speech production: data and theory, *Journal of Phonetics* 14, 29-59.

[10] Ohala, J. & H. Kawasaki (1984). Prosodic phonology and phonetics, *Phonology Yearbook* 1, 113-129.

[11] Stevens, K.N. (1985). Evidence for the role of acoustic boundaries in the perception of speech sounds, in V. Fromkin (ed.), *Phonetic linguistics: Essays in honour of Peter Ladefoged*, New York, Academic Press, 243-257.

[12] Zue, V. (1983). Proposal for an isolated-word recognition system based on phonetic knowledge and structural constraints, in Cohen & Broecke (eds.), *Abstracts of the Tenth International Congress of Phonetic Sciences*, Dordrecht, Foris, 299-307.

PHONETICS AND NATURAL PHONOLOGY

WOLFGANG U. DRESSLER

Institut für Sprachwissenschaft
Universität Wien

ABSTRACT

Introduction to topics and principles of mutual relevance for phonetics and the model of Natural Phonology (NPh)

Among various concrete and/or phonetically oriented phonological theories (4, 10, 27, 33, 50) there is the model of Natural Phonology (NPh) as founded by Stampe (48) and later developed in various directions (49, 11, 13, 3, 8, 18, 19, 21, 38, 39). The selection of our topic for a symposium gives evidence for the congress organizers' recognition that phonetics and NPh hold a particular relevance for one another. In this introduction to our symposium I want to briefly explore this mutual relevance.

1) NPh deliminates the field of phonology in a way which facilitates cooperation with phoneticians:

a) Boundaries to morphonology are defined rather sharply (notwithstanding diachronic transition phenomena), sharper than e.g. Lexical Phonology (40): phonological processes of a given language represent constraints on production and perception (cf. already 5, 46), morphonological rules do not. Thus both phonemic and allophonic processes of final obstruent devoicing in many languages (e.g. German and Polish) are phonological, whereas Russian final devoicing may approach a transitional stage towards morphonology (cf. final voiced stops in bab ! and abbreviations such as MID). The consequences for phonetic experimentation are obvious.

b) No boundaries are set towards phonetics, there is no despise of 'low level rules' or of 'performance factors' (note NPh work on speech errors as in 36, 22), i.e. there is no puristic dichotomy between phonology and phonetics as often in structural phonemics or Natural Generative Phonology (50). Therefore the division between phonetics and phonology is rather a division of labor.

2) In contradistinction to Generative Phonology (2, 25) NPh acknowledges and investigates the extralinguistic bases of its principles (18, 19,9). These bases lie generally in phonetics, but also in neurological, psychological, and social phenomena, and in general properties of semiosis. Thus phonetics is a founding discipline of phonology. This comes out in NPh as a great concern for phonological universals.

3) NPh and phonetics share a functional perspective (34, see Basbøll and Bertinetto at this symposium).

a) In order to avoid proliferating functional ad hoc assumptions it seems useful to subsume functional considerations into a semiotic metatheory (18, 19; for other semiotic approaches to phonology see 31, 1, 47). Clearly all phonetic acts involved in production and perception are parts of semiosis, because acts of semiosis (or sign chains) mediate between "substance and form". In this perspective the pernicious dichotomy between sound substance as domain of phonetics and sound form as domain of phonology can be eliminated.

b) Any functional explanation has to account for functional conflicts: NPh has highlighted two contrasts well-known to many linguistic and phonetic schools for a long time:

A) phonological-phonetic vs. morphological tendencies, whose diachronic version has been heralded by the Neogrammarian "Lautgesetz und Analogie" (cf. now 45, 19, 51, 30);

B) tendencies towards "ease of articulation" vs. "optimal perception". NPh has elaborated on a classification of opposed universal phonological process types of fortition (foregrounding) vs. lenition (backgrounding) with opposed hierarchies (see 12, 24, 19). The single process types can be classified according to the following antagonistic subgroups: processes of dissimilation (polarization incl. dipthongization) vs. assimilation (incl fusion), strengthening vs. weakening, lengthening vs. shortening, insertion vs. deletion. The calculation of these conflicts needs psycholinguistic, sociolinguistic models, and phonetic operationalization (see I below).

c) function-form relations are typically many-to-many, as can be seen in overlapping phonetic explanations of e.g. vowel lengthening before voiced stops: This draw-back of functionalism unfortunately makes falsifiability more difficult (cf. 19). On the other hand functional considerations can prevent the assumption of ad hoc hypotheses (cf. Bertinetto at this symposium).

4) NPh does not assume an abyss between competence and performance or langue et parole which would be difficult to bridge. Thus NPh has been open to (and directly attackable by) experimental evidence, even before Ohala's important design of Experimental Phonology (cf. now 43). Nor does NPh restrict its domains to the analysis of citation forms (of native informants) or spelling pronunciations (of grammars and dictionaries). Since its very beginnings NPh has emphasized the importance of studying first language acquisition (48, 24), phonological variation (49, 14, 15, 20, 23, 26, 41) as well as other fields of external/substantive evidence (16, 18, 19, 8, 37, cf. 53), e.g. second language acquisition where NPh can better cooperate with phonetics than other models (cf. 28, 29, Dziubalska at this symposium).

Among other specific fields of common inquiry for phonetics and NPh I would like to single out the following ones:

I) Concepts of "ease of articulation" and "optimal perception" (3bB) have been shunned as too vague by many investigators. Lindner's (35) model of context free and context sensitive articulation gestures represents an important step forwards for the first concept. Clearly "ease" may mean either "local" application of one given phonological process in cross-linguistic and universalist studies, e.g. hierarchies of consonant palatalization or vowel centralizations and deletions irrespective of the phonotactic results, or "global" effects in terms

of prosody or consonant/vowel clusters (e.g. avoidance or elimination of consonant clusters resulting from vowel deletion).

More detailed phonetic investigations are needed to help phonologists differentiate process types, e.g. is intervocalic lenition of stops a weakening or an assimilation process ? Or should we rather envisage overlapping parameters ?

Optimization of perception ist studied by Angenot (partially referred to at this symposium, cf. also Basbøll and Dogil).

II) Following Beaudouin de Courtenay (5, cf. now 37, 9, 42) phonemes are defined as sound intentions within NPh. This entails selection of phonemes to occur on a more conscious level than other acts of phonological semiosis. One of many consequences can be seen in phonological variation: If speakers select a dialect phoneme instead of a standard phoneme then this choice can be much better controlled by speakers (and perceived/evaluated by hearers) than other phenomena of dialect mixing (23, 41). However this does not entail that phonetic realizations are entirely unconscious, automatic consequences, as studies of coarticulation have shown. Since all phonetic acts are parts of semiosis, all-or-nothing distinctions between phonemic intentionality and unintentional application of phonological processes must be wrong from a deductive point of view as well. However the fine-grading of "consciousness" or "intentionality" awaits further empirical elucidation. E.g. should one draw a clear line between segment deletion and physical articulatory undershoot ?

III) A similar gradience must be allowed for phonetic-phonological realization phenomena in studies on automatic speech processing/recognition (cf. e.g. 6). Such empirical studies may lead to a radical reformulation of some phonological processes, at least at "lower levels" (cf. Dogil at this symposium).

IV) NPh is the only model of process phonology where considerable attention has been paid to speech disturbances (17, 24, 52, 32, 44, 22). Here as well a similar gradience comes into play (cf. Dogil's unpublished work on aphasia vs. dysarthria).

Of all these and other interesting matters only some can be discussed in this symposium. More at the next Phonologietagung ! (Sixth International Phonology Meeting, Krems, july 1-4, 1988, organized by the Institut für Sprachwissenschaft der Universität Wien)

REFERENCES

(1) H.Andersen, "Phonology as semiotic", A Semiotic Landscape, ed.S.Chatman, Mouton, 1979, 377-381.

(2) S.Anderson, "Why phonology isn't 'natural'", Linguistic Inquiry 12,1981,493-540.

(3) J.-P.Angenot et al. ed., "Studies in Pure Natural Phonology and related topics", Univ.Federal de Santa Catarina Working Papers in Linguistics.

(4) C.-J.Bailey, "On the yin and yang nature of language", Karoma Press, 1982.

(5) J.Baudouin de Courtenay, "Versuch einer Theorie phonetischer Alternationen", Trübner, 1895.

(6) J.Bernstein, M.Kahn, T.Poza, "Speaker sampling for enhanced diversity", ICASSP 4,4,1985,1553-6.

(7) P.Bjarkman, "Towards a proper conception of processes in Natural Phonology",PCLS 11,1975,60-72

(8) D.Churma, "Arguments from External Evidence in Phonology", Garland, 1985.

(9) D.Churma, "On explaining the phoneme: Why (some of) phonology is natural", BLS 11,1985,25-38.

(10) D.Dinnsen, "Phonological rules and phonetic explanation", Indiana Univ. Linguistics Club 1978.

(11) P.Donegan, "On the natural phonology of vowels", Ohio State Univ Working Papers in Ling. 23,1978.

(12) P.Donegan, D.Stampe, "On the description of phonological hierarchies",CLS Book of Squibs 1977,35-8

(13) P.Donegan, D.Stampe, "The Study of Natural Phonology", D.Dinnsen ed. Current approaches to phonological theory, Indiana Univ. Press 1979, 126-173.

(14) W.Dressler,"Allegroregeln rechtfertigen Lentoregeln" , Innsbruck 1972.

(15) W.Dressler, "Methodisches zu Allegroregeln", Phonologica 1972, Fink, 1975, 219-234.

(16) W.Dressler, "Arguments and non-arguments in phonology: on the use of external evidence", PICPhS 9,1979,93-100.

(17) W.Dressler, "A classification of phonological paraphasias", Wiener ling.Gazette 29,1982,3-16.

(18) W.Dressler, "Explaining Natural Phonology", Phonology Yearbook 1,1984,29-51.

(19) W.Dressler, "Morphonology", Karoma Press, 1985.

(20) W.Dressler, J.Hufgard, " Études phonologiques sur le breton sud-bigouden", Österr.Akademie der Wissenschaften, 1980.

(21) W.Dressler, L.Tonelli eds. "Natural Phonology from Eisenstadt", CLESP, 1985.

(22) W.Dressler, E.Magno, L.Tonelli, "Phonologische Fehlleistungen und Paraphasien im Deutschen und Italienischen", GLS 26,1986,43-57.

(23) W.Dressler, R.Wodak, "Sociophonological methods in the study of sociolinguistic variation in Viennese German",Language in Society 11,1982,339-370.

(24) M.L.Edwards, L.Shriberg, "Phonology", College Hill Press, 1983.

(25) E.Gussmann, "Naturalness, morphonology, and the Icelandic velar palatalization", NJL 7,1984,145-63

(26) T.Herok, L.Tonelli, "Natural process phonology and the description of phonological variation", Wiener ling.Gazette 16,1977,43-63.

(27) J.B.Hooper, "An introduction to Natural Generative Phonology", Academic Press 1976.

(28) B.Hurch, "H. Eine Studie zur kontrastiven Phonetik und Phonologie", Parallela,Narr,1983,271-80.

(29) B.Hurch, "On aspiration and markedness or restriction on limitations", (21) 61-72.

(30) Interplay of phonology, morphology, and syntax, Chicago Linguistic Society, 1983.

(31) R.Jakobson, L.Waugh, "The sound shape of language", Indiana Univ.Press 1979.

(32) M.Kilani-Schoch, "Processus phonologiques, processus morphologiques et lapsus dans un corpus aphasique", Lang, 1982.

(33) S.Kodzasov, O.Krivnova, "Sovremennaja amerikanskaja fonologija", Izd.Moskovskogo Univ. 1981.

(34) B.Lindblom, "On the teleological nature of speech processes", Speech Comm.2,1983,155-158.

(35) G.Lindner, "Der Sprechbewegungsablauf", Akademie-Verlag, 1975.

(36) E.Magno, L.Tonelli, "Syllabic constraints on phonological speech errors in Italian",(21) 73-88.

(37) A.Manaster-Ramer, "The Ułaszyn-Halle paradox and modern phonology", (21) 89-100.

(38) E.Mayerthaler, "Unbetonter Vokalismus und Silbenstruktur im Romanischen", Niemeyer, 1982.

(39) W.Mayerthaler, "Markiertheit in der Phonologie", T.Vennemann ed. Silben, Segmente, Akzente, Athenaion, 1982, 205-246.

(40) K.Mohanan, "The Theory of Lexical Phonology", Reidel, 1986.

(41) S.Moosmüller, "Diskrimination und Einschätzung des dialektalen input-switch /a/ <--> /ɔ/", Wiener ling.Gazette 35-36,1985,75-94.

(42) G.Nathan, "Phonemes as mental categories", BLS 12,1986,212-223.

(43) J.Ohala, J.Jaeger ed., Experimental Phonology, Academic Press, 1986.

(44) W.Perkins ed. Phonological-articulatory disorders, Thieme, 1983.

(45) E.Ronneberger-Sibold, "Sprachverwendung - Sprachsystem: Ökonomie und Wandel",Niemeyer 1980.

(46) E.Sapier, "La réalité psychologique des phonèmes", J de Psych.Norm.et Path. 30,1933,247-265.

(47) M.Shapiro, "The sense of grammar", Indiana Univ.Press 1983.

(48) D.Stampe, "The acquisition of phonetic representation", PCLS 5,1969,443-454.

(49) D.Stampe, "A dissertation on Natural Phonology", Garland, 1980.

(50) T.Vennemann, "Phonology as non-functional non-phonetics", Phonologica 1980,Innsbruck 1981,391-402

(51) W.Wurzel,"Phonologie - Morphonologie - Morphologie", Akad.Wiss.DDR, 1982.

(52) W.Wurzel, R.Böttcher, "Konsonantenkluster, phonologische Komplexität und aphasische Störungen", M.Bierwisch ed. Psychologische Effekte sprachlicher Strkturkomponenten, Akademie-Verlag,1979, 401-445.

(53) A.Zwicky, "The strategy of generative phonology", Phonologica 1972,Fink, 1975, 151-168.

REVIEW OF FRENCH WORK ON VOCAL SOURCE - VOCAL TRACT INTERACTION

R. CARRE

Laboratoire de la Communication Parlée, ICP Unité Associée au CNRS,
INPG-ENSERG, 46 Avenue Félix Viallet, 38031 GRENOBLE CEDEX

ABSTRACT

The French research work on vocal source - vocal tract interaction was mainly developped in Grenoble and began as early as 1974 with various studies on vocal tract modeling and on source modeling. We have studied both detailed and simplified models of source and tract in order to assess the interaction effects. Modeling of the subglottal cavities has been also carried out. Today, various spectral analysis allow us to complete this approach, to have new data and to formulate new interpretations.

MODELING OF SOURCE - TRACT COUPLING

Vocal tract impedance loading the vocal source.

For a long time, the vocal tract impedance was considered to be low compared with the source impedance. Thus, time-varying vocal tract was considered having no influence on the source.

In reality, several studies do not support these simplification (FANT - 1960, HOLMES - 1976). The coupling phenomenon can be studied by computing the vocal tract input impedance seen from the vocal source. With a model of the vocal tract including all the different losses (wall vibrations, viscosity, heat and radiation), we may suppose that this model gives a good approximation of the vocal tract impedance. The coupling effect must be maximum for high values of the vocal tract impedance. This impedance was calculated for the first three formants and for different French vowels (MRAYATI, GUERIN, BOE - 1976). The resistive part of the input impedance of the vocal tract is given in table 1 for the first two formants.

Judging from the resistive part of the input impedance at resonance, the coupling would, at first, affect the first formant of /a/, /ɔ/, /ɑ/, and the second formant of /ɑ/, /i/.

Modeling is a useful tool for studying the coupling phenomenon ; we therefore developed a simplified model of the vocal tract load impedance, suited for simulations in the time domain of the glottal flow shape. Foster type circuits were introduced, each representing the properties of a formant (GUERIN, MRAYATI, CARRE - 1976). Such simulations with one or two formant circuits have generally been used in these studies.

A simplified representation of the vocal tract impedance is given in figure 1. The elements (L1, R1, C1, L2, R2, C2) of the two Foster circuits are related with the formant frequencies F1 and F2.

This impedance was coupled with a two mass model of the vocal cords (ISHIZAKA, FLANAGAN - 1972), and the effect of the coupling was studied on different parameters :

- the fundamental frequency of the vowel (GUERIN, DELOS, MRAYATI - 1976, GUERIN, BOE - 1980) : the coupling cannot take into account the intrinsic pitch of the vowels, its effect is opposite ;
- the intensity of the vowels (GUERIN, BOE - 1977) : the coupling effect is negligeable ;
- the formant frequencies and the bandwidths (CARRE, GUERIN - 1980) : when coupling is strong, the formant frequency increases due to the glottal inductance effect ; on the other hand, the bandwidth increases due decreasing source resistance.

The fundamental frequency variations are given in figure 2 for different vowels.

Modeling the subglottal cavities (AL ANSARI, GUERIN - 1981) has small effects on the fundamental frequency, but greater effects on the dissymmetry factor and the opening factor of the glottal flow shape.

Finally, the dynamic impedance of the two mass model was measured (CARRE, GUERIN - 1980) : the impedance value corresponding to an average glottal area is not representative of the dynamic

effect. The dynamic impedance of the source was found to be small enough to permit a strong coupling effect in the case of high load impedance. This effect was noticed on formant frequencies.

Figure 3 shows the evolution of the harmonic frequency number 6 of the source spectrum for a F0 increasing in relation with the variation of the vocal cords tension in two situations : two mass model (1) short circuited, (2) loaded by the vocal tract. In the case of the situation 1, harmonic amplitude evolution is linear. In the situation 2, the harmonic number 6 is attenuated at the place of the formant. This attenuation can be used to calculate the impedance of the source for this frequency value. The formant frequency increase, due to the coupling effect, can also be measured.

Moreover, a synchronisation effect on the highest harmonic around the formant can be observed. The two maxima around the formant correspond to a negative resistance and thus to an amplification. Is it a means for adjusting the formant on the most important harmonic ?

A simulation of a two beam model of the source does not improve very much the coupling modeling (PERRIER, GUERIN, AULOGE - 1981).

As a main result, the two mass model is always the best simple representation of the vocal cord behavior.

STUDY OF THE GLOTTAL FLOW SHAPE WITH COUPLING, MALE AND FEMALE VOICE

By inverse filtering, the opening factor of the glottal flow was studied and was considered stable whatever the coupling may be (CHENG, GUERIN - 1985). On the other hand, the dissymmetry factor is a decreasing non linear function of the first formant frequency.

A strategy was developped to calculate the fundamental frequency, the asymmetry factor and the opening factor of the glottal flow, from the coupling, for male and female voices. Three independant parameters are used (subglottic pressure, vocal cord tension and first formant frequency). Three dependant parameters are then obtained (fundamental frequency, opening factor and asymetry factor). High quality synthesis was obtained by this means.

STUDY OF THE VOWEL SPECTRUM FOR INCREASING F0

Stable vocal tract shape

The conditions of the coupling simulation for increasing F0 (two mass model and vocal tract loading) was reproduced for natural vowel production (CARRE - 1981a, 1981b). The vocal tract was kept as stable as possible by the speaker during the F0 variations. The stability was controled by closed glottis analysis. As in the simulation, we observed an increase of the formant frequencies and bandwidths for a strong coupling when comparing the transfer function of the vocal tract obtained for closed glottis conditions and the envelope of the harmonic components. Moreover, in some cases, the harmonic evolutions do not exactly follow the formant envelope when coupling is important. For example, for the vowel /a/, male voice, a deep of about 10 dB appears near the first formant for the harmonic number 6 (figure 4). Does this correspond to the synchronisation effect observed in the simulation ? The formant frequency could be synchronized on the highest harmonic. Formant measurements by autocorrelation LPC techniques have shown such an effect. It could be also due to the analysis method.

Natural vocal tract shape

When the larynx is moving freely during the F0 increase, the larynx generally rises up and an increase of the first formant frequency is observed when the source tract coupling is known to be important. In this case, the male vocal tract changes into a female tract size by reduction of the length on the source side (CARRE, LANCIA, WAJSKOP - 1968). A clear correlation between the impedance of the vocal tract load and the male/female formant frequency factors exists.

CONSEQUENCES IN PRODUCTION, ANALYSIS AND SYNTHESIS OF VOWELS

The results reported by SUNBERG (1982) on singing show that, for a better energy transmission, some adjustments (larynx displacements for an important source tract coupling, modifications of the tract shape...) are carried out to equate the formant and the biggest harmonic. This is specially true for female voice, for high F0. CHENG (1986) obtained the same results for synthetic vowels in the case of high fundamental frequencies : listeners prefer high F0 synthetic vowels for strongest harmonic closed to formant. With all the possibilities of adjustment, the synchronization effect described above could be added.

The formant analysis of the vowels produced by female speakers could have to be weighted by the fundamental frequency measurement. Such an interaction has to be taken into account for speech recognition. In the case of speech synthesis, the acoustic and physiological coupling effects cannot be ignored for a better synthesis quality. CHENG (1986) used special formant circuits, the damping of which being controled by the shape of the source signal. With such circuits, the synthesis quality is improved.

Diphone synthesis which is the result of sound concatenation obtained for a specific F0, is not well adapted to take into account the modifications due to various coupling.

CONCLUSIONS

We have now a better knowledge about the source-tract coupling : variations of vowel characteristics have been studied (intrinsic pitch and intensity, formant and bandwidth variations) and the two mass model is always the best simple representation of the vocal tract behavior.

Coupling characteristics (due to F0 variations) have to be taken into account for analysis and recognition. The perceptual importance of coupling characteristics have also to be tested for synthesis.

BIBLIOGRAPHY

AL ANSARI A., GUERIN B. (1981)
Effet du couplage source - conduit vocal sur les caractéristiques de l'onde de débit.
Actes des 12èmes Journées d'Etude sur la Parole, MONTREAL.

CARRE R. (1980)
Modèles auditifs et modèles de production de parole.
10th International Congress on Acoustics, SYDNEY.

CARRE R. (1981)
Vocal source - vocal tract coupling. Effects on the vowel spectrum.
Proceeding of the IV th. FASE Symposium, VENEZIA.

CARRE R. (1981)
Couplage conduit vocal - source vocale. Conséquences en analyse de parole.
Actes des 12èmes Journées d'Etude sur la Parole, MONTREAL.

CHENG Y.M., GUERIN B. (1985)
Introduction du couplage source - conduit vocal dans un synthétiseur à formants.
Actes des 14èmes Journées d'Etude sur la Parole, PARIS.

CHENG Y.M., GUERIN B. (1985)
Control Parameters in male and female glottal sources.
Fourth International Vocal Fold Physiology Conference, NEW HAVEN.

CHENG Y.M. (1986)
Etude du concept source - filtre interactif pour la synthèse de la parole. Analyse des voyelles nasales.
Thèse de l'Institut National Polytechnique, GRENOBLE.

DELOS M., GUERIN B., MRAYATI M., CARRE R. (1976)
Study of intrinsic pitch of vowels.
91th ASA Meeting, WASHINGTON.

FANT G. (1960)
Acoustic theory of speech production.
Mouton. THE HAGUE.

GUERIN B., BOE L.J. (1977)
Etude des caractéristiques acoustiques intrinsèques des voyelles françaises.
9th International Congress on Acoustics, MADRID.

HOLMES J.N. (1976)
Formant excitation before and after glottal source.
IEEE Conf. on Acoustics, Speech and Signal Processing, PHILADELPHIA.

ISHISAKA K., FLANAGAN J.L. (1972)
Synthesis of voiced sounds from a two - mass model of the vocal cords.
Bell System Tech. Journal, 51, 1233-1268.

MRAYATI M., GUERIN B., BOE L.J. (1976)
Etude de l'impédance d'entrée du conduit vocal. Couplage source - conduit vocal.
Acustica, 35, 330-340.

PERRIER P., GUERIN B., AULOGE J.Y. (1982)
Simulation d'un modèle continu de la source vocale.
Actes du Symposium FASE/DAGA, GOTTINGEN.

Vowels	/u/	/o/	/ɔ/	/ɑ/	/a/	/ɛ/	/e/	/i/	/y/
R-F1(Ohms)	50	128	275	330	265	102	95	38	50
R-F2	175	37	51	350	125	107	93	348	62

TABLE 1. Input impedance of the vocal tract at F1 and F2, for 9 different French vowels.

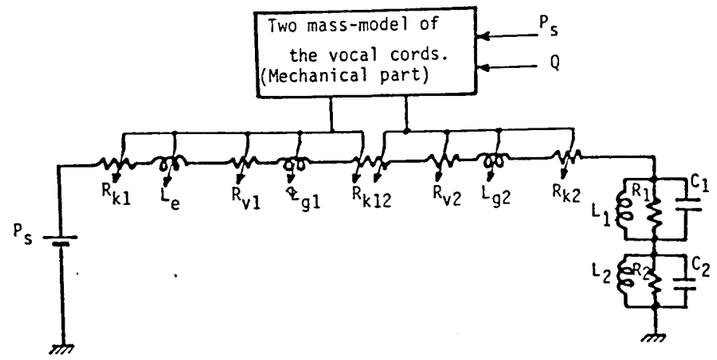


Figure 1. Simulation of the tract load impedance and a two mass model.

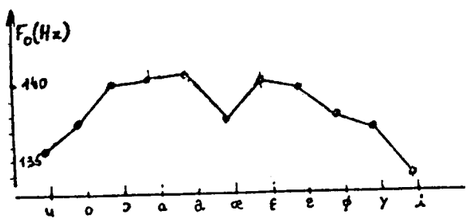


Figure 2. F0 variations for different vowels.

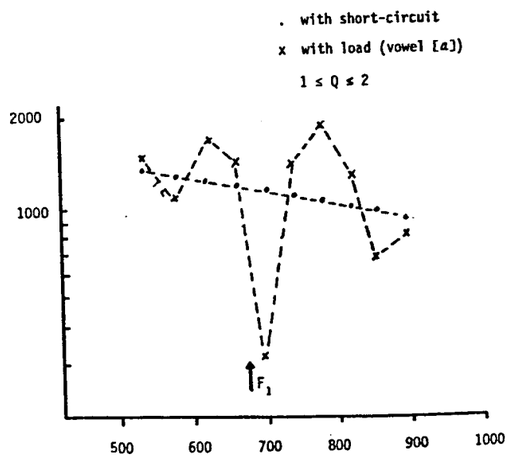


Figure 3. Variations of the harmonic amplitude for a two mass-model : when short circuited, when loaded by a vocal tract (vowel /a/).

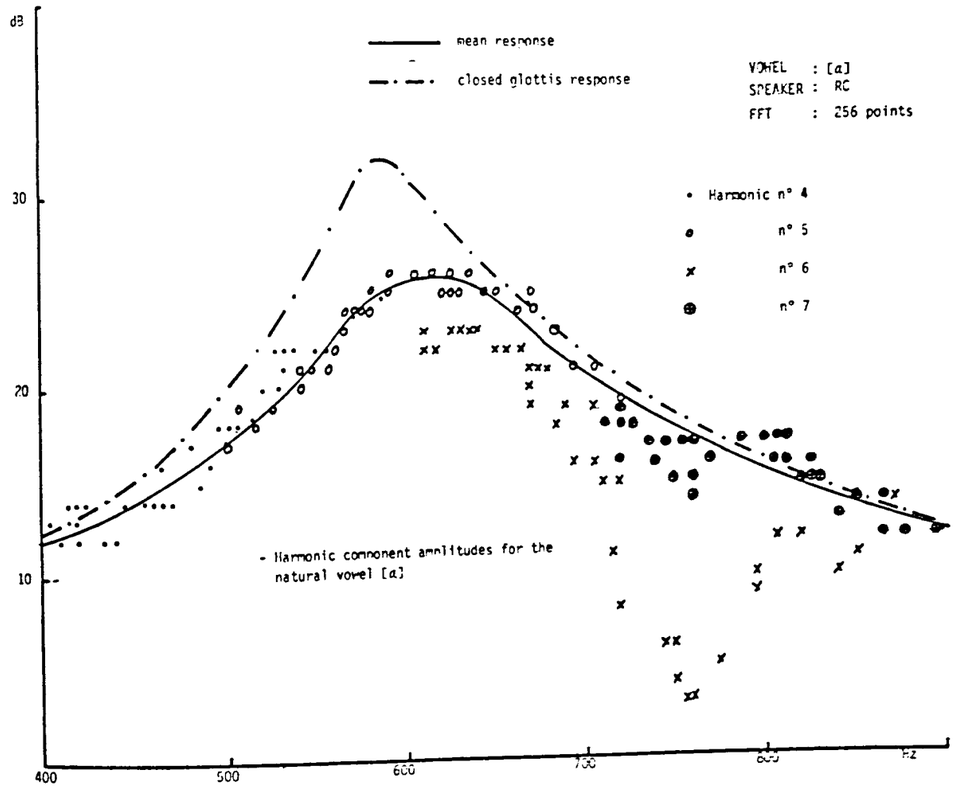


Figure 4. Harmonic component amplitudes for the natural vowel /a/ and F0 increase.

INTERACTIVE PHENOMENA IN SPEECH PRODUCTION

GUNNAR FANT

Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH), Box 70014
S-100 44 Stockholm, Sweden

ABSTRACT

A brief overview of interactive phenomena on several levels of speech production modeling has been attempted. Special attention has been devoted to the dependent covariation of phonation and articulation and the implications for a source-filter decomposition of speech. The growing insight in the voicing mechanism and voice source dynamics provides a broader basis for description of segmental as well as prosodic features.

INTERACTION AND THE SPEECH CODE

Speech production processes are inherently interactive in the sense that component parameters and descriptors seldom function independently. Interaction has thereby become a key word in theoretical issues concerning the speech code of almost the same dignity as variability and invariance and is related to these topics.

There exists a large literature on invariance and variability, e.g., the volume edited by Perkell and Klatt /1/. The various standings on these issues seem to reflect consequences of varying definitions and interpretations of terminology rather than true divergencies. I must admit that there exists a similar vagueness in the interpretation of the term interaction which I will use in a rather general sense. One obvious comment on the invariance issue is that we must first accept that phonological transformations and deletions frequently interact with the planning of an utterance, thus accounting for a deviant set of phonological representations in less precise speech. The expected phonemes may simply not be there. Apart from this extreme, there is a continuity of the extent of information-bearing properties in the speech signal ranging from weakly induced traces to the presence of well-defined phonetic segments and cues.

However, it is not fair to refute the invariance issue by reference to either missing or weakly manifested features or to the listeners' complimentary "top down" expectancy. The situation has been neatly summarized by Lindblom /2/ who points to comprehension as the ultimate level of invariance. Personally, I do not favor any specific definition of invariance but I feel it has an important role in the discussion of distinctive features /3/ in their literary sense as constituents of the full speech code. It is also important to point out the relational basis of features. To search for distinctive elements in the speech wave is not a matter of hunting for a very specific golden grain of information that should always be there. It is rather a matter of finding context-biased manifestations of relational contrasts. A feature is just as much a matter of what could have been present in the speech wave as what is actually found.

In practice, we thus have to make up for the frequent lack of direct invariance by resorting to a rule-oriented analysis of variabilities which in the far end preserves an output more or less appropriate for the specific situation with its constraints and demands. Production has to recruit a substantial amount of coordinated interaction within the system to accomplish its complex task.

How do we now define interaction? In a general sense, interaction is an interdependency of constituents of a descriptive system applicable to complex transformations and departures from linear orthogonal relations. A variation of one parameter usually implies a nonlinear influence on the values and variational limits of other co-varying parameters and the extent to which each of the parameters influences the final output.

The many-to-one and one-to-many relations between linguistic and acoustic entities, e.g., relating a sequence of phonemes or a bundle of phonological features to acoustic segments and events and vice versa /4/ has its parallel in transforming from one level of speech production to a previous or to a following one, e.g., from neuromuscular activity to articulatory movements and further on to vocal tract area functions, aerodynamical events and speech wave patterns.

The movement of a single articulator is generally an interplay of several muscle functions displaying synergism or antagonism, with a large allowance for individual variations, combining with other articulatory activity to preserve an adequacy of the final output. Sensory feedback adds to the complexity of interactions, see the contribution of V. Sorokin to this symposium. Let's hope that the now popular "action theory" /5-7/ will find a sufficient close tie to neurophysiology so that we at some stage may transform our present hypothetical generalizations into a more complete insight in actual speech motor behavior.

I have often complained about our lack of speech analysis data. For speech production modeling, the need is even greater. Research has dealt more with tracking of the movements of specific reference points of articulators than a mapping of complete time-varying area functions and aerodynamic states. There remains a great deal of work to map cavity dimensions and speaker specific topologies. We need more insight in the general relations of speech production and speech wave patterns, e.g., with respect to consonants. Cavity-formant relations are complex but these can be handled with appropriate models /8/. The lack of descriptive physiological data remains the bottleneck.

Speech production is a key to the understanding of the speech code. Speech production research is now enjoying a renaissance as a support of speech perception theory and also offers intriguing potentialities for a more natural articulatory-based synthesis. Although a complex of coordinated activities of several articulators may be involved in securing a specific auditory-perceptual effect, the opposite can also be true. What appears to be a complicated set of context-dependent, perceptually interacting segments and cues in the speech wave can often be related to a single production parameter.

An example is the role of a vocal fold adbuction-adduction gesture determining a sequence of associated events in the speech wave of an unvoiced stop including possible aspiration and preaspiration which we may contrast with a voiced stop. Presence or absence of a voice bar, the initial F0 and F1 at relase, F1 cut back and a shorter duration of a preceding vowel are all functions of one and the same underlying glottal gesture. Preaspiration usually terminates the preceding vowel prior to supraglottal closure inducing breathy termination of the vowel which may end with a consonantal occlusion noise.

Other factors than abduction-adduction may contribute. Thus, the phonemic contrast above may contain covarying elements along the tense-lax dimension. Anyhow, this example illustrates that what may seem quite complex in a pure perception-oriented analysis may have a simple correspondence on the production level. Such relations support motor theories of speech perception /9/.

Apart from top-down effects, I would prefer to conceive of speech perception not as a process of looking for equivalent production patterns but rather as involving direct responses to complex auditory patterns which we have learned to associate with linguistic entities. These may not entirely conform to speech-motor patterns, the full equivalence being reached at a higher level of message representation only. An association of auditory patterns to one's own motor capacity could be of importance in the learning stage /10/.

Compensatory modes of articulation have not been studied extensively. Compensation is never complete if we look at fine acoustic details but has to satify perceptual criteria.

The output-oriented function of speech production is often illustrated by the classical bite-block experiment of Lindblom, Lubker and Gay /11/. A speaker aiming at the vowel [i] compensates for an unnatural fixed high jaw opening by raising the tongue to an appropriate position. It is an open question whether the execution relies more on an invariant command for anchoring the tongue blade in a certain contact position than a recalculation of what the tongue has to do with respect to the jaw.

Coarticulation is generally a matter of complex interactions which might obscure the interpretation of spectrographic patterns. Thus, tracking the transition in the release phase of a labial stop, it might be hard to catch the initial delabialization cues and keep them apart from the more slowly progressing main tongue body movements. An insight in production mechanism is apparently at an advantage. Still it is questionable whether this reference also operates in normal speech perception.

Speech output norms vary with the situational demands. Vowel space shrinks in casual style and is expanded in "hyper-speech" modes (ref. /2/). This is analogous to the relation between unstressed and stressed vowels in Swedish /12/. A related observation is that of Zhang Jialu who in his paper for this session reports shifts of formant frequencies and F0 as a function of voice output level.

An adequate theory of prosody must take into account systematic interactions of stress and emphasis with most speech production parameters.

Words within sentence focus display an increased articulatory or "dynamic" contrast whereas unstressed words will be produced with less contrast between successive segments. With emphasis, vowels and unvoiced consonants increase in intensity, whereas voiced consonants display decreased intensity due to more effective constrictions. With emphasis, an otherwise voiced [ɦ] tends to loose its voicing due to a more extreme abduction of the vocal folds, and noise generation takes over. In a destressed position, a voice bar of a stop tends to turn into a semi-vowel with but little contrast to adjacent vowels /13/. Fig. 1 illustrates the degree of contrasts within a Swedish word "behålla", [behɔl:a], uttered in sentence focal position and prefocus. The oscillogram and the voice source excitation parameter display similar contours which bring out the difference in dynamic contrast. We are now engaged in more general studies of how voice source parameters enter prosody.

PHONATORY AND ARTICULTORY INTERACTION. THE HUMAN VOICE SOURCE

A decomposition of the acoustic stage of speech production into a source function and a filter function has a counterpart in the terms phonation and articulation but the correspondence is not perfect. The lack of coherence is in part a matter of terminology, in part a matter of physical interaction.

We may thus speak of laryngeal or glottal articulations as determinants of the voice source as well as of quality changes related to accompanying changes in vocal tract configurations (e.g., a "throaty voice"), or we could imply glottal stops. In connection with Fig. 1, we have already noted that a highly constricted supraglottal articulation impedes the glottal flow which causes appar-

ent changes in glottal pulse shape and intensity /14/. Furthermore, a glottal abduction induced by an [h]-sound or appearing at the boundary towards an unvoiced segment causes changes in formant frequencies and bandwidths in addition to changes in glottal pulse shape all of which contribute to the breathiness or the local aspiration. Thus, both articulatory and phonatory processes may influence the voice source whilst the filter function is determined by articulatory as well as by phonatory adjustments including lung pressure variations. The validity of the last statement, however, may depend on the particular definitions adopted for source and filter. These are not self-evident.

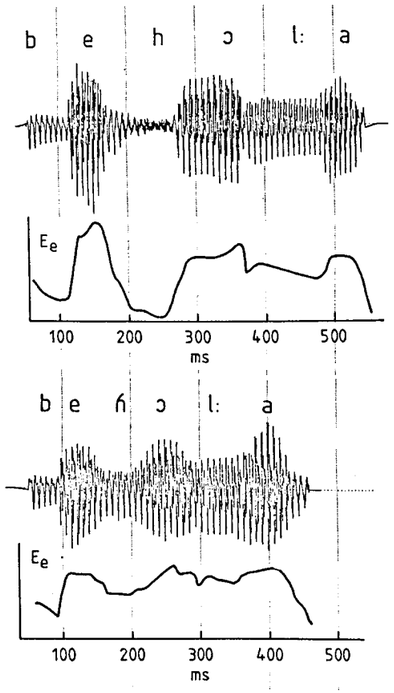


Fig. 1. Oscillogram and equivalent source amplitude E_e of the word "behålla" [beh'ɔl:a]. Above the word in sentence focus, below in prefocus position.

One approach is to refrain from decomposing the speech into a source and a filter function. This is the basis for the Flanagan vocal tract analog which includes a two-mass self-oscillating representation of the vocal folds /15/. It has been a most important and influential tool for simulations.

Over the last eight years, a substantial amount of voice source studies and modeling has been carried out in our department /16-24/. Our recent modeling has been based on a definition of the source as the actual air flow passing through the glottis. The filter function is, accordingly, defined as the supraglottal volume velocity transfer function relating the output flow at the lips, or with radiation included, the sound pressure in front of the speaker to the glottal flow. In inverse filtering, this transfer function is canceled which ensures an output of true glottal flow. It should be observed that the source becomes a property of the entire system just as any flow or pressure within the vocal tract whilst the filter function excludes the glottal and subglottal impedance. Its sole function is to translate from glottal flow to output flow.

A consequence is that the instantaneous resonance frequencies of the whole system may differ somewhat from the corresponding resonance frequencies of the supraglottal system. Also, the rate of formant damping is enhanced during the glottal open period. These circumstances as well as the nonlinearity of glottal impedance and the presence of distributed excitations within the glottal cycle account for a modulation of the instantaneous frequency, phase and damping of formants during the open period. This interaction is usually a second-order effect. However, it puts the burden on the voice source to introduce these modulations in combination with the constant noninteractive filter function. The result is a ripple superimposed on an otherwise smooth glottal pulse shape and the presence of a pattern of distributed zeros in the source spectrum.

These irregularities are especially enhanced by the superposition of formant oscillations from previous voice periods which may occur at a high F0. They enter as components of the instantaneous pressure above and below the vocal folds and thus, to the transglottal pressure drop which has a square-law relation to the resulting flow. This nonlinearity accordingly accounts for an interaction between the existing flow-pressure state and a following excitation.

A more basic instance of vocal tract - source interaction is the tendency of a delay of glottal flow towards the end of the glottal open phase. The main pulse shape is "skewed" to the right in comparison with the profile of the time-varying glottis opening. A consequence is a greater steepness of the flow pulse at closure /25/. This steepness quantified by the maximum flow derivative at closure becomes a scale factor of formant excitations. The larger the negative flow derivative, the larger values the formant amplitudes will be. The maximum glottal flow amplitude (or more precisely, the total volume of the pulse) is a main determinant of low-frequency energy, e.g., the amplitude of the voice fundamental. Increased flow derivative at closure under the condition of constant pulse amplitude thus increases the level of formant amplitudes versus the fundamental. The pulse skewing increases with overall vocal tract inductance, i.e., with the length and inversely with the cross-sectional area of the main vocal tract constriction. Therefore, there is a small difference in inherent voice source strength of vowels. The [i] and [a] and [u] will thus gain about a decibel compared to less constricted vowels (see ref. /19/). These relations can be upset at high F0 values.

All these acoustic interaction phenomena display a seemingly random pattern of perturbations of the voice pulse shape which presumable adds to naturalness /26-27/. They are illustrated in Fig. 2 which shows glottal pulse shapes and spectra under two conditions, the source without any load and with the full load of sub- and supraglottal systems, glottal inductance and viscous resistance included.

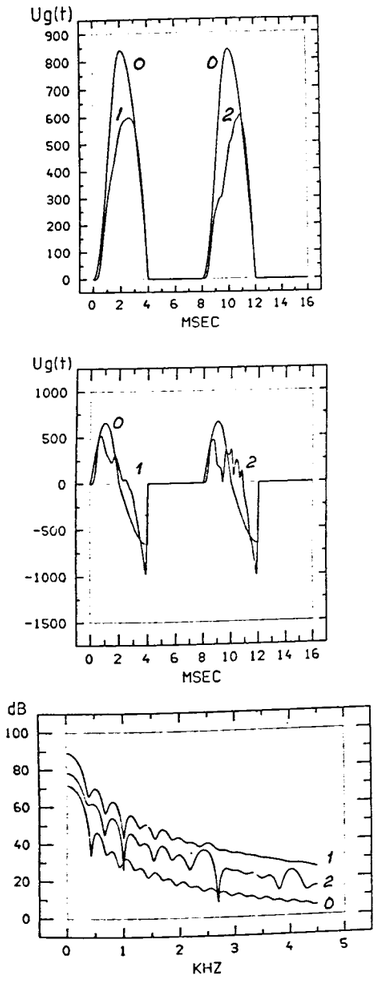


Fig. 2. From top: glottal flow, flow derivative, and flow derivative spectra from modeling of the vowel [i]. 0 stands for source without load, 1 for the first pulse and 2 for the second pulse with load.

The interaction ripple is larger in the second pulse than in the first pulse because of the nonlinear superposition effects. We may also observe spectral cancellation and reinforcement effects in the vincity of F2 and F4 of the vowel [i] which reflects a redistribution of spectral energy to fit the specific source-filter model.

In Fig. 3, pertaining to the vowel [a], we observe a zero between F1 and F2 in the vowel spectrum and an extra peak between F2 and F3 also associated with the nonlinear superposition /28/. The tendency becomes enlarged at large glottal openings and small losses within the vocal tract, and when a formant is much dependent on cavity structures close to the larynx. Figs. 2 and 3 originate from systematic simulations with our model. In true speech, we occasionally observe similar effects of extra peaks between formants which are not related to nasalization. The origin is the nonlinearity element in the source-filter system, see further the contribution of René Carré to this symposium.

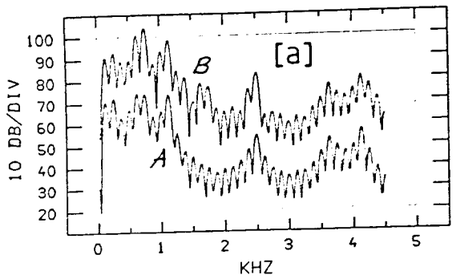


Fig. 3. Modeling of the spectrum of a vowel [a]. A=without interaction, B with acoustic interaction.

Another approach to defining source and filter which relies on approximations suitable for synthesis work is to start out by assuming a specific pulse skewing effect, i.e., the main shape of a glottal pulse which is incorporated in a constant volume velocity source feeding into the vocal tract terminal which is loaded by the glottal impedance /29/. Various alternatives exist such as incorporating the subglottal impedance also or to resort to a time average only of the load.

In formant synthesis one may take into account the variable loading by a modulation of formant bandwidths and frequencies within the glottal open period. This has been successfully exploited by Cheng and Guèrin /30/. However, available experimental data to assess the subjective gain of various solutions is still meager.

Summarizing interaction phenomena in voice production, we have described an acoustic interaction related to the dependency of the excitation mechanism on the instantaneous value of transglottal pressure drop in which components of formant oscillations gain prominence when the impedance of the supraglottal system is comparable to or larger than the glottal impedance. The main objection to selecting the equivalent constant current trans-

formation is the nonlinearity of the glottal impedance. To this acoustic interaction adds the mechanical interaction, related to the change in the aerodynamic forces, affecting the vocal folds as a consequence of a supraglottal constriction which may impede the flow as earlier described and in general causes perturbations of both voice fundamental frequency and flow pulse shapes, see further the contribution of K.N. Stevens who also treats interaction phenomena in the generation of unvoiced sounds.

Acoustic interaction alone can explain an interesting phenomenon in soprany singing. An articulation maintaining F1 close to F0 will not only maximize acoustic output but will also minimize the air consumption (see refs. /22-23/).

At increasing F0 and constant vocal tract filter function, formant amplitudes display periodic amplitude variations, the range of which is lowered by the extra damping associated with interaction. At the same time, the fluctuations of F2 amplitudes are no longer determined by the F2/F0 ratio only and appears to be influenced by the F1 component of transglottal pressure. This is demonstrated in Fig. 4. The full effects observed experimentally by Fant et al. /31/, probably include the vocal fold sound pressure mechanical interaction (see also ref. /20/).

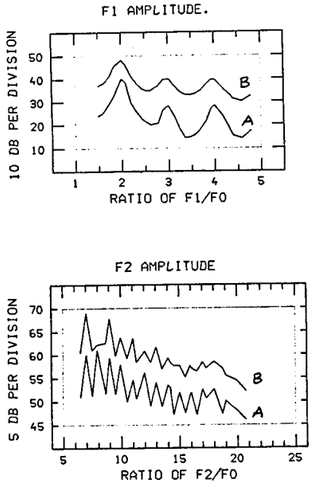


Fig. 4. F1 and F2 amplitude variations as a function of F0 in A: noninteractive, B: with acoustic interaction model. Vowel [ɛ].

More about interactive voice source effects will be reported by René Carré. The Grenoble group has also contributed to other aspects of source filter or rather source vocal tract interactions. One is related to the problem of the origin of inherent F0 of vowels. Guérin and Boe have shown that the aerodynamic forces on the vocal folds tend to lower F0 when F1-F0 is small and positive as could be the case for narrow front vowels /32/.

The subject of inherent pitch has also been treated by Zhang Jialu in his presentation to the congress. He finds that inherent pitch operates in the Chinese language and fairly independent of both distinctive tones and speaker sex. Inherent pitch differences are greatest at high F0. Zhang concludes, as is now generally accepted, that the mechanism of inherent pitch is effected by vocal fold tension passively induced by tongue elevation.

Experiments in France with the Flanagan type vocal tract analog have verified the raise of F0 with subglottal pressure of the order of 2-4 Hz per cm H_2O pressure increase /33-35/. The role of the subglottal system appears to be rather small. It adds slightly to the source excitation parameter and has a quite small effect on F0. Our experiments in Stockholm point to rather small influences of the subglottal system on formant frequencies and bandwidths except when the abduction is relative large and the subglottal pressure low.

There is evidence that an increase of subglottal pressure alone is followed by an approximately square-root dependent increase of maximum vibratory glottal area, see Fig. 8 of Flanagan, Ishizaka and Shipley /36/. Since particle velocity is proportional to the square root of pressure, and the volume velocity is the product of glottal area and particle velocity, it follows that glottal peak flow should increase in direct proportion to subglottal pressure. The accompanying shortening of the pulse base length and the increase of F0 accounts for an additional 3 dB increase in formant amplitudes, i.e., a doubling of subglottal pressure is associated with 9 dB overall spectral level gain (ref. /20/).

In the shift towards a pressed voice, there is an increase of maximum flow derivative at closure and thus of formant amplitudes at constant or even reduced glottal peak flow and a decrease of the open quotient.

We are now engaged in a project of parameterizing the voice source and tracking source parameters in connected speech /ref. 17/. One important parameter is the projection of the initial slope of the return phase on the time axis. This is a measure of the effective duration of the interval from maximum flow discontinuity in the closing branch to complete closure (see refs. /22-23/). This parameter is expecially apparent in breathy phonation. It is associated with reduced excitation and extra formant damping whilst the maximum flow may increase. These studies are also directed to the mapping of individual and of age- and sex-related specifics.

It appears to be fruitful to incorporate voice source parameters as correlates to prosodic categories. Rule-oriented studies are now under way to sort out segmentally induced interactions from underlying prosodic patterns. An example was given in Fig. 1. It is apparent that both prosodic-suprasegmental and inherent-segmental structures are related to all factors of speech production, articulation as well as phonation, and thus source as well as filter functions (ref. /13/).

ACKNOWLEDGMENTS

Christer Gobl has contributed with Fig. 1 and Lin Qiguang with Figs. 2-4.

REFERENCES

/1/ J. Perkell, D. Klatt, Eds., "Invariance and variability of speech processes", Lawrence Erlbaum, New York 1986.
/2/ B. Lindblom, "Phonetic invariance and the adaptive nature of speech", lecture at 30th Ann. of the IPO, Eindhoven, 1987.
/3/ R. Jakobson, G. Fant, M. Halle, "Preliminaries to speech analysis. The distinctive features and their correlates", MIT Press, Cambridge, MA, 7th ed. 1967.
/4/ G. Fant, "Descriptive analysis of the acoustic aspects of speech, Logos 5, 3-17, 1962.
/5/ F.J. Nolan, "The role of action theory in the description of speech production", Linguistics 20, 287-308, 1982.
/6/ B. Lindblom, P. MacNeilage, "Action theory, problems and alternative approaches", J. of Phonetics 14, 29-60, 1986.
/7/ J.A.S. Kelzo, E.L. Saltzman, B. Tuller, "The dynamical perspective on speech production: data and theory", J. of Phonetics 14, 29-60, 1986.
/8/ G. Fant, "The relations between area functions and the acoustic signal", Phonetica 37, 55-86, 1980.
/9/ A.L. Liberman, I.M. Mattingly, "The motor theory of speech perception revised", Haskins Lab., SR-82/83, 63-93, 1985.
/10/ G. Fant, Auditory patterns of speech", in W. Wathen-Dunn, ed., Symp. on models for the perception of speech and visual form 1964, M.I.T. Press, Cambridge, Ma, 1967.
/11/ B. Lindblom, J. Lubker, T. Gay, "Formant frequencies of some fixed mandible vowels and a model of speech motor programming by predictive simulation", J. of Phonetics 7, 147-162, 1979.
/12/ G. Fant, U. Stålhammar, I. Karlsson, "Swedish vowels in speech material of various complexity", in G. Fant, ed., Speech communication, Vol. 2, Almqvist & Wiksell Int., Stockholm.
/13/ G. Fant, L. Nord, A. Kruckenberg, "Segmental and prosodic variabilities in connected speech. an applied data-base study", paper XI ICPhS, Tallinn, 1987.
/14/ C.A. Bickley, K.N. Stevens, "Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies", J. of Phonetics 14, 385-392, 1986.
/15/ K. Ishizaka, J.L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell System Techn. J. 51, 1233-1268, 1972.
/16/ G. Fant, "Vocal source analysis - a progress report", STL-QPSR 3-4/1979, 31-54 (KTH, Stockholm).
/17/ G. Fant, "Voice source dynamics", STL-QPSR 2-3/1980, 17-37 (KTH, Stockholm).
/18/ G. Fant, "Preliminaries to the analysis of the human voice source", STL-QPSR 4/1982, 1-27 (KTH, Stockholm).
/19/ T.V. Ananthapadmanabha, G. Fant, "Calculation of true glottal flow and its components", Speech Communication 1, 167-184, 1982.
/20/ G. Fant, T.V. Ananthapadmanabha, "Truncation and superposition", STL-QPSR 2-3/1982, 1-17 (KTH, Stockholm).
/21/ L. Nord, T.V. Ananthapadmanabha, G. Fant, "Signal analysis and perceptual tests of vowel responses with an interactive source filter model", STL-QPSR 2-3/1984, 25-52 (KTH, Stockholm).
/22/ G. Fant, Q. Lin, C. Gobl, "Notes on glottal flow interaction", STL-QPSR 2-3/1985, 21-45 (KTH, Stockholm).
/23/ G. Fant, J. Liljencrants, Q. Lin, "A four-parameter model of glottal flow", STL-QPSR 4/1985, 1-13 (KTH, Stockholm).
/24/ T.V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics", STL-QPSR 2-3/1984, 1-24 (KTH, Stockholm).
/25/ M. Rothenberg, "An interaction model for the voice source", STL-QPSR 1/1981 (KTH, Stockholm), 1-17.
/26/ G. Fant, "Glottal flow: models and interaction", J. of Phonetics 14, 393-400, 1986.
/27/ L. Nord, T.V. Ananthapadmanabha, G. Fant, "Perceptual tests using an interactive source filter model and considerations for synthesis strategies", J. of Phonetics 14, 401-404, 1986.
/28/ Q. Lin, G. Fant, "Complete simulation of voice source - vocal tract interaction", paper, Int.Conf. on Information Processing, China, 1987.
/29/ G. Fant, "The voice source-filter concept in speech production", STL-QPSR 1/1981, 21-37 (KTH, Stockholm).
/30/ Y.M. Cheng, B. Guérin, "Dynamically controlled excitation source for a time-varying formant synthesizer, 2003-2006, ICASSP 86, Tokyo, 1986.
/31/ G. Fant, K. Fintoft, J. Liljencrants, B. Lindblom, J. Mártony, "Formant amplitude measurements", J.Acoust.Soc.Am. 35, 1753-1761, 1963.
/32/ B. Guérin, L.J. Boe, "Etude de l'influence du couplage acoustique source-conduit vocal sur F0 des voyelles orales", Phonetica 37, 169-192, 1980.
/33/ B. Guérin, L.J. Boe, "A two-mass model of the vocal cords: determination of control parameters and their respective consequences", 583-586, IEEE-ICASSP, 1977.
/34/ B. Guérin, D. Degryse, L.J. Boe, "Acoustical consequences of parameters controlling of a vocal cord model coupled with the vocal tract", Report from Symp. on articulatory modelling, Grenoble, 1977.
/35/ B. Guérin, Effects of the source-tract interaction using vocal fold models", in J.R. Titze, R.C. Scherer, eds., Vocal Fold Physiology, The Denver Center for the Performing Arts, Denver, 1985, 482-499.
/36/ J.L. Flanagan, K., Ishizaka, K.L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell System Techn. J. 54, 485-506, 1974.

COORDINATION OF MUSCLES AND ARTICULATORS

VICTOR SOROKIN

Institute for Information Transmission Problems
Academy of Sciences, USSR
Moscow 101447

ABSTRACT

Internal models are important elements of the articulation control system. These models provide fullfilment of required targets both for the articulators position and speech signal parameters. Examples of the internal model action for the movement control of the jaw and lips are described.

Analysis of problems connected with the use of sensory information for control of human locomotion and articulation leads to necessity of an internal model existence in control systems. An internal model shapes control commands for a redundant mechanical system and provides adaptation to changing conditions - movements restrictions and various criteria of optimality.

Let us consider two tasks of movement control.

1. Movement of the finger tip from one point of space to another one along the given trajectory. Position of the finger tip during its movement is controlled by visual system and mechanoreceptors. Measured difference between visible position of the finger tip and the required trajectory must be transformed to muscle efforts of every link controlling the finger tip position in space. This is possible only in case the control system has an ability to predict the finger tip position for any set of muscle efforts even though approximately, i.e. there is an internal model of movements.

The situation is quite the same for the processing of feedback signals from mechanoreceptors. Every mechanoreceptor transmits an information on the muscle length or the joint angle only for one link of multylink system. To determine the finger tip position the information must be recalculated accordingly to kinematic connections, i.e. again there is a great need for an internal model.

2. Voice imitation of a melody not heard before. Here control of the length and tension of the larynx muscles that results in the required time change of the pitch of the vocal source. Difference between the given and the actual melodies is estimated in the perceptive analyser and can be transformed into muscle efforts only provided of an internal model existence which is "aware" of quantitative relationships between muscle efforts and voice pitch.

Therefore there is a necessity of an internal model for feedback closure in the control system. An internal model for the articulation control must have kinematics and dynamics equations for each level of the control system : motor units, muscles, articulators, vocal tract shapes, acoustical parameters of speech signal. An internal model also provides solutions for a set of important tasks as well as feedback closure. The main task of an internal model is the control shaping using mechanical system redundancy to satisfy imposed restrictions and optimality criteria.

Internal model notions have been under development for some time thus making contribution to the creation of the basis for a theory of purposeful movements [1,2, 5,6,8,9,10,12].

Let us consider an internal model functioning on an example of lips closure. Distance between lips h depends on the position of three articulators : the upper and lower lip and the jaw

$$h = y_3 - \Delta y_1 - \Delta y_2,$$

where Δy_1 is a vertical displacement of the jaw, Δy_2 is a vertical displacement of the lower lip, y_3 is a coordinate of the upper lip. As it is seen, this condition can be satisfied under different values of Δy_1, Δy_2 and y_3 but there is a set of restrictions both for articulator displacements and muscle efforts.

First of all let us consider joined movement of the jaw and the lower lip. Let h is the required height of the lower lip. Then equation of kinematics is

$$\Delta y_1 + \Delta y_2 = h$$

with restrictions

$$0 \leq \Delta y_1 \leq a_1, \qquad 0 \leq \Delta y_2 \leq a_2$$

and initial conditions $\Delta y_1 = 0$, $\Delta y_2 = 0$

If the criterion of optimality is minimum of muscle efforts

$$\min \{ F_1 + F_2 \}$$

then, considering only elastic forces

$$F_1 = c_1 \Delta y_1$$
$$F_2 = c_2 \Delta y_2$$

have

$$\min \{ c_1 \Delta y_1 + c_2 \Delta y_2 \}$$

where c_1, c_2 - coefficients of elastic resistance. From $\Delta y_2 = h - \Delta y_1$ obtaine criterion

$$\min \{ (c_1 - c_2) \Delta y_1 + c_2 h \} \qquad (1)$$

with new restrictions

$$\max (0, h - a_2) \leq \Delta y_1 \leq \min (a_1, h)$$

As it is seen, (1) is a straight line having the minimal value on one of the boundaries of the displacement range. In case of $c_1 > c_2$ the minimum is reached for $\Delta y_1 = 0$, $\Delta y_2 = h$, if $h \leq a_2$, and for $\Delta y_1 = h - a_2$, when $h > a_2$.

In case of $c_2 > c_1$ there is inverse relashionship - the minimum is achieved for $\Delta y_1 = a_1$, $\Delta y_2 = h - a_1$ if $h > a_1$ and for $\Delta y_1 = h$, $\Delta y_2 = 0$, if $h \leq a_1$.

Let us consider now contrary lips movements. The condition of the lips closure is $y_3 - y_2 = 0$, where y_2 is a vertical coordinate of the lower lip, y_3 is a coordinate of the upper lip. Restrictions for displacements are

$$a_2 \leq y_2 \leq b_2$$
$$a_3 \leq y_3 \leq b_3$$

with initial conditions $y_2 = a_2$, $y_3 = b_3$.

Elastic forces are

$$F_2 = c_2 (y_2 - a_2)$$
$$F_3 = c_3 (b_3 - y_3)$$

and the criterion of the force minimum is

$$\min \{ (c_2 - c_3) y_2 + c_3 b_3 - c_2 a_2 \} \qquad (2)$$

with new restrictions

$$\max (a_2, a_3) \leq y_2 \leq \min (b_2, b_3).$$

Again (2) is a linear function. For vertical displacements of the upper lip it is required more efforts then for displacements of the lower lip since the upper lip is lowered only due to orbicularis oris contraction. Thus $c_3 > c_2$ and the minimum of efforts is reached for

$$y_2 = \min (b_2, b_3)$$

i.e. for $y_2 = b_2$, since $b_3 > b_2$, $y_3 = 0$.

Therefore, if the criterion of optimality is a minimum of efforts then both during joined movements of the jaw and the lower lip and contrary movements of the lips the most displacement has the articulator with the least elastic resistance.

If not effort but work $A = F \Delta y$ is minimized then displacement for each articulator depends also on the amount of the required displacement. For joined movement of the jaw and the lower lip the criterion of optimality is

$$\min \{ (c_1 + c_2) \Delta y_1^2 - 2 c_2 h \Delta y_1 + c_2 h^2 \}$$

If

$$(c_1 + c_2) \Delta y_1^2 + c_2 h^2 > 2 c_2 h \Delta y_1$$

then minimum is achieved for $\Delta y_2 = a_2$, $\Delta y_1 = 0$. If

$$(c_1 + c_2) \Delta y_1^2 + c_2 h^2 < 2 c_2 h \Delta y_1$$

then minimum is achieved for

$$\Delta y_1 = c_2 h / (c_1 + c_2)$$

if $h \leq a_1$ and for $\Delta y_1 = \max (0, h - a_2)$ in opposite case.

For contrary movements of lips the criterion of optimality is

$$\min \{ (c_2 - c_3) y_2^2 + (c_3 b_3 - c_2 a_2) y_2 \}$$

which is reached in the point $y_2 = b_2$ if $y^* - a_3 < y^* - b_2$, and in the point $y_2 = a_2$, if $y^* - a_3 > y^* - b_2$, where

$$y^* = (c_2 a_2 - c_3 b_3) / 2 (c_2 - c_3).$$

It is seen from the condition of lips closure, that if for some reason the change of any coordinate y_i is impossible or the range of displacement has reduced, then for the given criteria of optimality the control system will immediately recalculates the required efforts F as soon as it gets an information on new conditions. This event is virtually observed for various paralysis of facial muscles and in experiments with jaw movements restriction by means of bite-block or electromechanical device [3,7] . As it is known, in these cases lips closure is achieved by means of the change of another articulators movement, if such a change is physically possible.

An internal model governs joined movements of articulators in space as well as in time in such a way that a certain configuration of the vocal tract for the given time interval could be shaped. Prediction-type coarticulation is possible only if an internal model "knows" not only kinematics but dynamics of articulators as well. For instance, using dynamics equation

$$my'' + by' + cy = F(t) \qquad (3)$$

an internal model can determine an articulator position in any moment of time for the given target position $y = y_0$ and restrictions on velocity y' and acceleration y''.

The faster movement is required according to coarticulation conditions, the greater muscle effort is spent. The economy criterion of efforts or fulfilled work appears in continious speech in such a way that even for the same target posi-

tions of articulators development of muscle efforts for initial sounds is twice or thrice as slow as for sounds inside words [11]. The muscle is nonlinear system in regard to the contraction velocity, thus a simple change of rate of articulation leads to considerable reorganization of control - other motor units and muscles are activated. For example, masseter is activized only for fast articulation of meaningful utterances ; slow articulation of / t / is provided by both the tongue and the jaw movements, but fast articulation - mainly by the jaw movements [12].

It is known, that for the certain area of the vocal tract turbulent noises appear and determine phonetical quality of sounds. However, the control system must avoid turbulent noises during articulation of close vowels like / i / and a closure during fricatives articulation. This requirement imposes restriction on the constriction velocity in the vocal tract. Indeed, measurements show that for the jaw velocity of movement up is slower then velocity of movement down [4,12].

In comparison with other modes of control an internal model has the advantage of guaranteed achievement of any physically realized target without test movements if it has complete information on kinematics, dynamics and current conditions of the control system. For example, required force F(t) in (3) can be calculated straight from the given displacement, velocity and acceleration.

An internal model in the speech production control system is aware not only of the connection between physical levels but of the code structure of speech flow as well. That code structure allows to correct errors of articulation and speech signal distortions. Thus each speaker can speak different style depending on circumstances, i.e. can use different control commands to achieve the requiered intelligibility under the given conditions.

Important function of an internal model consists in the estimation intraspeaker speech using both mechanoreceptors signals and results of the perceptive process. Then it is quate natural to suppose that the function is used for other peoples speech recognition. Accepting this hypotesis we maintane the physiological basis for regeneration the perception motor theory on a new level.

The idea concernig an internal model permits to skip endless process of experimental investigations of every separate level of the articulation control system and come to an analysis of the whole system of speech production.

1. Adams J.A. Issues for a closed-loop theory of motor learning. Motor Control London, Academic Press, 1976, p. 87-107.

2. Gurfinkel V.S., Levik Y.S. Sensory complexes and sensorymotor integration. Human Physiology, 1979, N 3, p. 399-414, (in Russian).

3. Folkins J.W., Abbs J.H. Lip and jaw motor control during speech. Responses to resistive loading of the jaw. JSHR, 1975, N 18, p. 207-220.

4. Imagava et all. Comparison of velocity and duration between open to close and close to open vowel transition. Ann. Bull. RILP, Tokyo, 1983, N 17, p. 33-36.

5. Kelso S., Stelmach G.E. Central and peripheral mechanisms in motor control. Motor Control, London, Academic Press, 1976, p. 33-40.

6. Korenev G.V. A target and adaptation of movement. 1974, (in Russian).

7. Lindblom B., Lubker J., Gay T. Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. J. of Phonetics, 1979, N 7, p. 147- 161.

8. Russel D.G. Spatial location cues and movement production. Motor Control, London, Academic Press, 1976, p. 67-85

9. Schmidt R.A. The schema as a solution to some persistent problems in motor learning theory. Motor Control, London, Academic Press, 1976, p. 41-66.

10. Schmidt R.A. The schema concept. Human Motor Behavior, ed. Kelso S., 1982, p. 219-238.

11. Sorokin V.N. Influence of articulation rate onto neuromotor processes in facial muscles. Human Physilogy, 1981, N 1, v. 7, p. 40-45, (in Russian).

12. Sorokin V.N. The speech production theory, 1985, (in Russian).

INTERACTION BETWEEN ACOUSTIC SOURCES AND VOCAL-TRACT CONFIGURATIONS FOR CONSONANTS

KENNETH N. STEVENS

Research Laboratory of Electronics and
Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge MA 02139 USA

Abstract

When a narrow constriction is formed in the vocal tract, there are substantial interactions between the vocal-tract configurations and the sources of sound in the vocal tract. In the case of liquids and glides, the interaction causes the amplitude of the glottal source to decrease somewhat relative to its amplitude for vowels. When the constriction is narrow enough to cause frication noise, there are stron interactions between vocal-tract configuration and both the noise source and the glottal source. A theoretical analysis of some aspects of this interaction is presented, together with some illustrative data.

1. Introduction

It is usually assumed that the acoustic source due to vocal-fold vibration for vowels is largely independent of the vocal-tract configuration. That is, the shape and frequency of the pulses of volume velocity passing through the glottis are not significantly influenced by the vocal-tract configuration, and the acoustic properties imposed on the sound by the vocal tract are largely independent of the glottal source. Recent analytic and experimental work has indicated that there are some interactive effects, between source and vocal tract [1, 2]. The interactive effects on the source become strongest when the vocal tract is constricted and also when the fundamental frequency is in the vicinity of the first formant frequency (F1). The effect is to modify somewhat the waveform of the airflow pulses emerging from the glottis, whereas the effect on the pattern of mechanical vibration is relatively small.

The basic mechanism of the interaction is that, during the open phase of the glottal cycle, the airflow is determined not only by the impedance of the glottal opening, which is largely resistive, but also by the impedance of the sub- and supraglottal airways, which is largely reactive in the frequency range of glottal vibration. This reactive impedance, which is determined by the acoustic mass of the airways, produces a skewing of the volume-velocity waveform relative to the waveform of glottal area versus time. There is also a small effect of the vocal-tract configuration on the frequency of glottal vibration. This is thought to be due to mechanical coupling between displacement of the tongue root and rotation of the thyroid cartilage, leading to a change in vocal-fold tension [3].

The approximation of independence between acoustic sources and vocal-tract shapes is far from being valid when the airways between the glottis and the vocal-tract output contain a narrow constriction, as is the case for many consonants. When there is vocal-fold vibration during the interval in which the airways are constricted, the shape and frequency of the pulses of glottal airflow can be strongly dependent on the supraglottal configuration, including the size of the constriction, the impedance of the vocal-tract walls, and the way the walls are displaced during the constricted interval. Under some circumstances, the airflow through the constriction is sufficient to generate a source of turbulence noise, and the characteristics of this noise are clearly dependent on the shape of the vocal tract. The aim of this paper is to explore the nature of the interaction between acoustic sour ces and vocal-tract shapes for these more constricted, consonantal configurations.

2. Glottal source for liquids and glides

The articulatory configuration for liquids and glides is generally characterized by a constriction or narrowing at some point along the vocal tract such that the cross-sectional area is smaller than that for vowels. Measurements have been made to determine the effect of this type of constricted configuration on the glottal source [4, 5].

In one type of measurement, the spectrum of the sound was compared at two points within each of a number of utterances containing liquids and glides: (1) when the vocal tract was constricted during the liquid or glide and (2) when the vocal tract was relatively unconstricted in a vowel adjacent to the consonant. Measurements of the amplitude of the fundamental component were made in these regions, and appropriate corrections were made for the influence of the first-formant frequency on the amplitude of

this component in the glottal spectrum. A change in the amplitude of this component in general indicates a change in the amplitude of the glottal pulses. The results for six speakers showed that when the vocal tract was maximally constricted for a liquid or glide, there was a reduction of about 3 dB in the amplitude of the glottal pulses, on the average. This relatively modest change can presumably be ascribed, at least in part, to an acoustic interaction between the glottal airflow and the increased impedance of the airway for the constricted configuration.

In addition to the acoustic measurements, an electroglottograph was used to obtain estimates of glottal opening and closing times for a number of glottal cycles in the same liquids and glides and in the adjacent vowels. The aim was to determine whether there was evidence for an influence of the vocal-tract constriction for the consonant on the mechanical aspects of vocal-fold vibration. The measurements showed that there was an average increase of about 10 percent in the glottal open time within the vibratory cycle during the constricted interval for liquids and glides relative to that for a vowel.

Models of vocal-fold vibration predict effects in the same direction, i.e., a decreased amplitude of the glottal pulses and an increased duration of the glottal opening, but current models are not sufficiently refined to predict accurately the magnitude of the effects. In any event, it seems clear that when the vocal tract is constricted to form a liquid or a glide, there is a modest influence on the glottal volume-velocity source. This influence is in a direction that tends to reduce the overall amplitude of the sound during the consonant, and thus can be considered to enhance the contrast between the syllabic peaks for vowels and the intensity minima for nonsyllabic segments.

3. Turbulence noise source for fricatives

The generation of turbulence noise in the vocal tract for a fricative consonant is accomplished by forming a constriction in some region along the supraglottal airways and directing the airstream against an obstacle or surface in the vocal tract. In a sense, then, the properties of the source are a direct result of interaction of the vocal-tract shape and the airflow in the vocal tract.

A typical configuration of the vocal tract for a fricative consonant is shown in Fig. 1a. The factors influencing the properties of the source of turbulence noise can be explored by examining the properties of the sound that is generated when air flows through a mechanical model like that in Fig. 1b. The rapid airflow through the downstream constriction generates turbulence noise, and if an obstacle like that shown in the figure is present in the airstream, this sound source is concentrated in the vicinity of the obstacle. The constriction at the left of the model represents the glottal opening. A number of measurements have been made of the sound generated by airflow in a model of the type shown in Fig. 1b [6].

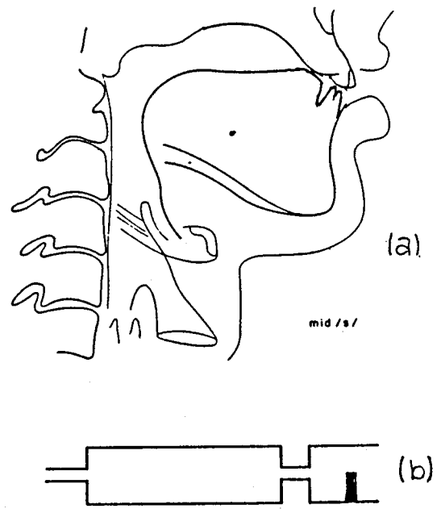


Fig. 1(a) Midsagittal configuration of vocal tract for a fricative consonant. **(b)** Mechanical model of fricative noise generation with an obstacle downstream from the constriction.

Figure 2 shows the spectrum of the sound pressure measured some distance from the model with and without the obstacle downstream from the constriction. The difference in level is as great as 30 dB in some frequency regions. Turbulence noise in the vocal tract is never entirely free of some obstacle or surface, but clearly large differences in source strength are obtained for different configurations of obstacles in the airstream. This fact is evidently exploited in the selection of articulatory gestures for producing feature combinations and contrasts for use in language.

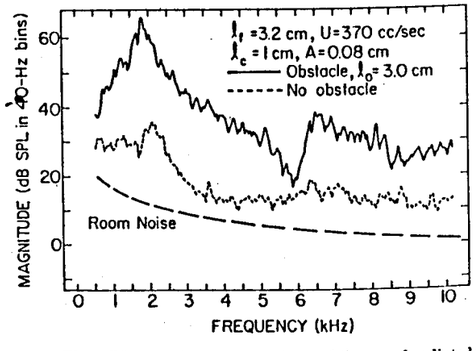


Fig. 2 The solid curve shows the spectrum of the sound radiated from a model like that in Fig. 1(b). Front-cavity length and distance from constriction to obstacle are both about 3 cm. The dotted curve is the spectrum for the same configuration, but with the obstacle removed. (From Shadle, 1985).

The results of a series of experiments with models of the type shown in Fig. 1b, together with theoretical studies of turbulence noise generation, have led to procedures for calculating the amplitude and spectrum of the sound-pressure source p_s for different values of airflow and constriction size. The relation that has been developed is

$$p_s = K_1 U^3 A^{-\frac{3}{2}} = K_1 \left(\frac{2\Delta P}{\rho}\right)^{\frac{3}{2}} A^{\frac{1}{2}} \qquad (1)$$

where U = volume velocity through the constriction, ΔP = pressure drop across the constriction, ρ = density of air, A = cross-sectional area of the constriction, and K_1 is a constant that depends on the configuration of obstacles and surfaces against which the airflow impinges.

Equation (1), together with standard equations relating pressure and airflow in constricted tubes, can be used to calculate the levels of the noise sources at the glottal constriction and at the supraglottal constriction in the model of Fig. 1b as the cross-sectional area of the supraglottal constriction is manipulated. Figure 3a shows the results of this type of calculation when the cross-sectional area of the glottis is fixed at 0.2 cm². The amplitude of the source at the downstream constriction has a broad maximum for constriction sizes in the range 0.05 to 0.15 cm². Over this range, the amplitude of the turbulence noise source is relatively insensitive to the size of the supraglottal constriction. This result suggests that, in producing a fricative consonant, a speaker is not required to adjust the size of the constriction precisely in order to obtain a fixed maximum amplitude for the noise source. The figure also shows that as the supraglottal constriction becomes larger, the amplitude of the noise generated at that constriction decreases, and the noise source at the glottis becomes dominant. This glottal source in the model corresponds to aspiration noise in the vocal tract.

One other type of interaction has been ignored in estimating the source levels in Fig. 3a. Since the walls of the vocal tract are not rigid, an increase in pressure behind the constriction will cause the walls of the vocal tract to displace outwards in response to the pressure. This displacement of the walls can have an influence on the size of the constriction. In the configuration of Fig. 1a, for example, the heightened pressure behind the alveolar constriction exerts a force that causes a downward displacement of the tongue blade that is sufficient to cause an increase in the size of the constriction. If the constriction is adjusted to have a "resting" cross-sectional area in the absence of an applied subglottal pressure, then the cross-sectional area when the pressure is applied will become larger. It is possible to calculate approximately this displacement, since some data are available on which to make estimates of the mechanical compliance of the vocal-tract walls [7]. Based on these calculations, we have recomputed the curves in Fig. 3a, except now we have plotted on the abscissa the resting area of the constriction, rather than the actual area after the subglottal pressure is applied. The new curves are given in Fig. 3b. We again observe a maximum in the amplitude of the noise generated near the downstream constriction, except that now the maximum is much broader. That is, the level of the noise is even more stable in response to perturbations in the area of the constriction. This example, as well as others not discussed here, illustrate that properties of the vocal-tract walls can play a significant role in determining the characteristics of sound sources in the vocal tract.

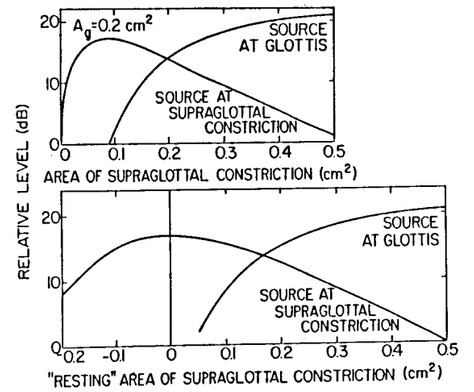


Fig. 3(a) Calculated levels of the turbulence noise sources at the supraglottal and glottal constrictions as a function of the area of the supraglottal constriction. The area of the glottal constriction is fixed at 0.2 cm². (b) Same as (a), except that the size of the constriction is modified by the presence of the intraoral pressure. The abscissa is the area that the supraglottal constriction would assume before application of the subglottal pressure.

4. Source characteristics for voiced fricatives

A voiced fricative consonant is produced with a supraglottal configuration similar to that for a voiceless fricative, described above. The difference, however, is that the vocal folds continue to vibrate over at least part of the time interval when the vocal tract is in the constricted configuration. If noise is to be generated at the supraglottal constriction and if glottal vibration is to continue, then the pressure in the space between the glottis and the constriction must be intermediate between the subglottal pressure and atmospheric pressure. In order to achieve this condition , the speaker must maintain a rather careful balance between the configuration of the glottis and the cross-sectional area of the supraglottal constriction.

If the average pressure in the vocal-tract volume between the glottis and the supraglottal constriction is P_m and the subglottal pressure is P_{sub}, then the transglottal pressure

is $P_{trans} = P_{sub} - P_m$, and the pressure across the constriction is P_m. This pressure drop P_m is the value of ΔP that is used to esimate the amplitude of the turbulence noise source at the constriction, in equation (1). When the vocal folds are in a configuration appropriate for voicing, experimental observations and theoretical analysis have shown that the amplitude of the volume-velocity pulses at the glottis is roughly proportional to $P_{trans}^{\frac{1}{2}}$, i.e., proportional to $(P_{sub} - P_m)^{\frac{1}{2}}$ in this case [8]. However, when the transglottal pressure becomes less than a particular threshold value, usually considered to be about 2-3 cm H_2O, vocal-fold vibration is no longer maintained.

In Fig. 4 we estimate the level of the turbulence noise source at the constriction and the level of the glottal source as a function of the cross-sectional area of the supraglottal constriction. The average area of the glottis is assumed to be fixed (at 0.13 cm^2 in this case). The noise level is plotted relative to the maximum level that would be obtained for a voiceless fricative consonant produced with a glottal opening as indicated in Fig. 3. The amplitude of the glottal source is plotted relative to the amplitude that is obtained with the same glottal configuration but with no supraglottal constriction. When the constriction size decreases to about 0.07 cm^2, the transglottal pressure drops to about 2 cm H_2O, and vocal-fold vibration can be assumed to cease for smaller constrictions.

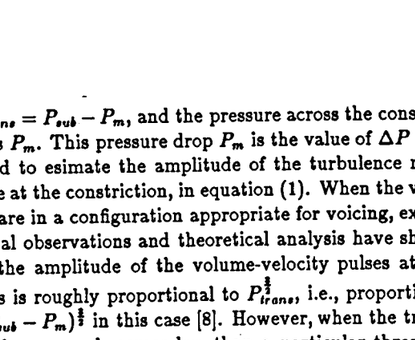


Fig. 4 Calculated amplitudes of noise source and glottal source for a model of a voiced fricative consonant, as a function of the cross-sectional area of the supraglottal constriction. The average glottal area is assumed to be fixed at about 0.13 cm^2. Calculated voicing amplitude is given relative to the amplitude when the vocal tract is unconstricted, as it might be for a vowel. Noise amplitude is given relative to the amplitude for a model of a typical voiceless fricative. The vertical line at about 0.07 cm^2 indicates the constriction area below which vocal-fold vibration cannot be sustained.

Figure 4 indicates that there is some instability in adjusting the supraglottal constriction size. If strong voicing is to be maintained (say around a constriction size of 0.2 cm^2), then the amplitude of the noise source becomes small, whereas a strong noise source can only be achieved at the expense of weakened or even cessation of voicing. It often happens that voicing is maintained only over some part of the interval when the vocal tract is constricted for a voiced fricative.

Examples of measurements of the voicing amplitude and the noise amplitude are shown in Fig. 5 for the intervocalic voiceless and voiced fricatives [s] and [z]. In the case of the voiceless fricative, the amplitude of the noise remains rather stable throughout the interval, and the glottal source turns off and on rather abruptly. There is a brief reduction in noise amplitude just before voicing onset, as the constriction is released before the vocal folds have adducted to a position appropriate for voicing. For the voiced fricative, there is considerable variation in both noise amplitude and glottal source amplitude. The noise amplitude tends to be 5-10 dB less than that for the voiceless cognate, over much of the constricted interval.

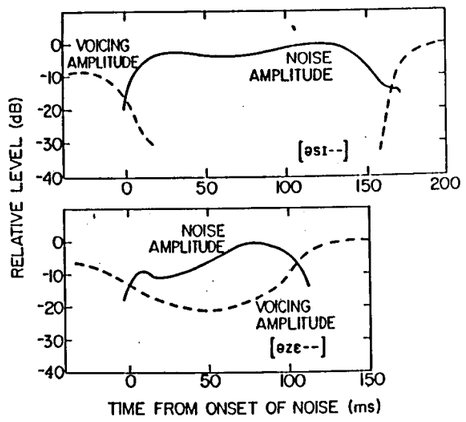


Fig. 5 Measurements of noise amplitude (peak spectrum amplitude above 3.5 kHz) and voicing amplitude (peak spectrum amplitude at *F1*) as a function of time for an intervocalic voiceless fricative [s] (top panel) and voiced fricative [z] (bottom panel). The voicing curves are plotted with 0 dB as the maximum level, and the noise amplitude curves are plotted with the reference of 0 dB being the peak level for the voiceless cognate.

The reduced amplitude of the glottal source for the voiced fricatives [v] and [ð] has been observed for a variety of utterances by Bickley and Stevens [5]. They report an average reduction in amplitude (relative to that for a vowel) of about 10 dB for these fricatives.

5. Concluding remarks

When a constriction in the vocal tract becomes narrower than the minimum cross-sectional area for a vowel, there is a substantial increase in the interaction between the vocal-tract shape and the characteristics of the source. For liquids and glides, the contriction is not sufficiently narrow to produce turbulence noise, but the impedance of the constriction is sufficient to cause a decreased amplitude of the glottal source. When the constriction is made narrower, turbulence noise is generated in the vicinity of the constriction. For a fixed glottal opening, the amplitude of this noise is reasonably stable over a substantial range of supraglottal constriction sizes. Simultaneous generation of both glottal vibration and turbulence noise at a supraglottal constriction requires a rather careful adjustment of the supraglottal and glottal constrictions. It is to be expected, therefore, that one or other of the sources may not continue throughout the entire consonantal interval for a voiced fricative.

This analysis has indicated that the interaction between acoustic sources and vocal-tract shapes tends to be much greater for constricted or consonantal configurations than it is for more open vocalic configurations. We still have much to learn about the characteristics of the glottal and turbulence noise sources for different types of consonants, particularly for obstruent consonants for which there is an increased pressure in the intraoral space.

6. Acknowledgements

This work was supported in part by a grant from the National Institute of Neurological and Communicative Disorders and Stroke and the National Science Foundation.

7. References

1. M. Rothenberg. *Acoustic interaction between the glottal source and the vocal tract.* In **Vocal Fold Physiology**, K.N. Stevens and M. Hirano, eds. Tokyo: University of Tokyo Press, 305-318 (1981).

2. G. Fant. *Preliminaries to analysis of the human voice source.* **Quarterly Progress and Status Report** (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm), STL-QPSR 4, 1-27 (1982).

3. K. Honda. *Relationship between pitch control and vowel articulation.* In **Vocal Fold Physiology: Contemporary Research and Clinical Issues**, D.M. Bless and J.H. Abbs, eds. San Diego: College-Hill, 286-297 (1983).

4. C.A. Bickley and K.N. Stevens. *Effects of a vocal-tract constriction on the glottal source: Experimental and modelling studies.* **J. Phonetics, 14**, 373-382 (1986).

5. C.A. Bickley and K.N. Stevens. *Effects of a vocal-tract constriction on the glottal source: Data from voiced consonants.* In **Laryngeal Function in Phonation and Respiration**, T. Baer, C. Sasaki, and K.S. Harris, eds. Boston: College-Hill, 239-253 (1987).

6. C. Shadle. **The acoustics of fricative consonants.** Technical Report 506, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge MA (1985).

7. K. Ishizaka, J.C. French, and J.L. Flanagan. *Direct determination of vocal tract wall impedance.* **IEEE Trans. on Acoustics Speech and Signal Processing, ASSP-23**, 370-373 (1975).

8. A. Bouhuys, J. Mead, D.F. Proctor, and K.N. Stevens. *Pressure-flow events during singing.* In **Sound Production in Man, Annals of the New York Academy of Sciences, 155**, 165-176 (1968).

THE INTRINSIC FUNDAMENTAL FREQUENCY OF VOWELS AND THE EFFECT OF SPEECH MODES ON FORMANTS

JIALU ZHANG

Institute of Acoustics, Academia Sinica
Beijing, CHINA

ABSTRACT

Two experiments have been carried out to explore the interaction phenomena. It is shown that 1) The intrinsic fundamental frequency of vowels is also found in Chinese, 2) The difference of intrinsic fundamental frequency between high and low vowels are related to the pitch level and a linear relationship was found in a certain dynamic range of register beyond that nonlinear relation will appear, 3) speech efforts influence not only the F0 but also the F1-F2 pattern of vowels.

INTRODUCTION

In recent years a more profound view of the speech production process is emerging and interactive models taking into account the interaction between source and vocal tract have been proposed [1,2,3,4]. In these models more attention has been paid to the acoustic interaction. However some investigations [5] shown that the mechanical interaction is prominent in some instances.

We still need, however, to incorporate more knowledge of source-filter interaction and speaker/speaking particulars to improve the model of speech production. In order to get an insight into the model, we have to get much more experimental data in dynamic process of connected speech.

The present paper deals with two interaction phenomena. The first concerns the relation between intrinsic fundamental frequency (henceforth IF0) for vowels in different tonal environments and different syllable structures, with groups of both adult male and female speakers. The second source of interaction studied comes from the dependent of vowel formant frequencies on speech modes.

EXPERIMENT I

A great deal of research has been devoted to the analysis and quantification of IF0 in several languages. However, none of these studies were concerned with the roles of pitch level and the syllable structures and word position as determinants of IF0. The speech material used in this study consists of two parts, 400 monosyllables and 509 disyllabic words.

In order to make all test items occur in the same phonetic environment and approch the situation of connected speech, all the monosyllables and disyllabic words were embedded in a frame sentence "Wǒ dú ___ zì." (I utter the character ___.) and "Wǒ dú _____ zhège cí." (I utter the word _____.) respectively. Ten speakers (5 male and 5 female) who speak "Putonghua" (standard Chinese) were recorded.

The measuring points of F0 are at the middle point for level tone T1; at lowest point T2-1 and highest point T2-2 for rising tone T2; at starting point T3-1 and lowest point T3-2 for dipping tone T3 (because in connected speech the tone contour of T3 will change from falling and rising into falling and low level except T3 is followed by another T3); at highest point T4-1 and lowest point T4-2 for falling tone T4.

The mean IF0 for each of nine Chinese vowels at different tonal points, and the IF0 difference between other vowels and /a/, △F0,derived from 400 monosyllables, averaged across consonantal contexts, and for 5 male and 5 female speakers respectively are listed in Table 1.

Table 1. Mean IF0 and IF0 difference between other vowels and /a/,△F0, at different tonal points for 5 males and 5 females respectively.

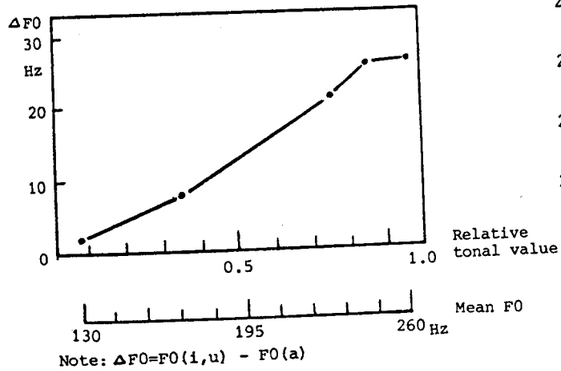
F0 and F0, (Hz), 5 males

	T1		T2-1		T2-2		T3-1		T3-2		T4-1		T4-2	
	F0	△F0	F0	△F0	F0	△F0	F0	△F0	F0	△F0	F0	△F0	F0	△F0
i	175	21	118	7	167	16	113	5	89	6	197	22	97	0
ɿ	181	27	122	11	171	20	116	8	90	7	208	33	99	2
ʅ	179	25	116	5	169	18	115	7	90	7	195	20	101	4
y	180	26	119	8	175	24	115	7	90	7	197	22	101	4
u	181	27	117	6	168	17	112	4	90	7	206	31	105	8
e	164	10	114	3	156	5	114	6	88	5	187	12	101	4
o	168	14	117	6	160	9	116	8	90	7	184	9	100	3
ɤ	170	16	116	5	170	19	122	14	88	5	178	3	100	3
a	154	0	111	0	151	0	108	0	83	0	175	0	97	0

5 females

i	291	15	205	7	265	10	219	-8	169	-2	312	10	180	-7
ɿ	302	26	206	8	271	16	214	-13	172	1	326	24	182	-5
ʅ	295	19	200	2	264	9	216	-11	168	-3	319	17	192	5
y	300	24	209	11	278	23	219	-8	171	0	318	16	176	-11
u	307	31	209	11	289	34	218	-9	172	1	335	33	184	-3
e	289	13	202	4	270	15	215	-12	170	-1	315	13	183	-4
o	278	2	200	2	270	15	213	-14	170	-1	310	8	183	-4
ɤ	302	26	200	2	274	19	209	-18	161	-10	314	12	182	-5
a	276	0	198	0	255	0	227	0	171	0	302	0	187	0

From Table 1. it can be seen that Chinese, which is a language with multitone system, also exhibits vowel IF0. Some negative values of F0 appeared at tonal points T3-1, T3-2, and T4-1 for females, this is due to the problems of F0 extraction. And it is worth to note that: 1) The IF0 difference between high vowels and low vowel /a/, F0, are related to the vowel pitch level; 2) There are no significant differences in the values of F0 between males and females. The IF0 difference, F0, is mainly dependent on the tonal value. The relation between F0 and pitch level is shown in Fig. 1. Two different kinds of tonal scale were used as abscissa, one is relative or normalized tonal value, which is defined as the average tonal value to tonal register ratio; the other one the average value of F0 corresponding to these tonal values over all speakers, both male and female.



Note: △F0=F0(i,u) - F0(a)

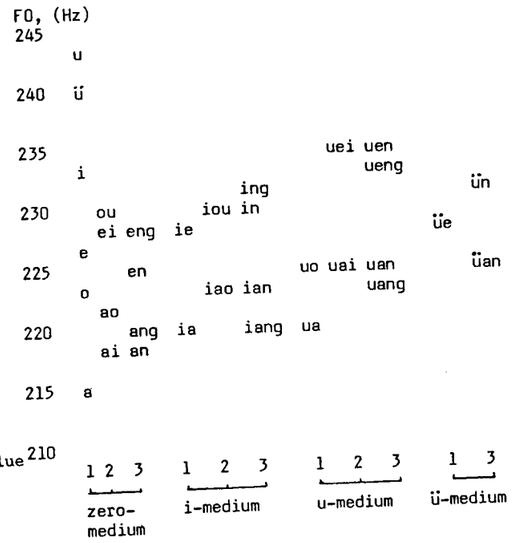
Fig. 1. △F0 versus mean F0 for male and female speakers at different tonal points.

From Fig. 1. it can be seen that F0 increases linearly with F0 for about an octave and after the linear section some nonlinearity or saturation of △F0 appears. Some similar phonomena have been observed in Italian where the accented syllables displaying greater IF0 than unaccented ones [6]. IF0 is redused in the final sentence positions with a lowered F0 [7]. In summary, a larger IF0 difference is generally related to a higher F0 (tonal value).

In order to show the influence of vowel combinations and of vowel nasalizations on the IF0, the F0 of "Yùnmǔ" (finals), with tone T1 averaged across consonantal contexts for 5 male and 5 female speakers, are drawn in Fig. 2. according to their structures.

Fig.2. shows that: 1) The nasalization of vowels reduce the IF0 difference between high vowels and low vowels, some similar results were found in French [5]; 2) "Yùnwěi" /i,u/ (vowel final endings) tend to increase the IF0 of main vowels; 3) "Jiémǔ" (medium vowels) /i,u,ü/ increase the IF0 of main vowels and main vowels with final endings (both vowel endings and nasal endings). In other words, diphthongs and triphthongs tend to reduce the IF0 difference. This is perhaps due to the fact that the supraglottal configuration is changed continuously and quickly during the pronunciation of compound vowels and the mechanical interaction has a certain time constant (due to the inertia of the muscles) so that F0, which is a laryngeal parameter, can not be changed as fast as the supraglottal tongue movement.

There have been various hypotheses concerned with the cause of vowel IF0. But none of them take notice of the "linearity" and "nonlinearity" of



Note: (1) /ao,iao/ are actually /au,iau/ in phonetic value.

(2) The number 1, 2, and 3 stand for the "yùnmǔ" without endings, with vowel endings, and with nasal endings in the corresponding column respectively.

Fig. 2. The influence of vowel combinations and vowel nasalizations on vowel IF0.

vowel IF0 in relation to larynx frequency. It seems that the tongue pull [8] or mechanical interaction theory has greater significance than other propositions. Here I try to give a probable interpretation from the point of mechanism of the vocalis muscle itself. According to Ohala's theory the tongue pull gives rise to increased vertical tension in the vocal folds through the mucous membrane and other soft tissues. We could assume that there must be a corresponding structural change in the mucous membrane and the soft tissues, and finally in the vocalis muscle itself thus causing a increased tension. The relationship between the tension T and the elongation X of the vocalis muscle can be approximated as [9]:

$$T=aEXP(bx) \qquad (1)$$

and to a first-order approximation the fundamental frequency F0 of the vibration of vocal folds as

$$F0=c0\sqrt{T} \qquad (2)$$

Then the incremental tension per unit elongation can be expressed as

$$\partial T/\partial X = bT \qquad (3)$$

If we neglect that c0 varies slightly with X, then we obtain

$$\partial F0/\partial X = (1/2)bF0 \qquad (4)$$

Formula (4) shows a linear relationship between Δ F0 and F0, in other words, the same incremental elongation Δ X due to tongue pull could cause a larger increment in tension Δ T, thus leading to a larger increment of fundamental frequency Δ F0, at high F0 than low F0. However, it seems that there is a dynamic range (about one octave in our data) for speakers control of their vocal folds. Beyond the dynamic range "nonlinearity" appears, perhaps, parameters a,b, and c0 in formulas (1) and (2) are not constant and the vibration of the vocal folds is associated with "overloading" near the upper end of the register.

EXPERIMENT II

Generally, it is assumed that formant frequencies are entirely determined by the vocal tract and independent from the voice source. And a target specified by the formant frequency of a vowel was considered as an invariant attribute. Recently some experimental results on dynamic spectra of speech sounds show that speech efforts and speech speed give strong influence not only on F0 but also on the F-pattern [10]. The testing material is a well designed sentence "Tā qù Wúxīshì, wǒ dào Hēilóngjiāng."(He comes to Wuxishi, I go to Heilongjiang.), in which three primary vowels /i,u,a/ are included. 12 speakers (6 male and 6 female) who are natives of Beijing uttered the testing sentence repeatedly in different speech modes in an anechoic chamber. The speakers were asked to change their speech efforts in five levels: 55, 60, 65, 70, and 75 dB SPL measured at 1 meter in front of the speaker's lips and to speed up the speaking rate by a factor of 2 at 65 dB SPL.

The F1-F2 plane for speech levels 55, 65, 75 dB and the high speed version are drawn in Fig.3. The variation of F0 with the SPL of speech for three syllables with different pitch levels are shown in Fig. 4.

Table 1. The average increments of F1 and F2, Δ F1 and Δ F2, for different vowels from 55 to 75 dB SPL of speech over 6 male and 6 female speakers.

	i	ü	u	-i	a
Δ F1, Hz	104	90	82	70	202
Δ F2, Hz	279	158	128	137	230

Note: -i stands for either [ɿ] and [ʅ].

From Fig. 3.a. it can be clearly seen that the vowel triangle is shifted and enlarged regularly as the SPL of speech increases. This suggests that both jaw and tongue movements are enlarged when a talker speaks loudly. On the other hand when the tempo is speeded up all vowels tend to be centralized and the vowel triangle is reduced (see Fig. 3.b.). The increments of F1 and F2 from 55 dB to 75 dB SPL of speech for different vowels are given in Table.1.

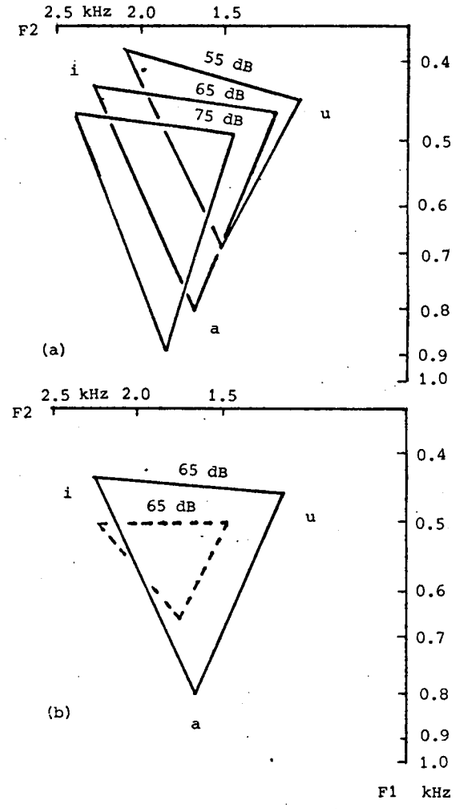


Fig. 3. F1-F2 plane for different speech efforts (a) and different utterance speeds (b).

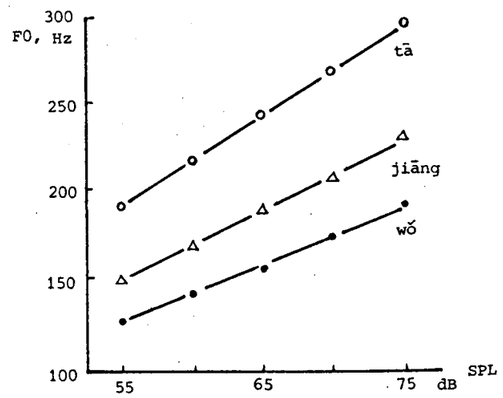


Fig. 4. F0 of main vowels with different pitch levels versus SPL of speech.

As the speakers increase their effort F1 shifts up in frequency by about 100 Hz for all vowels except /a/ which is raised by about 200 Hz. F2 shifts up in frequency somewhat more than F1. As for F3 the variations of all vowels are irregular.

Fig. 4 shows that the F0 increases linearly with the SPL of speech but that the steepness varies with the pitch levels of the vowels. The syllable "Tā" with level tone T1 in initial sentence position has the highest pitch level and the largest slope in F0 with increasing SPL of speech. The syllable "Wǒ" with dipping tone T3 has the lowest pitch level in the sentence and the smllest slope. As for the syllable "jiāng" with T1 in the sentence final position, it gets a mid pitch level and a middle slope in F0 because of declination. The average increment of F0 with increasing SPL of speech from 55 dB to 75 dB over 6 male, and 6 female speakers is about 111 Hz for high pitch vowel, about 98 Hz for mid pitch vowel and about 73 Hz for low pitch vowel. However, the relative increment of F0 or the increment in log F0 scale for the three syllables are nearly the same about 60 per cent i.e. F0(75dB)= 1.6F0(55dB).

According to Fant [1] there is a net gain of 9.5 dB included all factors accompanying the doubling of lung pressure. Starting out from F0=149 Hz, SPL=55 dB, say lung pressure P1=5 cm H2O and increasing SPL of speech to 75 dB, an increase of P1 should be to 20 cm H2O. And the increase of P1 by 15 cm H2O cause an increment of F0 by 73 Hz for low pitch vowel and by 111 Hz for high pitch vowel, in other words the incremental rate of F0 covers from 4.8 to 7.4 Hz/cm H2O. That is a reasonable value, but somewhat higher than the predicted 3.5 Hz/cm H2O [11].

Comparing the increments of F0 of vowels of different pitch levels caused by increasing lung pressure with the IF0 difference related to F0 of vowels, some similarity appears which might have a common physiological basis.

So when a speaker increases his/her lung pressure during speaking, both F0---source parameter and F1,F2---vocal tract parameters are simultaneously increased. This is due to the increased air push force which makes the glottis shift upward, then the tension of vocalis muscle is increased and the lenghth of vocal tract is shortened. This effect adds to increasing mouth opening. We searched persistently an invariant attribute of vowel targets but we found some floating islands---moving vowel triangles instead. It tells us that relative position of vowels in perceptual space are very important.

CONCLUSION

Two different experiments have illustrated interactions of phonatory and articulatory mechanisms. It seems from the present results that the mechanical interaction is stronger than acoustic interaction under these conditions. A lot of experimental results show that the intrinsic fundamental frequency of vowels is universal and we find here that it is also exhibited in Chinese---a tone language with multitone system. The intrinsic fundamental frequency difference Δ F0 increases linearly with pitch level over a certain dynamic range and then saturates in a nonlinear region near the upper boundary of the tonal range. The tongue pull theory and the stress-strain relationship of muscle could account almost entirely for vowel IF0.

Speech effort not only influences the F0 but also the F1-F2 plane. It is hard to determine vowel targets in connected speech, and the vowel triangle is floating. The relative positions of vowels in the F1-F2 plane, however, convey the perceptual features of vowel quality and relate to the speech modes. The fact that F0 increments caused by speech efforts are related to pitch level is somewhat similar to the IF0 difference caused by tongue pull. And a subglottal air push hypothesis could explain the fact that both F0 and F1-F2 plane pattern are changed with speech effort.

There remains much work to be done on the theoretical modeling and in regard to the development of experimental techniques to establish an advanced interactive model of speech production.

ACKNOLEDGEMENTS

The author wishes to thank Professor Fant for his very valuable comments and enthusiastic encouragement in this work and to express his sincere appreciation of Professor Fourcin's help.

REFERENCES

[1] Fant, G. (1982A): "Preliminaries to analysis of the human voice source", STL-QPSR 4/1982, 1-28.

[2] Fant, G. (1982B): "The voice source---acoustic modling", STL-QPSR 4/1982, 29-49.

[3] Fant, G (1985): "A four-parameter model of glottal flow", STL-QPSR 4/1985, 1-13.

[4] Rothenberg, M. (1981): "An interactive model for the voice source", STL-QPSR 4/1981, 1-17.

[5] Delos, M. et al. (1976): "Study of intrinsic pitch of vowels", J. Acoust. Soc. Am. 59,S72(A).

[6] Ferrero, F. E. et al., (1975): "Some acoustic and perceptual characteristics of the Italian vowels", VIIIth Int. Cong. Phon. Sc., (Leeds),mimeographed.

[7] Shadle, C. H. (1985): "Intrinsic fundamental frequency of vowels in sentence context", J. Acoust. Soc. Am., 78, 1562-1567.

[8] Ohala, J. (1973): "Explanations for the intrinsic pitch of vowels", Monthly Internal Memorandum, Phonology Lab.,Univ. of California, Berkeley (January), 9-26.

[9] Fujisaki, H. et al. (1981): Analysis of pitch control in singing", in Vocal folds physiology, K. N. Stevens and M. Hirano (eds), 347-363, Univ. of Tokyo Press.

[10] Zhang, Jialu and Qi, Shiqian (1982): "On the dynamic spectra of speech", FASE/DAGA'82, 997-1000.

[11] Fant, G. and J. Liljencrants (1979): "Perception of vowels with truncated intraperiod decay envelops", STL-QPSR 1/1979, 79-89.