PHONETIC ASPECTS OF STUDYING PHONEMIC SYSTEMS AND SPEECH ACTIVITY

L.V.BONDARKO

Dept.of Phonetics, Leningrad State University
Leningrad, USSR, 199034

ABSTRACT

The goals of the phonetic analysis of speech activity are determined by the properties of the language as a means of communication. Production and perception of speech under normal conditions of communication can only be understood if one is aware of both the characteristics of simple acoustic signals, representing a set of allophones and the rules of their processing.
Of great importance is also a detailed study of phonetic variance of a particular language as well as information on phonetic structure of meaningful units of the language: morphemes and words. A phonetic fund of the Russian language has been described that combines the information specified above. The fund provides phonetic information for speech analysis and synthesis as well as for liguistic study of Russian sound system.

Phonetics as a science dealing with speech sounds can proceed along two distinct paths: one parallels phonology, whose concern is distinctive function of speech sounds, the other parallels psychophysiology, studying mechanisms of production and perception of sound sequences. Phonology has already devised rather strict methods of analysis enabling linguists to study any sign system. Phonology's traditional refusal to analyze phonetic reality has become now a universal characteristic of phonological studies, where the authors either absolutely deny the importance of physical properties of speech sounds or are satisfied with rather primitive phonetic information.

During the 16 years separating us from the XIIth Congress of Phonetic Sciences when Dr D.B.Fry accused linguists of neglecting scientific knowledge little has been changed. Up to now, experimental phonetic studies of speech activity have been non-essential for phonologists, because it is assumed that by contrast with the systematic character of language, speech is individual and, as a consequence, unsystematic. Many present-day phonological concepts exist absolutely independently of phonetic knowledge, are "nourished" by their own postulates, and it seems that no new phonetic information obtained in experimental studies can shake the stability of those postulates.

Another approach to speech sounds is represented in studies dealing with speech production and perception. During the last decades a wealth of research work has been done, where the properties of man, allowing him to use speech so effectively in communication, were of utmost importance. Interest in this information is shown first of all by those research workers who, with respect to linguistics, may be called representatives of neighbouring sciences - physiologists, psychologists, research workers in speech communication and automatic speech recognition, as well as those studying problems of artificial intelligence. This trend using the most perfect experimental methods and statistical analysis has made an important contribution to our concepts, both in the physiology of speech production and in psychophysiology of speech perception, beginning with peripheral processing of speech signals and ending in procedures making decisions by central parts of the hearing system (for a detailed account of a similar approach and extensive bibliography on this subject see, for example, the work by Bernard Delgutte /13/). However, the material used in most of these studies seems to be rather limited, if considered from the point of view of linguistics. For instance, in studying speech perception such simple sound sequences as CV or CVC are often used. Many researchers,on the whole, prefer using synthetic speech-like stimuli which allow them to manipulate the parameters under study, no matter how far their characteristics are from those of real speech signals.

As a result of the development of such diametrically opposed sciences as the phonology and psychophysiology of speech, sciences using their own strict methods and having specific areas of application, the speech activity of man, who used speech signals for communication, is beyond the interests of both the former and the latter trends. Phonologists, as has been said, are not interested in the real manifestations of speech. The psychophysiologists' concern, on the other hand, is limited to the phonetic properties of simple sound sequences.

It becomes expedient, therefore, to study speech activity on the basis of both phonemic concepts and the knowledge of phonetic mechanisms. It is desirable that such studies should be more intenseve than they are today. From a perceptual point of view, information contained in the auditory system of any native speaker may be compared to a curious "puff-pastry", in which without fail there are the following layers:

(a) Certain universal properties of auditory system that are common both to man and animals.

For example, the ability to classify synthetic speech-like vowels according to the values of FI and FII and ascertain "phoneme boundaries"/I6/ was found in experiments on dogs, which allows us to assume that "phoneme boundaries between vowels are determined by some fundamental properties of man's auditory system, not by his linguistic competence" /I/.

(b) Some properties of the auditory system that are determined by man's linguistic ability and his use of articulate speech.

These are properties enabling speakers of various languages to discriminate between the vowels of the basic triangle, to use on- and off-glides of vowels for the identification of adjacent consonants to define the accentual structure of a sound sequence, etc. To these abilities, common to all people, one might add sound symbolism, i.e. the presence of certain psychological and sound associations /22, 24/.

(c) Some specific properties of the auditory system that depend on the speaker's own sound system.

These properties are determined not only by the number of phonemes and their allophonic variation but also by the whole sound system. For example, in experiments on Russian subjects estimating the distance between pairs of sounds it was found that 2 vowels were similarly rated on the basis of the regular alternation they take part in (/l'es/- /l'isá/;/dom / - /damá/), rather than on closeness of their FI and FII values.

No doubt it is very difficult, or even impossible, to find the exact boudaries of the layers. As has been said above, the ability to identify adjacent consonants by on- and off-glides of vowels is a common feature of man (we may assume that animals can acquire this ability as well). However, Russian subjects easily identify hard and soft consonants on the basis of on-glides, because in Russian hard and soft consonants are in phonological opposition, but they show poor discrimination of the place of hard consonants /p, t, k/ and /b, d, g/. French and American subjects, on the other hand, as is well known from the classical studies of the early '50s/12/ do this very well, but the /i/-glides of Russian vowels are not used by them as reliable cues for correct identification of preceding consonants/7/ because softness in these languages is something unknown and phonologically irrelevant.

In any case, investigation of speech activity should be based on the results of experimental psychophysiological studies, but the main function of speech, i.e.conveying meaning, should also be properly considered. This very function allows or even provokes variation of speech signals and hinders successful modelling of man's perceptual properties in automatic speech recognition.

To demonstrate the degree of divergence between physiological and psychophysiological data, on the one hand, and the results of speech activity, on the other hand, two figures are given. In Fig.I(a and b) Russian consonants are shown in two different feature spaces. Fig.Ia demonstrates a geometrical arrangement of the consonants in a space of articulation features/I8/, which seemed to be a convenient way to show the relations between Russian consonants and their features. Fig.Ib demonstrates an arrangement of Russian consonants in a space of psychological features comparable with such oppositions as hard-soft and continuous-discontinuant/24/. What a great difference between the geometrical linguistic pattern and the real arrangement of the consonants in the perceptual space!

Fig.2 (a and b) shows schematic representation of the vowels used as stimuli in experimental phonetic studies: Fig.2a demonstrates synthetic four-formant stimuli used in numerous works aimed at ascertaining "phoneme boundaries"/I6/, Fig.2b shows Russian stressed and unstressed vowels. As can be seen from the comparison of steady-state synthetic vowels (400 msec long) and transitory natural vowels(varying in duration from 200 to 50 msec), the differences between them are so great that one cannot assume that in processing and identification of these two groups of stimuli the same mechanisms are used.

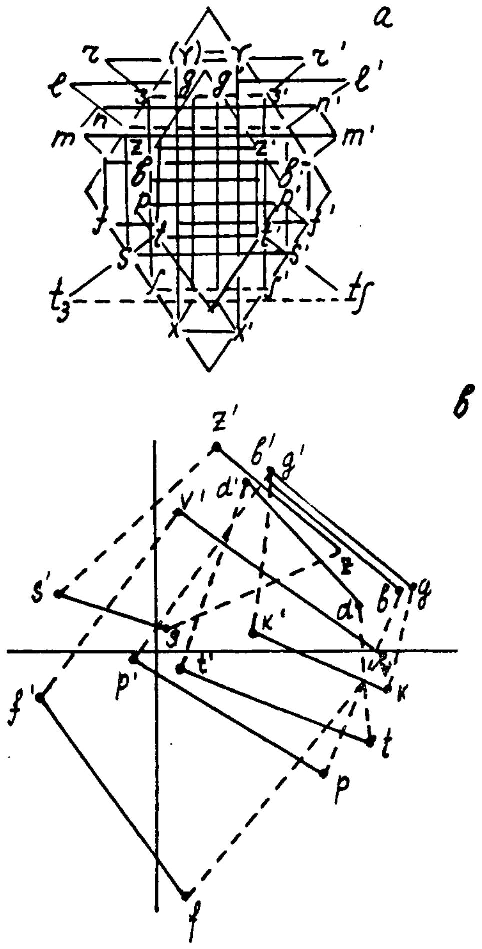


Fig.I. Russian consonants in a space of features.
(a) Russian consonants in a space of articulatory features, demonstrating the arrangement of the consonants within phonemic system/I8/;
(b) Russian consonants in a space of perceptual features related to the features "hard-soft" and "continuous-discontinuant"/24/.

Thus, in investigating speech activity, when natural languages are studied, one should consider the following: (I) psychophysiological properties of man, (2) how these properties are realized in a particular phonetic system, (3) in what way the phonemic system as one of the upper levels of the linguistic structure effects speech activity.

Such an approach to the study of speech activity will undoubtedly cause the disapproval of both phonologists and representatives of the natural sciences. Let us take courage and borrow what we need from these opposite provinces!

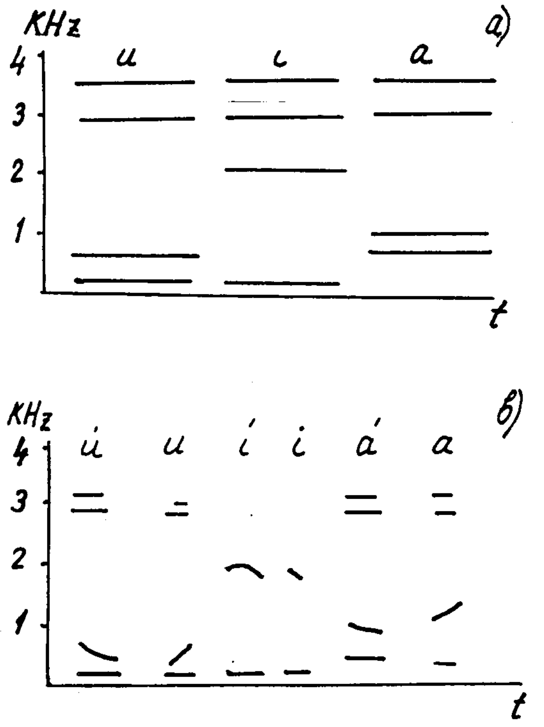


Fig.2 The scheme of formant characteristics of the experimental vowels.
(a) synthetic vowels
(b) natural Russian vowels, both stressed and unstressed, occurring in different phonetic contexts.

Phonemic terminology, due to thorough elaboration of the main concepts of the field, is more precise than psychophysiological one. Let us consider some of the terms.

I. The Phoneme is the minimal unit of the expression system which is able to constitute and distinguish meaningful units, i.e.words and morphemes /25/.The term "psychophysiological phoneme", as used by psychophysiologists, is less precise: psychological phonemes are defined as units corresponding to non-overlapping areas in the space of acoustic parameters of the speech signal. The number of these phonemes exceeds that of linguistic phonemes in any language.However, it is not known exactly how great this excess is /I6, p.82/. Fig.3 presents the phoneme boundaries of psychological vowel phonemes in relation to the arrangement of Russian vowels in FI-FII plane(Fig.3a), as well as data on possible changes in FI and FII of the vowels as a result of coarticulation with adjacent consonants(Fig.3b). Comparison of these figures shows that psychological phonemes, as revealed in experiments on synthetic vowels, do not correspond to the arrangement of natural vowels based on their acoustic and perceptual characteristics.

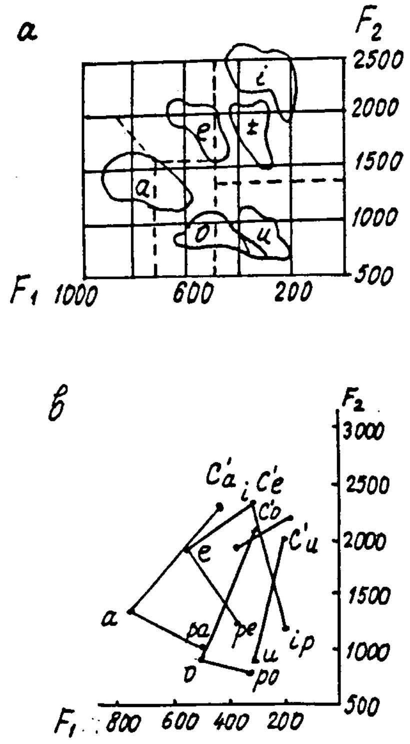


Fig.3 Areas of F-values of natural Russian vowels and "phoneme boundaries"

a) arrangement of Russian vowels in FI-FII plane and as related to "phoneme boundaries" obtained in experiment with synthetic vowels;

b) possible FI and FII values of glides with respect to stationary segments of the vowels:/a, o, e, ɨ ,u, i/ - stationary segments; /c'a/ etc. - i-glides of the vowels preceded by soft consonants; /pa/,etc.- glides of the vowels preceded by labials.

What are the correlates of phonemes in speech activity? From the viewpoint of speech production, the minimal unit of pronunciation is an open syllable(CV,CCV), in which the information about the consonant(s) and the vowel is contained nearly in the whole of the syllable /4, 6/; neither is it the minimal unit from the viewpoint of speech perception, because some phonemes and classes of phonemes cannot be identified without minimal phonetic context /10/. Finally, if we consider the main function of phonemes, which is to constitute and distinguish meaningful linguistic units, the phoneme does not appear to be their obligatory element: it is a well-known fact that it only seems to a subject that the two words differ in some sound segment /9/; it is also known that man can "hear" the sound in a sound sequence(more often in meaningful units) even if it is not present at all.

We may speculate that the phoneme as the minimal unit of the expression system is only necessary to put in good order conceptions about the structure (arrangement, set-up) of meaningful units, and such a conclusion gives grounds for the very bold but false claims that the phoneme as an operational unit of linguistic analysis bears no relation to speech activity of native subjects. Researchers studying speech activity have already gone through the period when the concept of the phoneme seemed to be a logical device which did not have any correspondence with speech material/20/. Now one can safely say that the phoneme is as real as other units of linguistic structure, such as the morpheme, the word, etc. Evidence of its reality for native subjects is quite plentiful and is discussed on a large scale in experimental phonetic studies. Let us consider some of the facts in the sequence that seem to be the most natural[x].

A phonemic system is represented in the brain of native subjects as an organized structure /4, 17/. Phonemic classification is used by native subjects for systematization of sound units (in speech perception), which greatly vary in their parameters, and for coding programs of essential articulations (in speech production). Phonetic realization of a phonemic sequence, as a specific phenomenon of any language, is regulated by a whole set of rules (the articulatory basis)and leads to certain peculiarities of perceptual processing of acoustic signals(the perceptual basis)/8/.

2. The phoneme and its distinctive features. Since the middle of the 20th century, this problem, due to the scholarly work of Jakobson, Fant, Halle,etc, has become central inphonological discussions and experimental phonetic studies. Linguists concern themselves first of all with the idea of regarding a distinctive feature as an independent unit of the expression system/3, 14/. Of utmost importance for phoneticians is the study of articulatory and acoustic correlates of distinctive features, as well as procedures for obtaining information about distinctive features in speech perception /6, 19/.

No less important, however, is the problem of the degree of manifestation of linguistic and proper phonetic characteristics of distinctive features in native subjects' speech activity. Is the phoneme represented by a constant set of distinctive features or does it vary from one context to another? As a matter of fact, the answer to this question is closely connected with a different problem: is the set of distinctive features of a phoneme based only on the phonemic oppositions existing in a given language or does the phonemic system itself effect the procedure of attributing distinctive features to phonemes? For example, are the phonemes / k', g', x'/ in the words /rúk'i/ "hands", /g'imn/ "anthem" and /x'ítrыj/ "cunning" soft or is their softness an allophonic variation determined by the character of the following vowel? Are the affricates /c/ and /č/ voiceless or are they lacking characteristics of the feature "voiceless/voiced"? Experiments on speech activity of Russian subjects demonstrate that the set of distinctive features of each phoneme is ascertained on the basis of knowledge of the phonemic system as a whole, and if the feature in question is distinctive for most phonemes, it is also attributed to the phoneme which is not opposed to others by this feature. Thus, /n/ is a forelingual nasal phoneme, though in Russian there is no opposition of forelingual and backlingual nasal consonants; backlingual /k', g', x'/ in the words given above are soft phonemes but not the allophones of hard /k, g, x/. This conclusion is supported not only by numerous experiments where subjects make phoneme discriminations of such sounds, but also by the indisputable ability of the subjects to mark the "unnaturalness", "anomaly" of those stimuli which satisfy our phonological concepts about distinctive features but do not meet the phonetic requirements concerning the correlates of the distinctive features. It is noteworthy that distinctive features are abstractions: each distinctive feature has a great number of phonetic correlates, and native subjects can use any combination of these correlates for the identification of the distinctive feature in question. The abstract nature of distinctive features is also supported by the fact that the character of phonemic oppositions is determined not by the degree of phonetic manifestation of distinctive features but by phonemic relations proper. For example, Russian nasal and soft consonants having distinct phonetic characteristics are in phonemic oppositions to each other as unmarked and marked members, the fact having been definitely confirmed in perceptual experiments on Russian subjects/5/.

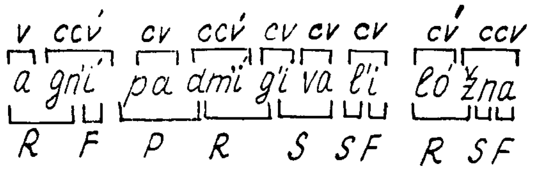
It follows from what has been said above, that, on the one hand, native subjects behave contrary to the phonological conceptions about phonological operation (which have been developed in phonology). On the other hand, being tolerant to the varying charateristics of speech sounds, native subjects use an effective set of rules allowing them to proceed from a variable phonetic picture to a sequence of phonemes, thus constituting the expression of meaningful units. This, in turn, means that native subjects use their own phonemics, which only partly coincides with that of a phonologist.

3. The Phoneme and the Morpheme.

From the viewpoint of classical phonology one of the main functions of the phoneme is its ability to discriminate morphemes. Morphemic criteria are also used both in determining the independent status of a phoneme and in making decisions as to mono- or biphonemic interpretation of a sound sequence, as well as in classifying phonemic oppositions. Indeed, the morpheme is the minimal meaningful linguistic unit and the ability of the phoneme to function as the morpheme's exponent is a very important evidence of the linguistic segmentation of the acoustic continuum into minimal segmental units, i.e. phonemes.

It is necessary to point out that experimental phonetic studies are very rarely based on conceptions that combine both phonemic and morphemic levels of analysis.

But it is quite clear that a description of human speech activity dealing with natural coherent utterances should not ignore the principal rules that govern the sound(phonetic) structure of morphemes. Russian language studies have excited an ever-growing interest in this problem. Every chain of sounds can be represented phonetically, for Russian at least, as a sequence of open syllables, and from a morphological viewpoint, as a sequence of morphemes: affixes, roots and inflections (Fig.4). Segmentation of the utterance into open syllables is used in applied studies and is confirmed by experimental data/2/.



V CCV́ CV CCV́ CV CV CV CV́ CCV
a gn'í pa dm'í g'i va l'i l'ó žna
R F P R S SF R SF

Fig.4 A sound sequence segmented into open syllables (at the top) and morphs (at the bottom).
R-root, P-prefix, S-suffix, F-flexion

[x] Taking this opportunity to acquaint wide circles of phoneticians with studies little known outside this country, I will mainly mention here the results of studies of Soviet phoneticians.

In order to gain an understanding of how this segmentation can be rendered morphologically, that is, how to transform a sequence of syllables into a sequence of morphemes, a special study was carried out.

Each syllable was considered from the point of view of its morphological segmentation, producing the morphemic syllable structure. This made it possible to formalize the transfer from syllable segmentation to morphemic segmentation/21/.

We have every reason to believe that the relation between the "morphemic structure of the syllable" and the "syllabic structure of the morpheme" has psycholinguistic correlation and it can be experimentally investigated as an element of human speech activity.

A close study of phonetic properties of morphemes revealed certain facts which are important in the evaluation of morphological criteria used in phonology.

First of all, not every morpheme is a meaningful unit. Secondly, many morphemes differing in their sound pattern have the same grammatical meaning (we are not considering root morphemes here, of course). These facts challenge the exclusiveness of morphological criteria in phonology.

Nevertheless, rules governing the combination of phonemes(sounds) into morphemes and their arrangement into word-forms are language specific; they form one of the building blocks of what is meant by "language comprehension" or "information about higher levels" in constructing speech recognition models.

Systematic studies of the Russian Language Dictionary[x] where each word is represented as a sequence of morphemes/23/, have made it possible to obtain quantitative data for linguistic interpretation of the predictability of phonemes both in a dictionary and in speech flow.

110 thousand words were organized into 10 thousand word-families having the same basic root.

The phonetic analysis[x] of these roots revealed the following:

1. Approximately half of the roots contain a stressed vowel.
2. The probability of the occurrence of a stressed vowel in the root depends on its quality:

stressed vowel	Number of syllables in the root					
	1	2	3	4	5	6
1	2	3	4	5	6	7
a	10250[xx]	3868	1032	126	22	2
e	6299	2401	572	67	7	
o	8845	3371	597	56	3	
i	3866	1461	641	75	1	
ɨ	1429	360	22	1		
u	4105	1049	279	26		

[x] The data presented below were obtained by computer analysis of the dictionary.

[xx] The absolute number of roots containing this vowel

The table shows that there is a consistent relationship between the number of syllables in the root and the frequency of roots: the longer the roots, the fewer their number. Of frequent occurrence in stressed syllables are /a/,/e/ and /o/.

3. The probability of occurrence of unstressed vowels in the root morpheme varies: the more frequent are /a/ and /i/, less frequent are /u/ and /ɨ/.
4. The description of root morphemes in terms of generalized phonetic structure(C and V) revealed 29 different combinations, the more frequent of them being CVC, CCVC, CVCC and CVCVC.
5. Historical alternations of vowels(i.e. changes in the phonemes of the root morphemes which cannot be explained by phonetic rules of modern pronunciation) occur in approximately 3% of all roots,alternations of consonants - in nearly 6%.

Most prefixes, as our investigation revealed, contain an unstressed vowel. This indicates that a stressed vowel in a prefix is an exception rather than the rule,which any Russian speaker can use in phonemic identification of a vowel in a prefix (the prefix *бес* = b'is for example, occurs in the dictionary 409 times, whereas the prefix *бéс* =/b'es/ only 3 times; the prefix *от* =/at/occurs 2045 times, whereas *óт* =/ot/ is found only 32 times).

The computer based dictionary has made it possible to determine the frequency of cases in which considerable vowel reduction occurs and, as a consequence, the simplification of the phonemic sequence. In Russian vowel reduction is often found in post-tonic parts of the word. A special computer programme enabled us to extract all unstressed fragments given in the dictionary; 67% of word-forms contain such fragments in their structure; every 311 fragments out of the 1200 which are possible occur in 90% of all word-forms having post-tonic parts. Research is under way to establish the relationship between the phonetic and morphological properties of these fragments.

These studies may seem to have no direct reference to the investigation of human speech activity, but this is not so.The "language competence" of a speaker implies not only his ability to make use of phonemic and phonetic distinctions of his language, but also to understand the meaning of the phonetic complexes. The system of basic knowledge which forms the mechanisms of recoding sounds into meaningful units includes also the comprehension of rules of word-formation which enable the speaker to make lexical and grammatical interpretation of a phonetically vague series of sounds.

The study of regularities governing the formation of the phonetic structure of an utterance in a particular language is one of the necessary constituents in the investigation of human speech activity. The study of speech perception, exhaustive as it might be,will give us information only about the potential capabilities of human speech activity, whereas information about the predictability of occurrence of phonetic patterns of meaningful units makes it possible to put forward a reasonable hypothesis about the mechanisms which enable the listener to predict one element of speech by the other and the abilities of the listener on which the speaker can rely when he allows himself certain deviations from the "ideal" phonetic pattern of the utterance he produces.

In fact, the problem of defining the acoustic cues for the transformation of the acoustic continuum into a succession of discrete elements in speech perception or automatic recognition by a computer cannot be solved without reference to all possible modifications of the whole word. These modifications are governed by certain rules. This means that in order to give a thorough and comprehensive phonetic description of the sound system of a particular language, it is necessary to take into consideration both allophonic modifications caused by the phonetic environment and modifications due to tempo variation, the intonation pattern and the placement of the word in the phrase (variability caused by deviations from standard pronunciation is the subject of a special study).

So the problem is to create a phonetically representative speech material that will enable us to obtain necessary information.

We will use the Russian language to illustrate how it can be done.

As mentioned above, there are statistical data on the open syllable in Russian: 200 most frequently occurring syllables account for about 80% of any Russian text /15/. These are sequences of CV, CCV and CCCV, both stressed and unstressed.

Fig.5(a,b,c) shows the relative frequencies of syllables with various vowels (in per cent) and the relative frequencies of stressed and unstressed vowels in CV, CCV and CCCV sequences.

It is evident that syllables with the vowels /a/,/i/ and /u/ prevail in the group of most frequently occurring syllables; the number of syllables containing stressed /o/ and /e/ is considerably greater than that of syllables with unstressed vowels; other vowels were more frequently found in unstressed syllables.

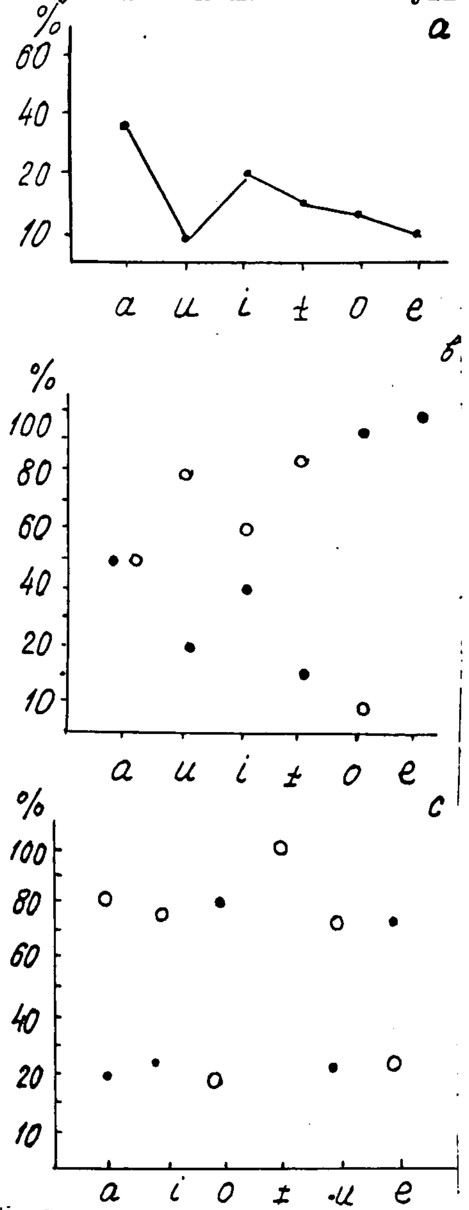


Fig.5 Relative frequencies of syllables in Russian:
(a)containing various vowels
(b)and (c)containing either stressed or unstressed vowels (filled and unfilled circles). Data on CVsyllables are given in (b), and on CCv and CCCV syllables in (c).

The occurrence of consonants and their clusters in these syllables is in accord with the known statistical data for the Russian language /II/. The creation of a phonetically representative text is necessary not only for experimental studies of speech activity but rather it should serve as a component of the bank of phonetic data obtained for any language in various computer techniques for processing and storage of phonetic information. This text, together with simple phonetic sequences like CV and CCV, will provide necessary information both for theoretical research and applied studies of speech signals. In its "ideal" form this bank of phonetic data must contain the following four blocks(Fig.6):

I. Block of physical(acoustic) information proper, which characterized distribution of acoustic parameters at the allophonic level as well as their combination within a word-form,

II. Block of phonetic properties of final constituents of the word-form(i.e. morphemes).

III.Block of phonetic properties of the word-form as a combination of morphemes: it allows sequences of sounds which are impossible within a morpheme.

IV. Block of phonetic properties of a text of any length.

The first of these blocks seems to be the simplest since it transforms the recorded text into digital representation and performs segmentation of the computer version into "fragments" in accordance with the prescribed transcription. One of the disputable questions here is the number of informants necessary for obtaining a statistically adequate and reliable corpus. They may be few ,but a preliminary selection with the help of an experienced phonetician is necessary, since he is able to assess both the standard of pronunciation and the degree of its individual variability. The computer version of phonetic material makes it possible to obtain any information which may be interesting for a phonetician and also makes possible accurate comparison of data obtained by other linguists.

The second block in which the information about phonetic properties of morphemes is stored, also requires the use of the computer based dictionary segmented into morphemes and computer programmes which make it possible to obtain the necessary information.

The realization of the third block is also impossible without the computer based dictionary. One of the best examples of such dictionary is the above mentioned Russian Derivational Dictionary by Dean S. Worth (et al.) which gives information about predictable combinations of derivational morphemes in Russian .

And, finally, the block of phonetic properties of a text which in fact is the algorithm for an automatic transcription which converts any orthographic recording into a sequence of phonetic symbols. Since every phonetic symbol is assigned its possible acoustic realizations in the first block, such a transcriber should provide an optimal synthesis of the text.

The realization of the bank of phonetic data as it is described here is a very difficult and responsible task.Only a few fragments of each of the four blocks have been realized up to now.But our confidence in the necessity of this work is justified by the interest aroused by this idea in linguists and representatives of applied sciences. In some respect, to create a bank of phonetic data means to construct a model of human speech activity.

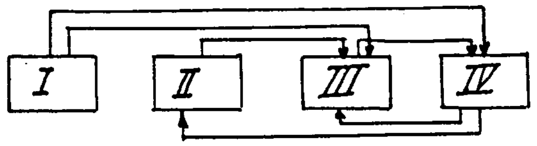


Fig.6 A scheme of phonetic base data with linguistic information considered I,II,III, and IV - blocks of phonetic properties.Upper lines with arrows indicate the most closely tied blocks providing for analysis and synthesis of speech. Lower lines with arrows indicate the direction of information transmission in linguistic processing of speech.

Fig.6 shows the structure of the bank of phonetic data and the relations that seem important both from the linguistic point of view and from that of the investigation of human speech activity.

The block of acoustic data which contains information about the realization of sound units may provide data for a reliable and thorough description of the acoustic cues of the distinctive features and for the description of standard pronunciation. Segments from this block may serve handsomely as transcription symbols, since each of them is assigned information about the position of the corresponding allophone. Phonetic transcription provides information about potential phonetic variability of each phoneme. It is important that these segments can also be used for comparison as "ideal" models.

Classification of final constituents of word-forms - morphemes - in terms of phonetic and phonological units is extremely important for linguistic analysis proper, since we know very little as yet about the quantitative aspect of the relationship of the two types of linguistic units, the phoneme and the morpheme. How often does a phoneme perform its distinctive function, i.e. how many morphemes are distinguished by the phoneme alone? Which phonemes are the more active in this respect and which are less so? How often do the morphemes which differ in various respects have the same phonemic make-up? How many morphemes with the same grammatical meaning differ in their phonemic make-up? Even the listing of these problems makes it clear that information cannot be obtained without the use of computer techniques which are employed not just because of fashion but as vital research necessity.

From the linguistic point of view, information about the phonetic properties of a word-form as a combination of morphemes is also of some interest, since it enables us to obtain quantitative data that characterize processes of forming a phonetic pattern of lexical items. The occurrence of definite classes of phonemes in definite positions within a word-form is a universal phenomenon, but only by comparing inherent phonological properties of sound units with their functions within the word-form and the morpheme can we obtain new data in this respect. These phenomena which occur within the word-form may even give specialists in the field of diachronic phonetics something to think about.

Finally, an automatic transcriber performs the analysis of any text in terms of the first three blocks, and thus not only verifies the various properties of sound signals but also enriches the content of these blocks with the data of the text.

In conclusion, I would like once again to draw your attention to the necessity of the investigation of those specific aspects which are pertinent to human sppech activity. The development of new and reliable methods is only beginning. To these we may refer the investigation of the perception of foreign language sounds (familiar and unfamiliar to the listener), the comparison of results of the identification of the same speech stimuli(synthetic sounds, for example) by speakers of different languages, the analysis of perceptual abilities of speakers of those languages which have different rules governing the combination of phonemes into meaningful units (Russian compared to Turkish, with its law of vowel harmony). The modifications of Russian sound units produced by the speakers of different languages is a good model of the influence of one's native language on one's speech activity in a foreign language.

How to investigate these fine mechanisms of the influence of the linguistic system on human speech activity is the problem which requires close attention of all specialists interested in obtaining new data about properties of speech production and perception.

REFERENCES

I. Baru A.V.,"Slukhovyje centry i opoznanije zvukovyh signalov"(Ear Centers ans Sound Signals Recognition),Leningrad, I978, I92 p.

2. Beliavskij V.M. Algoritm avtomatičeskoj segmentatsii reči na otkrytyje slogi"(Algorithm for Automatic Segmentation of Speech into Open Syllables), Trudy ARSO-9, Minsk, 1976,p.56.

3. Benveniste E.,"Problème de Linguistique générale"(Problems of General Linguistics), Paris, I966.

4. Bondarko L.V.,"Fonetičeskoje opisanije jazyka i fonologičeskoje opisanije reči"(Phonetic Description of Language and Phonological Description of Speech), Leningrad, I98I, 200 p.

5. Bondarko L.V. Nekotoryje zamečanija po povodu markirovannosti-nemarkirovannosti členov fonetičeskih protivopostavlenij"(Some Remarks on Marked-Unmarked Members of Phonetic Oppositions), in: Issledovanija po fonologii, Moskva, I966,p.394-600.

6. Bondarko L.V., "The Syllable Structure of Speech and Distinctive Features of Phonemes", Phonetica, I969, 20,pp.I-40.

7. Bondarko L.V., Kiushkina O.M. Pertseptivnaja baza jazyka i obrabotka fonetičeskoj informatsii"(Perception Principles of language and Processing of Language Information),Ivanovo, I98I,pp.I4-24.

8. Bondarko L.V., Lebedeva G.N.,"Opyt opisanija svojstv fonologičeskogo sluha"(Some Principles of Description of Phonological Perception),Voprosy jazykoznanija, I983, 2, pp.9-I9.

9. Bondarko L.V., Verbitskaya L.A. Factors underlying phonemic interpretation of phonetically non-defined sounds in: Auditory Analysis and Perception of Speech, AP, 1975,p.I77-I90.

I0.Bondarko L.V., Svetozarova N.D. O vosprijatii bezudarnyh slogov(On Perception of Unstressed Syllables), in: Fonetica, fonologija, grammatika, Moskva, I97I,p.30-43.

II.Bondarko L.V., Zinder L.R.,Stern A.S. Nekotoryje statističeskije harakteristiki russkoj reči(Some Statistical Characteristics of Russian Speech), in: Sluh i reč v norme i patologiji, Leningrad, I977,p.3-I6.

I2.Cooper F.S., Delattre P, Liberman A.M. Borst J.M.,Gerstman L.J. Some Experiments on the Perception of Synthetic Speech Sounds", JASA, I952,vol24, p.597-606.

13. Delgutte B., "Codage de la parole dans le nerf auditif", These de doctorat d'etat des Sciences Naturelles, Presentee a L'Universite Pierre et Marie Curie pour obtenir le grade de docteur des Sciences, 1984.
14. Džaparidze Z.N. O merizmatičeskom urovne lingvisticeskogo analiza(On the merismatical level of linguistic analysis), in: Zvukovoj stroj jazyka, Moskva,p.98-103.
15. Jolkina V.N.,Judina L.S. Statistika slogov russkoj reči(Statistics of Russian Speech Syllables).In: Vycislitelnyje sistemy, 10, Novosibirsk, 1964,p.58-78.
16. Fiziologija reči. Vosprijatije reči čelovekom(Speech Ariculation and Perception),L.A.Čistovič et al.,Leningrad, 1976, 386 p.
17. Galunov V.I. Issledovanije subjektivnogo predstavlenija gruppy russkih soglassnyh metodom semantičeski protivopolozhnyh par(Research on Subjective Perception of a Group of Russian Consonants by means of Semantically Opposed Pairs).In: Analiz rečevyh signalov čelovekom.,Leningrad, 1971.
18. Moisejev A.I. Sootnositelnost' soglasnyh zvukov v sovremennom russkom jazyke(Correlation of Consonantal Sounds in Modern Russian).In: Studia Slavica(Hungary), 1969,t.XV,p.207 - 218.
19. Problemy i metody eksperimentalno-fonetičeskogo analiza reči(Problems and Methods of Experimental Phonetic Analysis of Speech), Leningrad, 1979, 151 p.
20. Reč.Artikulacija i vosprijatije. (Speech, Articulation and Perception L.A.Čistovič et al.,Moskva-Leningrad, 1965, 241 p.
21. Urovni jazyka v rečevoj dejatelnosti. (Language Levels in A Speech Activity L.V.Bondarko, Leningrad University Press, 259p.
22. Voronin S.V. Osnovy fonosemantiki (Foundations of phonosemantics), Leningrad, 1982.
23. Worth D.S., Kozak A.S., Johnson D.B. Russian Derivational Dictionary, New York, 1970, 747 p.
24. Žuravl'ov A.P. Fonetičeskoje značenije (Phonetic Meaning), Leningrad, 1974, 160 p.
25. Zinder L.R. Obščaja fonetika (General phonetics), Moskva, 1979, 312 p.

ARTICULATORY MEASUREMENTS BY MAGNETIC METHODS

GEORGIOS PANAGOS HANS WERNER STRUBE

Drittes Physikalisches Institut, Universität Göttingen,
Bürgerstr. 42-44, D-3400 Göttingen, Fed. Rep. of Germany

ABSTRACT

We present two methods, which use alternating magnetic fields, for measuring articulatory activities. The first method uses homogeneous magnetic fields and as induction coils a flat flexible rectangular coil and a magnetic potentiometer. The second method uses inhomogeneous magnetic fields and small dipole receiver coils. Some examples of comparative measurements of various articulatory movements during speech production are presented here in order to demonstrate the applicability of the system.

INTRODUCTION

One of the most important problems in the phonetic sciences is the measurement and theoretical modelling of the articulatory motions and their relation to the acoustic speech signal.

For the direct registration and measurement of the articulatory motions several methods have been developed in the past. However, most of them are very expensive or/and have some undesirable bioeffects [2]. Some other techniques, based on pulsed-echo ultrasound [3], are not tissue invasive but disturb the articulation a little, and they are not suitable for measuring all the articulatory parameters.

We developed two magnetic methods, by which we are able to measure the tongue and jaw movements. These methods hardly disturb the speech production, are biologically safe and not expensive. We also present here some comparative investigations between various quantities related to the articulatory movements.

METHODS

Homogeneous fields

The first magnetic method uses homogeneous fields, generated by three orthogonal Helmholtz-coil pairs with different frequencies (15, 17.5 and 20 kHz) surrounding the head of the subject. One of the coil pairs is not quite "Helmholtz", but serves only for correcting purposes. The homogeneous fields do not allow to measure absolute positions, but vectorial distances can be measured more exactly. We used two types of receiver coils: a magnetic potentiometer (MPM) and a flat flexible coil (FC).

By the MPM, i.e. a long, thin and flexible coil, we can measure the vectorial distance between its ends with the help of the voltages induced by the three fields in the coil. These voltages depend only on the position of the MPM's ends (Fig. 1). We place and fix the coil's ends on the upper and lower incisors. Thus we can measure the distance of the upper to the lower jaw in the midsagittal plane.

The FC (with one or two rectangular windings, which are embedded between two flexible plastic sheets) is attached to the tongue surface in order to measure the tongue "curvature" and the angle of the tongue tangent relative to the Frankfort horizontal line in the midsagittal plane. By the induction of the same homogeneous fields as above, a two dimensional vectorial distance of the short edges can be measured (Fig. 2), which can be interpreted as the curvature and the tangential direction of a tongue surface element. An outline of the vocal tract with the positions of the coils is shown in Fig. 3.

A longer flat coil allows us to measure the distance palate-tongue, if one edge of the FC is attached to the immoveable palate and the other to the tongue. Thus we had the possibility to obtain measurements relative to the head of the subject. But this technique may be too disturbing during speech production.

Dipole fields

The second method uses four dipole transmitter coils placed on the edges of a square in the median plane of the subject's head (Fig. 4). The opposite coils are driven at the same frequency but opposite phase, so that the field strength equals zero at the centre of the square. Each coil has 22 cm distance from this centre. Another pair of transmitter coils with their axes perpendicular to the median plane are used for correcting purposes. This (circular) coil pair generates a nearly homogeneous field about the centre of the above-mentioned square, which is approximated with a 4th-order polynomial [4]. These coils correct the induced amplitude for a possible deviation of the receiver-coil axes from the normal of the median plane. The receiver coils attached to the tongue surface are about 1 mm thick and 3 mm long, with a ferrite core and about 400 windings. With such miniaturized dimensions of the coils no disturbance during speech is given (they are smaller than the pellets in [2]). They are pasted on a plastic strip, which is then attached to the tongue. This facilitates obtaining the proper orientation and position of the coils, protects the coil's leads and enhances reproducibility. Another receiver coil is placed on a immoveable point of the head (e.g. upper incisor) used as reference point. Since we use synchronous demodulators for the detection of the receiver signals, we can distinguish the sign of the field strenghts, unlike [1]. The field strengths are converted into Cartesian coordinates by a zero-detection iterative technique. The calibration constants can also be estimated by a similar iterative technique.

The electronic section of the apparatus consists of a transmitter (three sinewave oscillators) and a receiver lock-in amplifier, which separates the three induction voltages of the receiver coils by a synchronous demodulator and three 4th-order Bessel filters. The advantage of the lock-in circuit is that it keeps the disturbing voltages to a minimum and allows the distinction of the sign of the induced voltage.

The errors of both methods are below 1 mm. Simultaneously with the motions of the coils the speech signal is picked up by a Sennheiser microphone (MKH 105), in order to compare the signal with the articulatory parameters. For preventing any acoustical disturbances all the measurements have been done in an echo-free chamber. All the signals (speech and those from the lock-in amplifier) are digitized and fed into a laboratory computer.

MEASUREMENTS - DISCUSSION

With the homogeneous-field apparatus tongue and jaw movements have been measured in VCVCV utterances. Comparative investigations have been done of the time which is required from the start of the movements until the onset of phonation and the maximal velocity or the maximal amplitude of the movement; between the latter two quantities we found a roughly linear dependence of the data (Fig. 5). We measured the coordination between the jaw and tongue movements. Specifically, we did some investigations about the repetitive production of /d/, /t/, /p/, combined with /a/, /u/, /o/ and /au/. In the comparisons of tongue and jaw movement we found a positive correlation (an example is shown in Fig. 6). Jaw and tongue measurement examples by the dipole-field method will be presented at the Congress.

These methods have the following advantages: negligible disturbance during speech, they are inexpensive, and they are biologically safe. The disadvantage of the homogeneous-field method is that it is impossible to registrate parallel displacements of the coils (thus no absolute positions can be measured). This disadvantage is avoided by the second method (the inhomogeneous fields). Unlike the first method, we can thereby measure absolute positions in the mouth, with the transmitter coils, having fixed positions about the head of the subject, as reference frame. In further development of these methods we combine them with optical methods (measurements of the lip opening area [5]) and collect a large amount of data for application in articulatory and speech modelling.

REFERENCES

[1] van der Giet G. "Computer-controlled method for measuring articulatory activities", J. Acoust. Soc. Am. **61**, 1072-1076 (1977).

[2] Fujimura O. "Modern methods of investigation in speech production", Phonetica **37**, 38-54 (1980).

[3] Keller E., Ostry D.J. "Computerized measurement of tongue dorsum movements with pulsed-echo ultrasound", J. Acoust. Soc. Am. **73**, 1309-1315 (1975).

[4] Nagaoka H. "Magnetic Field of Circular Currents", Phil. Mag. **41**, 377-388 (1921).

[5] Kretschmar J. "New method for fast continuous measurement of the lip opening area", J. Acoust. Soc. Am. **62**, 474-476 (1977).

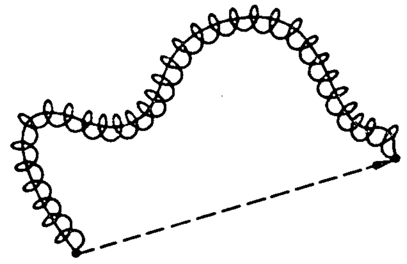


Fig. 1: Schematic of the magnetic potentiometer; dashed line indicates the vectorial distance between MPM's ends.

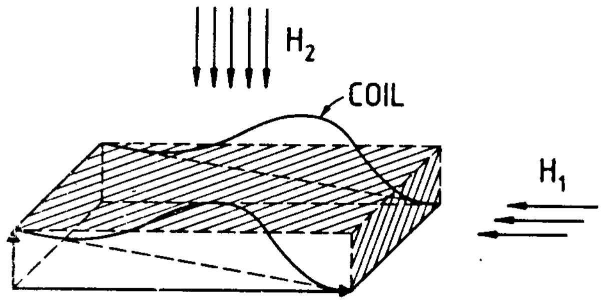


Fig. 2: Schematic of the flexible flat coil. The hatched planes indicate the projections normal to the fields. The dashed line indicates the measured vectorial distance.

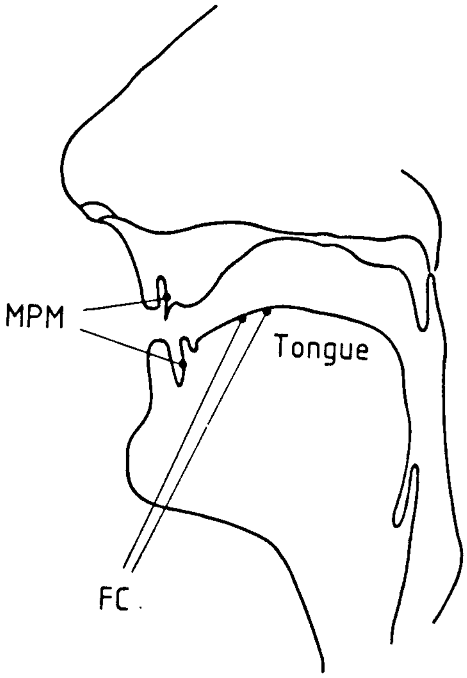


Fig. 3: Positions of the flat coil and the magnetic potentiometer in the vocal tract.

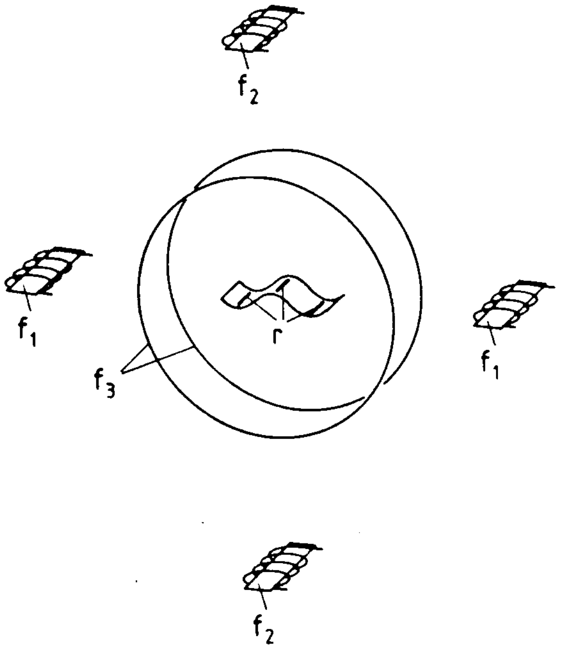


Fig. 4: The dipole coil system. Transmitter coil pairs with frequencies f1, f2, f3; r: three receiver coils on a plastic strip in the middle of the system.

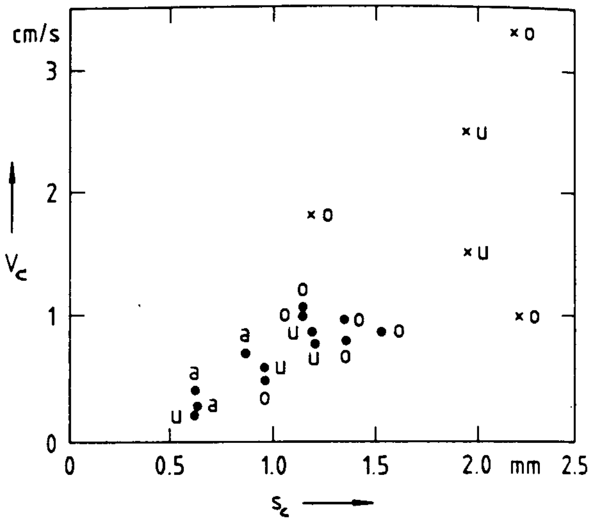


Fig. 5: Velocity v_c of tongue's "curvature" in dependence of the maximal amplitude s_c of the "curvature" of the vowels /a/, /o/ and /u/ combined with the consonant /d/; x, □ ,● : the measurement points of the three subjects.

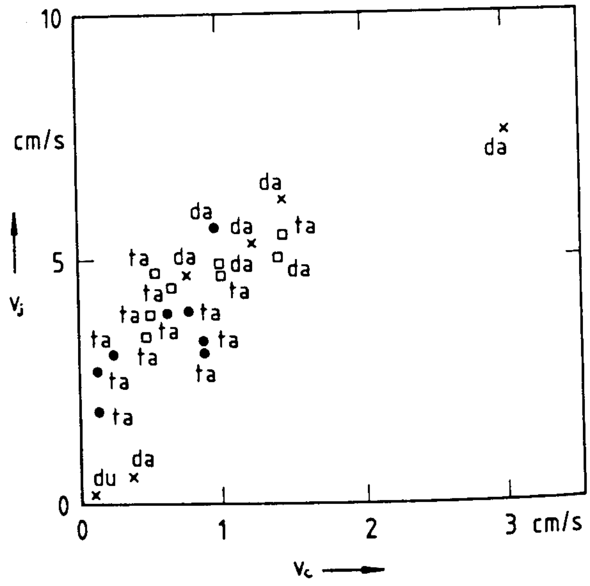


Fig. 6: Velocity v_j of the jaw movement in dependence of the velocity v_c of tongue's "curvature" (velocity: maximal velocity at the beginning of the vowel /a/); x, □, ●: the measurement points of the female and the two male subjects respectively.

ELECTROMAGNETIC ARTICULOGRAPHY *

A New Approach to the Investigation of Palatalization in Russian

Gabriel HONG, Paul W. SCHÖNLE, Bastian CONRAD

Department of Clinical Neurophysiology
University of Göttingen
3400 Göttingen, Federal Republic of Germany

12

ABSTRACT

Electromagnetic articulography (EMA) is a non-invasive method of investigating the movements of articulators inside and outside of the vocal tract. With this method, it is possible to register the articulatory movements of the tongue during on-going speech and to reveal the dynamic, speech-physiological aspects of palatalization in Russian.

INTRODUCTION

Palatalization is a linguistic phenomenon of wide occurrence. In a slavic language such as Russian, it is especially prevalant and has a (mor)phonemic significance. There are some theoretical linguistic investigations [1,2,4,5,7,8,9] and experimental studies [3,5,6] on palatalization in Russian. However, a direct recording of articulatory movements of the tongue, which modulate the vocal tract configuration underlying palatalization, is still lacking. EMA offers the possiblity of routinely sampling large amounts of speech data for empirically testing theories of palatalization specifically, as well as modelling speech production in general.

METHOD

EMA (Fig.1) is based on the physical principle that a magnetic field of an oscillating dipole decreases as a cubic function of increasing distance from its center [10,11,12]. The distance between a receiver and a transmitter coil can be determined by measuring the voltage induced in the receiver if the axis of both coils are parallel. Use of two transmitters allows calculation of the x / y coordinates of the receiver if the receiver does not tilt or twist. A third transmitter corrects falsification of the signal due to tilting or twisting of the receiver and enables precise localization of the receiver by iterative solution of non-linear equations.

The three transmitters (4 cm x 2 cm) are fixed on a helmet and positioned around the head of the subject in the midsagittal plane. The receiver coil (2 mm x 4 mm) can be attached to the tongue with a tissue adhesive (Histoacryl blau). The signals from the receiver are fed into the analog circuitry via a thin copper wire (0.13 mm).

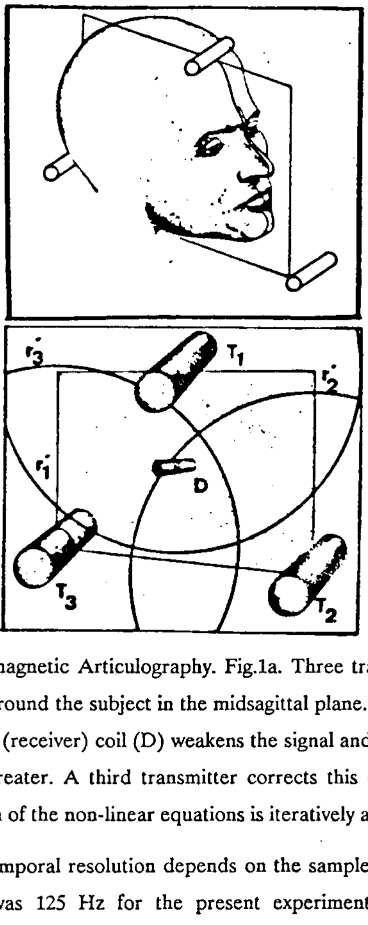


Fig.1. Electromagnetic Articulography. Fig.1a. Three transmitter coils (T) are fixed around the subject in the midsagittal plane. Fig.1b. Tilting of the detector (receiver) coil (D) weakens the signal and the radius (r) seems to be greater. A third transmitter corrects this effect and the unique solution of the non-linear equations is iteratively approximated.

The temporal resolution depends on the sample rate (up to 1 kHz), which was 125 Hz for the present experiment. The spatial resolution of the system is 0.5 mm in the major working range.

A physiological frame of reference is necessary for a reasonable orientation and localization while examining the movement trajectories of the tongue. One reference selected is the profile of the palate, which is obtained by sliding a receiver coil in the midsagittal plane along the palate (Fig.2). The overlay plot of Fig.2 demonstrates the reliability of the recordings. Another physiological reference is the occlusion plane of the subject, which is close to the resting position of the tongue. All data are presented in a coordinate system in which the x-axis is parallel to the occlusion plane of the subject.

In the present experiment two receiver coils were positioned upon the central furrow of the tongue, one about 2 cm from the tip of the tongue and the other at the dorsum of the tongue. The subjects, who are native speakers of Russian, were then requested to repeat various types of syllables in Russian, including: 1). non-palatalized consonant plus vowel <CV>, 2). palatalized consonant plus vowel <C'V> (the palatality is indicated by an "apostrophe"), 3). syllable with [j]-insertion between consonant and vowel <C(')jV>.

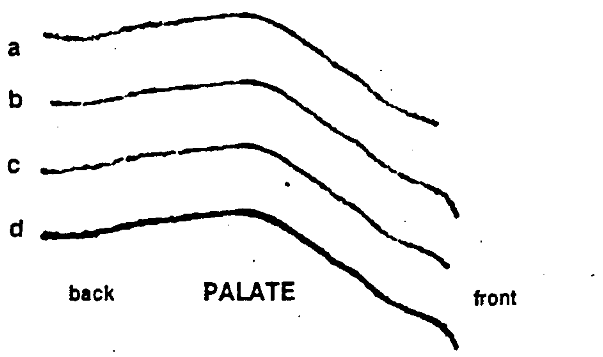


Fig.2. Three consecutive recordings of the midsagittal profile of the palate of a subject facing to the right. Fig.2a-c: Single plots. Fig.2d: Overlay plot.

RESULTS

The Russian vowels <i> and <y> deserve special attention with respect to palatalization in Russian: the consonant which precedes the vowel <i> is always palatalized, while the consonant which precedes the vowel <y> can never be palatalized. It is disputed whether they should be treated as allophones. Fig.3 compares the tongue positions of these two vowels.

For articulating the vowel <i>, the forward and upward movement occurs mainly at the front of the tongue, whereas for the vowel <y> the backward and upward movement occurs mainly at the back of the tongue. In order to determine the precise tongue position at a certain point in time, acoustic signals and movement signals, which are recorded synchronously during the experiment, are compared. For both vowels, the baseline of the tongue position shows the resting position of the tongue, which is approximately in the same plane as the occlusion plane, and is therefore parallel to the x-axis of the coordinate system. As a rule, both vowels are pronounced when the tongue reaches its highest position. The tongue position of <i> differs from that of <y> in that the tongue as a whole lies further forward. This is in accordance with the phonetic description that [i] is a high front vowel and [y] is a somewhat high and relatively back vowel. It is noticeable that the distance between the two receiver coils changes not only from vowel to vowel but also from time to time during speech movement.

In traditional static phonetics, the tongue tends to be simplified as a rigid mass, which moves at the same time in the same direction. In fact, the tongue is a heterogeneous mass so that different parts can move at the same time with different amplitudes and in different (or even opposite) directions.

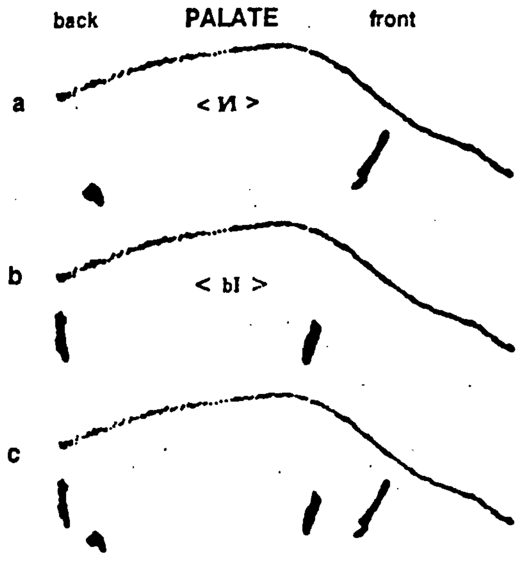


Fig.3. Tongue Positions of Russian Vowels. Fig.3a: <i>. Fig.3b: <y>. Fig.3c: <i> and <y>.

Since palatalization is closely related to high front vowels, it is generally treated as a regressive assimilation through high front "vowels", especially through the vowel [i] or the glide [j]. But the formulation of such a rule of palatalization is opaque in Russian, because, in Russian, palatalized consonants do not occur exclusively before high front vowels, and, furthermore, non-palatalized consonants also occur before high front vowels. Neeld (1973) suggested an addition of rules to solve the problem of opacity [8]. This means - in the case of Russian - an insertion of the glide [j] must be postulated. Yet the Russian phonology requires a fine distinction between palatalization and [j]-insertion, e.g.: <s'em'i> (gen. sig. of "seven") and <s'em(')ji> (gen. sig. of "family") are a minimal pair. This dilemma of static phonology can be solved by investigating the dynamic aspects of palatalization with EMA. Fig.4 and Fig.5 demonstrate the distictions between palatalization and [j]-insertion.

In the case of the palatal fricatives, the front of the tongue moves to a greater extent in comparison with the back of the tongue. The articulatory movements of the front of the tongue for the non-palatalized *palatal fricative* <Sa> are the parts of the trajectories, in which movement from the position for the palatal fricative <S> at the top directly down to the position for the vowel <a> occurs. The curved trajectories forward and upward are the preparatory movements of the tongue for pronouncing the target syllable. During the silent period, the tongue first returns to its resting position and then the front of the tongue moves backward and upward in order to reassume the initial position of the palatal fricative <S>. The trajectories for the palatalized <S'a> are loops which differ from those of the non-palatalized <Sa> in that the initial position for the palatalized consonant <S'> already lies more at the front at the very beginning of the articulatory movements.

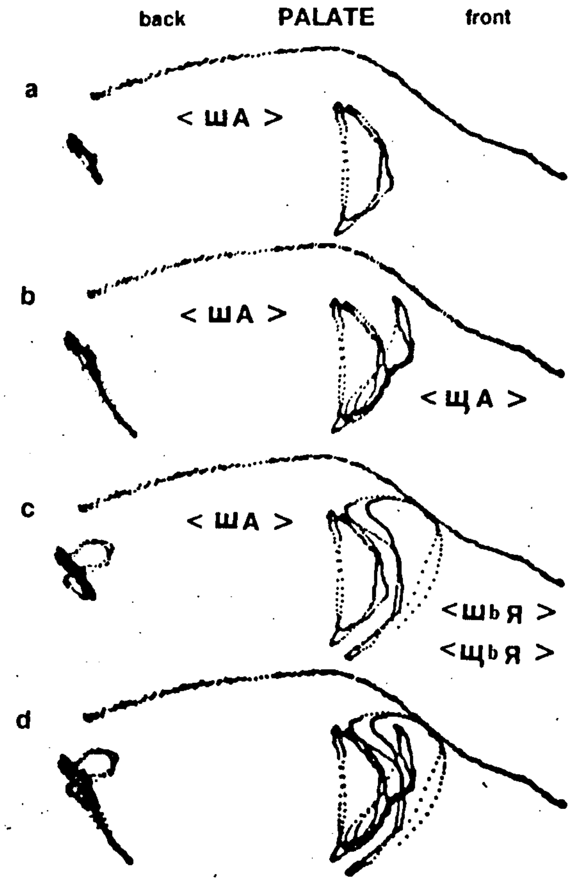


Fig.4. Palatalization of the palatal fricative <S>. Fig.4a. Non-palatalized <Sa>. Fig.4b. Non-palatalized <Sa> vs. palatalized <S'a>. Fig.4c. Non-palatalized <Sa> vs. <S(')ja> (with [j]-insertion). Fig.4d. <Sa>, <S'a> and <S(')ja>.

Although <Sja> and <S'ja> are written differently in cyrillic orthography, acoustic and articulatory investigations do not show any difference between them. Phonologically, the palatalized consonant and the corresponding non-palatalized consonant are neutralized in the position before an inserted [j]. The articulatory movement for <S(')ja> is no longer a nearly "straight line" but becomes a very bent curve. This indicates that the syllable contains three segments instead of two. Before the front of the tongue reaches its final goal, which is the area for the vowel [a], it passes an intermediate station, which is the area for the glide [j]. The initial position and the final position of the articulatory movements and even the preparatory movements of the trajectories of <Sa> and <S(')ja> coincide with each other. When we compare the trajectories of all three syllables, we find that the trajectories of the palatalized <S'a> lie just between those of the non-palatalized <Sa> and those of <S(')ja>.

All these three syllables <Sa>, <S'a> and <S(')ja> have the vowel [a] in common. The turning points at the bottom of the trajectories of the articulatory movements of these three syllables lie fairly close to each other. Yet at the same time there are still some deviations. A detailed study of the trajectories together with the acoustic signals shows that the vowel [a] is pronounced before as well as after the turning point is reached. Thus, there is not a single point but a whole area in which the vowel [a] may be produced. This means: speech production requires on the one hand precise tongue movement when sounds are to be differentiated, but allows on the other hand a certain degree of freedom when sounds are not to be differentiated.

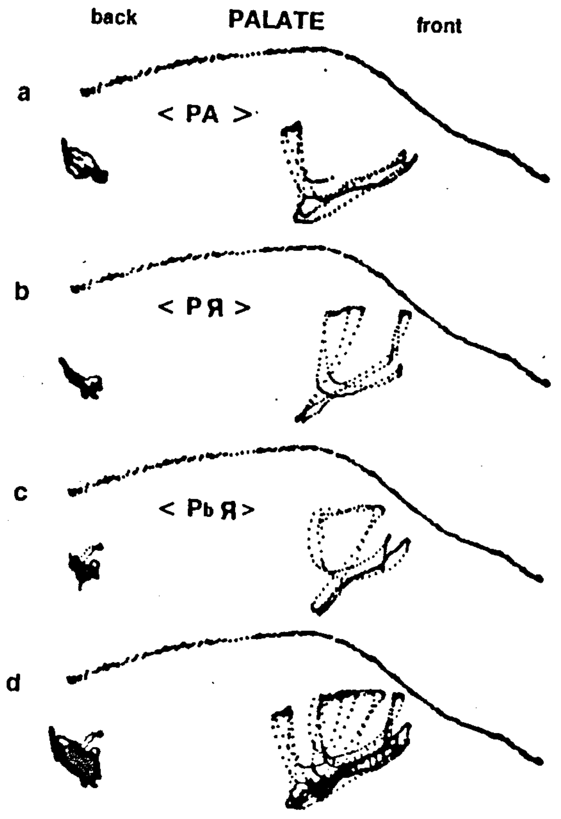


Fig. 5. Palatalization of <r>. Fig.5a. Non-palatalized <ra>. Fig.5b. Palatalized <r'a>. Fig.5c. <r(')ja> (with [j]-insertion). Fig.5d. <ra>, <r'a> and <r(')ja>.

The situation is similar when we compare <ra>, <r'a> and <r(')ja> (Fig.5). For articulating <ra>, the front of the tongue moves from its resting position first backward and then upward in order to form a constriction with the palate. Through interaction with the air flow, the tip of the tongue is set into vibration. Then it moves down to the area of the vowel <a> and then forward to its resting position. The trajectories of <r'a> differ from those of <ra> in that firstly, the constriction point of <r'a> lies (about 5 mm) more to the front and secondly, the front of the tongue moves somewhat forward in the direction to the position of the glide [j] after the tip of the tongue begins to vibrate. This forward movement is even greater for <r(')ja>. The difference is about 5 mm. On the whole, the trajectories of <r'a> lie again between those of <ra> and <r(')ja> and are more similar to those of <r(')ja> than to those of <ra>. This can be confirmed by data for many other consonants. Acoustically, palatalization and [j]-insertion resemble each other so much that most non-native speakers of Russian have difficulties distinguishing them.

DISCUSSIONS

The EMA investigation of the trajectories of tongue movements shows that there are similarities as well as differences between palatalization and [j]-insertion. It also shows that there are constants as well as variants in speech production. On the one hand, each sound requires a certain vocal tract configuration in order to be able to be distinct from other sounds in the language system. On the other hand, the various articulators are able to compensate for each other, so that each articulator has a greater degree of freedom. This speech-physiological interpretation of polymorphism supports the assumptions of generative phonology that even among distinctive phonemes in a language, there are still some overlapping of articulatory and acoustic elements.

In this sense, palatalization means the partial take-over of the acoustic and articulatory elements of the palatal glide [j] and, at the same time, differentiation from [j]-insertion. This means more exact spatial and temporal coordination between various articulators. Spatially, the trajectories of the palatalized consonant reach only the peripheral area of the glide [j], whereas those of the corresponding syllable with [j]-insertion pass through its center. The whole vocal tract is so configured for the palatalized consonant that it acquires the partial acoustic effect of a palatal fricative. At the same time, it is temporally so coordinated that the trajectories of the palatalized consonant pass the area for the glide [j] in approximately 20-30 msec less than the trajectories of the corresponding syllable with [j]-insertion. Thus, under certain circumstances the consonant and the short glide would be treated as a new consonant rather than two segments of a syllable.

In a broader sense, palatalization is a reduction of "extravagant" speech movements in the motor realization of the whole speech sequence. Since velar consonants and dental consonants require a greater extent of speech movement from the neutral position of the tongue than palatal consonants, they tend to be reduced to palatal consonants. This extended interpretation of palatalization can offer a unified explanation for palatalization at the phonetic level as well as palatalization at the historical, morphonemic level.

CONCLUSION

EMA investigation shows that palatalization can be treated in a wider framework of dynamic speech motor planning as an optimization of speech movement in the total planning of the whole speech sequence. This optimization of speech movements in various speech environments may result in a differentiation of the structure of a language.

REFERENCES

[1] Baudouin de Courtenay, J. 1894. Einiges über Palatalisierung (Palatalisation) und Entpalatalisierung (Dispalatalisation), Indogermanische Forschungen, 4, 45-52.

[2] Chomsky, Noam & Morris Halle. 1968. The Sound Pattern of English. Harper & Row: New York etc.

[3] Fant, C. Gunnar M. 1960. Acoustic Theory of Speech Production. Mouton: 's-Gravenhage.

[4] Fegert, Hermann. 1986. Die Formenbildung des Verbs im Russischen. Winter: Heidelberg.

[5] Halle, Morris. 1959. The Sound Pattern of Russian. Mouton: 's Gravenhage.

[6] Koneczna, H. & Zawadowski, W. 1956. Obrazy rentgenograficzne glosek rosyjskich. Panstwowe wydawnictwo naukowe: Warszawa.

[7] Lightner, Theodore M. 1972. Problems in the theory of phonology. Linguistic Research, Inc.: Edmonton, Champaign.

[8] Lunt, Horace G. 1981. The Progressive Palatalization of Common Slavic. Macedonian Academy of Science and Arts: Skopje.

[9] Neeld, Ronald. 1973. Remarks on palatalization. Ohio State University working papers in linguistics, 14, 39-49.

[10] Perkell, J. S. & M. H. Cohen. 1985. Design and construction of an alternating magnetic field system for transducing articulatory movements in the midsagittal plane. J. Acoust. Soc. Am., Suppl. 1,77, S99(A).

[11] Schönle, P. W., P. Wenig, J. Schrader, K. Gräbe, E. Bröckman, B. Conrad. 1983. Ein elektromagnetisches Verfahren zur simultanen Registrierung von Bewegungen im Bereich des Lippen-, Unterkiefer- und Zungensystems. Biomedizinische Technik. 28, 172-9.

[12] Schönle, P. W., K. Gräbe, P. Wenig, J. Höhne, J. Schrader, B. Conrad. 1987 . Electromagnetic Articulography: Use of Alternating Magnetic Fields for Tracking Movements of Multiple Points Inside and Outside the Vocal Tract. Brain and Language (in press).

* *Supported by Deutscher Akademischer Austauschdienst and Der Bundesmininster für Forschung und Technologie.*

A BIBLIOGRAPHY OF X-RAY STUDIES OF SPEECH

SARAH N. DART

UCLA Phonetics Laboratory, Dept. of Linguistics, Los Angeles, CA 90024-1543 U.S.A.

ABSTRACT

This paper reports the compilation of a bibliography of all studies of speech which include some x-ray data. The bibliography has been entered into a data base program for implementation on the Apple Macintosh computer and is currently being used by the UCLA Phonetics Lab Group.

Over the years, we at the UCLA Phonetics Laboratory have been compiling a bibliography of speech studies which contain some x-ray data. This has primarily been to gather a large data base of x-ray tracings and photographs for use in our research. During the past two years we have expanded this bibliography tremendously and entered it into the Microsoft File database program for implementation on the Apple Macintosh computer. This enables us to search entries with certain specific characteristics, such as language, author, or a certain segment of interest.

As a point of departure we took the existing bibliographies of Macmillan and Keleman (1952) [1] and Simon (1961 [2] and 1967 [3]) and reviewed each entry that we could locate, putting it into our database format. In addition, we have searched and reviewed many more sources not listed in those previous bibliographies and are still adding to the collection. Presently our bibliography consists of over 335 entries from 270 different sources (sources involving more than one language are listed in separate entries for each to facilitate searching).

FORMAT

We have organized each entry in our data base into ten "fields" according to the format shown below. These fields are: 1) author 2) year of publication 3) bibliographical reference 4) language involved 5) type of x-rays (i.e., still or cine-x-ray and if the latter, the frame speed) 6) segments covered (in the IPAPlus phonetic font developed at UCLA) 7) number of speakers filmed 8) location in our laboratory of the full publication 9) other data provided in addition to x-rays 10) a short abstract giving more specific information as to the type of data provided and the usefulness thereof, but not intended to be a summary of the author's claims or intent.

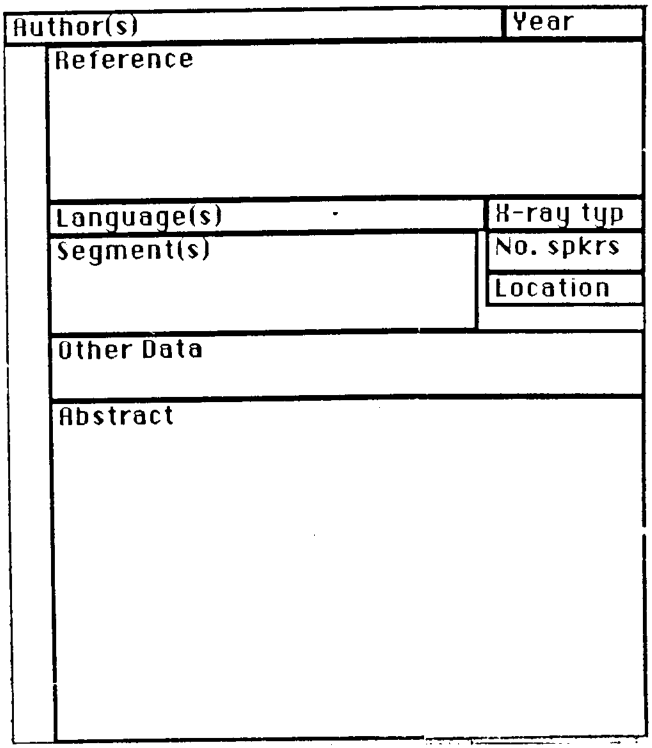


Figure 1. Blank format for each entry in the database.

A sample entry is given below to clarify the format.

Charbonneau, R. 1970

Le phonème /ɛ̃/ en français canadien. In B. Hala, M. Romportl and P. Janota (eds.), Proceedings of the Sixth International Congress of Phonetic Sciences (Prague 1967), pp.253-264. Prague: Academia.

French (Cdn.)	36 fps
ɛ̃,p,t,k,f,s	2
	X

spectrograms

Describes in detail the realization of [ɛ̃] in Canadian French. Two speakers were filmed at 36 frames/sec saying phrases consisting of 4 syllables, the last one containing [ɛ̃]. 33 composite tracings are given, showing successive frames of the syllables [pɛ̃, kɛ̃, fɛ̃, sɛ̃, pɛ̃:t, tɛ̃:t]. Spectrograms are also given of the same phrases and of the corresponding oral vowels.

Figure 2. Sample database entry.

APPLICATIONS

Each of these "fields" can be searched independently. Thus, for example, one can search for all entries from a particular language, or all those containing palatograms as well as x-rays, or those involving a particular segment. We have found this to be a useful tool in our research for easily locating articulatory data to compare segments or languages and check hypotheses. As an example, one of the laboratory members, Dr. Patricia Keating, was able to quickly perform a comparison of the differences between fronted velar consonants and true palatals by comparing x-rays from several different languages brought together for her by the x-ray bibliography database. Without this easy location of the relevant sources and the immediate knowledge of whether there even existed an appropriate body of data to examine this question, this study would have been tedious and time-consuming to the point of perhaps precluding the investigation altogether. With the exception of the location field, which refers only to our laboratory here at UCLA, this bibliography can also be useful to other phoneticians, either as a simple printout for reference or as a computer database program. We have no doubt that many participants at this congress know of sources of x-ray data that we are not aware of as yet. We look forward to widening our database from the contributions and suggestions of the other participants.

REFERENCES

[1] Macmillan, A.S. and G. Keleman (1952) "Radiography of the supraglottal speech organs: a survey". A.M.A. Archives of Otolaryngology 55:671-688.

[2] Simon, P. (1961) "Films radiologiques des articulations et les aspects génétiques des sons du langage". Orbis 10:47-68.

[3] Simon, P. (1967) Les consonnes françaises: mouvements et positions articulatoires à la lumière de la radiocinématographie. Bibliothèque Française et Romane, Série A, No. 14. Paris: Klincksieck.

ETUDE AERODYNAMIQUE DU SOUFFLE PHONATOIRE UTILISE DANS LA LECTURE D'UN TEXTE EN FRANCAIS

Bernard TESTON et **Denis AUTESSERRE**

Institut de Phonétique de l'Université de Provence 1
U.A. 261 du CNRS, Aix-en-Provence, France

RESUME

La respiration vitale comporte une alternance de phases d'inspiration et d'expiration qui se produisent avec des durées et des amplitudes bien connues.
Comment la respiration est-elle modifiée chez un sujet soumis à une tâche de lecture ?
Les résultats obtenus montrent que la répartition des pauses respiratoires est en grande partie guidée par la ponctuation du texte.
Quelles sont les nouvelles relations qui s'établissent entre les durées et les volumes d'air des phases successives d'inspiration et d'expiration ?
Nos résultats démontrent plutôt un contrôle souple qu'atteste une grande plasticité d'adaptation du système respiratoire aux contraintes d'organisation linguistique de l'énoncé.

INTRODUCTION

L'étude instrumentale des phénomènes aérodynamiques mis en jeu lors de la respiration et modifiés lors de la production de la parole, s'inscrit en France dans une longue tradition, plus que centenaire, jalonnée par les travaux d'Etienne MAREY, de l'abbé ROUSSELOT, de Marguerite DURAND et de Georges STRAKA. La présence, à la Faculté d'Aix, d'un kymographe acquis par Georges LOTE, a permis, dès la création d'un enseignement de phonétique expérimentale par Georges FAURE et Mario ROSSI, d'initier plusieurs générations à la recherche en phonétique physiologique. Dans un passé plus récent, la mise au point d'appareils de plus en plus perfectionnés nous a conduit à entreprendre de nouveaux travaux, avec une orientation plus quantitative, mais concernant toujours les modifications à court terme des phénomènes aérodynamiques : variations des débits d'air buccal et nasal, pression intra-orale (1).
La réalisation, toute récente, à l'Institut de Phonétique d'Aix, du dernier pneumotachographe appelé "Aérophonomètre III" élargit notre champ d'investigation aux phénomènes aérodynamiques de plus grande extension temporelle, plus directement en relation avec la ventilation pulmonaire : débits et volumes d'air inspirés et expirés lors de la respiration en phonation, durées relatives des prises d'air et des groupes de souffle (2).

Avant même d'entreprendre une analyse détaillée de la parole spontanée, but essentiel de ce programme de recherches, il nous a paru plus prudent, dans un premier temps, d'éprouver les possibilités de l'appareillage nouveau, en partant de l'étude de plusieurs lectures à haute voix d'un même texte : les marques de ponctuation y constituent autant de jalons susceptibles d'orienter le nombre et la répartition des prises de souffle. Nous aurons, alors, à répondre à trois questions essentielles :

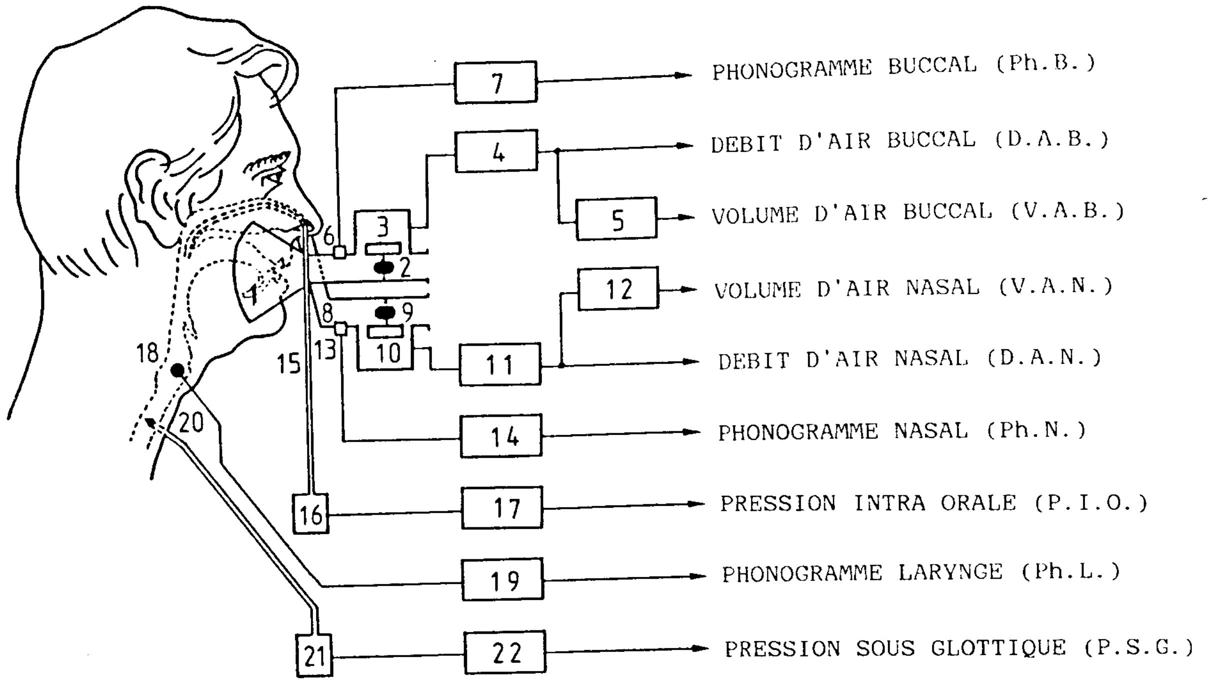
1. Quelles sont, par rapport à une respiration calme, les modifications de durée, de débit et de volume apportées aux phases successives d'inspiration et d'expiration, pendant la lecture d'un texte suivi ?
2. Cette réorganisation de la ventilation pulmonaire est-elle dépendante, et jusqu'à quel point, de l'organisation linguistique et du contenu du texte ?
3. Par là-même, et en dépit de la variabilité interindividuelle, est-il possible de prévoir les besoins en souffle nécessités par la lecture de ce texte ?

Nous n'avons pas la prétention de fournir des réponses définitives à ces questions : nous apportons plus simplement des résultats, ayant trait au français, et susceptibles de s'ajouter à ceux déjà obtenus pour d'autres langues (3) et (4). Ainsi, nous contribuerons à améliorer la connaissance des processus de production du langage articulé, dont les phénomènes aérodynamiques constituent le point de départ obligé : au commencement était le souffle !

PROCEDURE EXPERIMENTALE
CHOIX DES CRITERES D'INTERPRETATION

1. - Appareillage et paramètres enregistrés

L'aérophonomètre III (Fig. 1) se distingue des autres pneumotachographes utilisés dans les études de la respiration par sa très faible constante de temps : ceci a pour effet de fournir une bonne définition des variations à court terme des débits et des volumes d'air pendant la phonation.
De plus, et contrairement à la procédure habituellement suivie, les signaux aérodynamiques buccaux et nasaux ont été recueillis séparément à la sortie des orifices correspondants (Fig. 1). Ils sont donc représentés sur des lignes différentes lors de leur enregistrement oscillographique (Fig. 2). Ainsi, on recueille le débit d'air buccal (DAB) - ligne 1 -


PHONOGRAMME BUCCAL (Ph.B.)
DEBIT D'AIR BUCCAL (D.A.B.)
VOLUME D'AIR BUCCAL (V.A.B.)
VOLUME D'AIR NASAL (V.A.N.)
DEBIT D'AIR NASAL (D.A.N.)
PHONOGRAMME NASAL (Ph.N.)
PRESSION INTRA ORALE (P.I.O.)
PHONOGRAMME LARYNGE (Ph.L.)
PRESSION SOUS GLOTTIQUE (P.S.G.)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

1. Embouchure buccale. - 2. Pneumotachographe buccal. - 3. Capteur de pression buccale (+- 2 mB). - 4. Conditionneur du signal du débit d'air buccal (DAB). - 5. Intégrateur du volume d'air buccal (VAB). - 6. Microphone buccal. - 7. Conditionneur du signal du phonogramme buccal. - 8. Conduit de prélèvement nasal. - 9. Pneumotachographe nasal. - 10. Capteur de pression nasal (+- 2 mB). - 11. Conditionneur du signal du débit d'air nasal (DAN). - 12. Intégrateur du volume d'air nasal (VAN). - 13. Microphone nasal. - 14. Conditionneur du signal du microphone nasal. - 15. Sonde nasale de prélèvement de la pression intra-orale (PIO). - 16. Capteur de pression intra-orale (+- 70 mB). - 17. Conditionneur du signal de pression intra-orale. - 18. Laryngophone. - 19. Conditionneur du signal du larynx. - 20. Sonde trachéale de prélèvement de la pression sous-glottique (PSG). - 21. Capteur de pression sous-glottique (+- 70 mB). - 22. Conditionneur du signal de pression sous-glottique.

Figure 1

et le volume d'air buccal (VAB) expiré (VABE) - ligne 2 - ou inspiré (VABI) - ligne 3 - à l'aide d'une embouchure buccale souple, spécialement adaptée à la morphologie faciale de chaque locuteur. D'autre part, deux embouts, introduits dans les orifices narinaires, assurent l'enregistrement du débit d'air nasal (DAN) - ligne 5 - et des volumes d'air nasal (VAN) expiré (VANE) - ligne 6 - ou inspiré (VANI) - ligne 7 -.

Les signaux acoustiques sont captés à l'aide d'un microphone placé à l'intérieur de l'embouchure buccale : le phonogramme buccal correspondant (Ph.B.) - ligne 4 - permet la délimitation temporelle des séquences phoniques (la vitesse de défilement est de 50 mm/s).

En plus de cet enregistrement oscillographique, deux enregistrements magnétiques simultanés sont conservés sur les deux pistes d'un magnétophone Revox : le "son buccal" et le "son laryngé". Ce dernier est recueilli à l'aide d'un laryngophone (en vue d'analyses ultérieures de la courbe mélodique).

2. - Sujets et corpus

Dix sujets adultes, cinq hommes et cinq femmes, dont l'âge est compris entre 25 et 45 ans, ont été enregistrés. Une fois les embouchures buccale et nasales mises en place, il leur est demandé de se relaxer au maximum puis d'effectuer plusieurs cycles de respiration calme (RC). Lorsqu'ils se sentent tout à fait détendus ils peuvent aborder la lecture à haute voix d'un passage du roman de Claude Simon "La route des Flandres" (Editions de Minuit, 1960, p. 63). Ils terminent l'épreuve en revenant graduellement à leur respiration calme. L'extrait choisi comprend deux longues phrases séparées par un point : "Ils regardèrent le cheval ... retroussées. Il n'y avait que l'œil ... humide". A l'intérieur de ces deux phrases le lecteur est en partie guidé par la présence de deux points après "écumé" et de virgules à l'intérieur des phrases principales. Ces virgules individualisent des membres de phrase encore trop longs et le lecteur sera conduit à ménager des pauses et des prises de souffle supplémentaires, en dehors de ces marques de ponctuation. Ceci va introduire une assez grande diversité dans les lectures (cf. plus bas : résultats), chaque sujet ayant à répéter le texte cinq fois et toujours dans les mêmes conditions, précédé et suivi de cycles de respiration calme.

Les sujets sont convoqués, à quelques jours d'intervalle, pour réaliser plusieurs respirations profondes (RP) d'abord à leur convenance, puis à la demande et, successivement : buccale, nasale, nasobuccale (inspiration par le nez, expiration par la bouche).

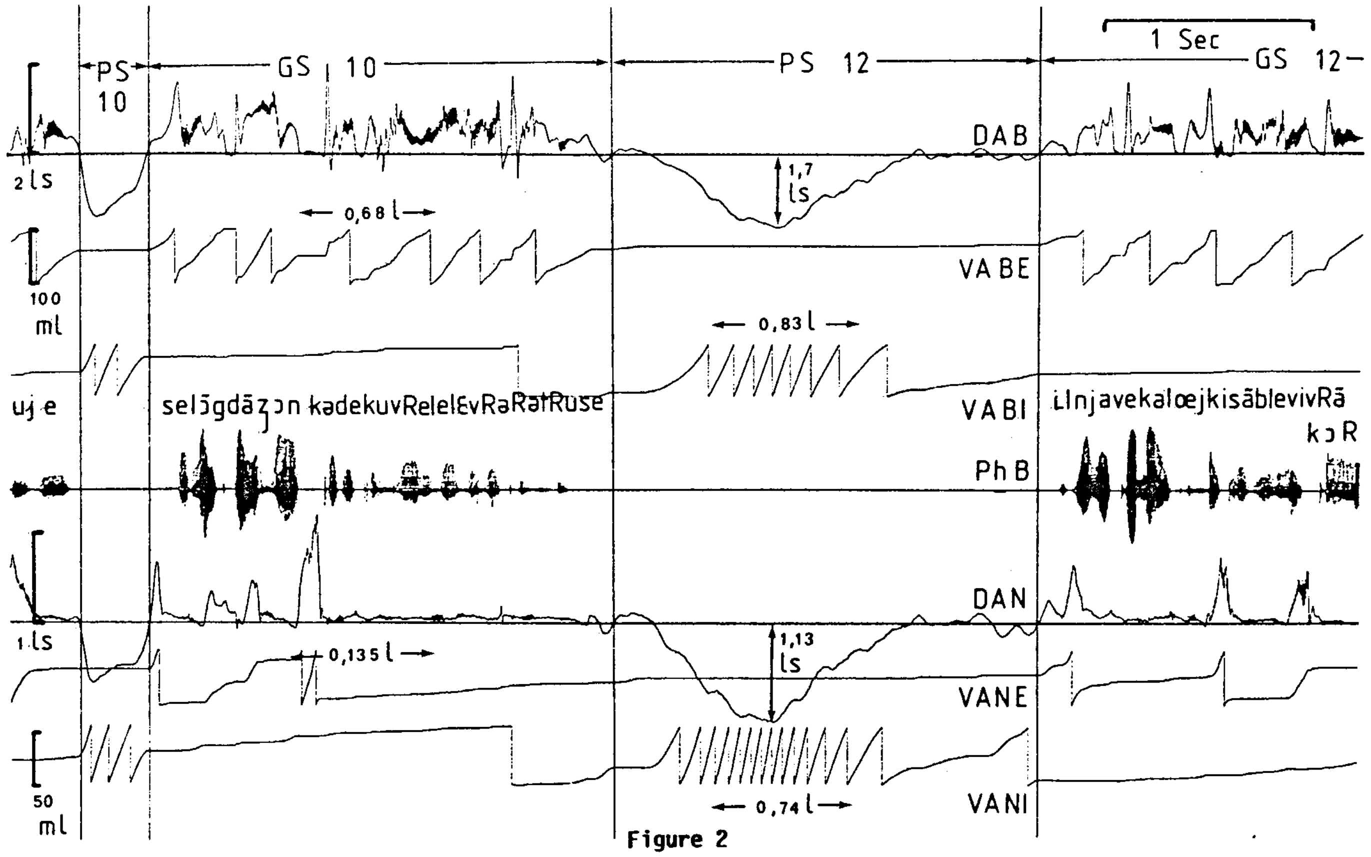


Figure 2

Enfin, une dernière épreuve consiste à prendre une inspiration forcée puis à émettre la voyelle /a/ tenue à intensité et hauteur constantes, pendant toute la phase d'expiration. L'expérience est répétée dix fois, seuls les meilleurs résultats sont pris en compte. Ils permettent de déterminer la capacité vitale en phonation (CVP) et le temps maximum de phonation (TMP).

3. - Dépouillement des données et mesures

3.1. - Délimitation temporelle :

La délimitation des phases alternées d'inspiration et d'expiration, lors de la respiration vitale (calme et forcée) ou en phonation, est effectuée à partir des lignes de base des tracés de débit d'air buccal et nasal (DAB et DAN). Ceci présuppose un bon repère des zéro correspondants. Lorsque l'expérience en phonation se poursuit plus longuement (lecture d'un texte ou parole spontanée), il est nécessaire de contrôler le zéro du DAB en se référant à de nouveaux repères tels que la partie finale de la tenue des consonnes occlusives non voisées (/p/ et /t/), de préférence en initiale de syllabe accentuée de type CV où V est une voyelle ouverte. De même, le zéro du DAN, plus difficile encore à stabiliser sera ajusté de proche en proche sur le tracé correspondant, durant la réalisation de séquences orales de type CV comportant des consonnes non voisées et des voyelles fermées telles /i/. Le début de toute phase d'inspiration, ou prise de souffle, est déterminé par la première inflection des courbes de DAB ou DAN lorsqu'elles ne sont pas simultanées. La fin de la prise de souffle est définie par le dernier retour à zéro du DAB ou du DAN. Les phases d'apnée intercalées entre inspiration et expiration, sont prises en compte avec la phase d'expiration qui suit (puisqu'elles en influencent le débit).

De nombreux problèmes de délimitations des groupes de souffle (GS) peuvent surgir lors de la respiration associée à la phonation. Il convient de distinguer soigneusement à partir du signal acoustique (Ph.B) le temps de phonation et le volume d'air exhalé correspondant, ainsi que les durées et les amplitudes des phases d'expiration. Ceci revient à isoler les segments silencieux à partir du signal acoustique aux bornes des groupes de souffle, et à repérer les pauses silencieuses internes, opération délicate en présence de consonnes ou d'agrégats consonantiques non voisés.

3.2. - Mesure des paramètres temporels :

Nous procédons à des mesures manuelles opérées sur des segments précédemment isolés :

- Durées totales des phases d'expiration (TE) et d'inspiration (TI).
- Durées partielles des temps de phonation (TEP) et des pauses silencieuses (S).
- Durées qui s'écoulent entre le début de la phase d'inspiration et le moment d'inflexion maximale du DABI et du DANI, (TIMB) (TIMN).
- Durées totales des voyelles /a/ tenues : temps maximum de phonation (TMP).

3.3. - Mesure des paramètres aérodynamiques :

Deux séries de mesures sont effectuées manuellement. Celle des débits en litres/sec et celles des volumes en millilitres BTPS ("Body Temperature and Pressure Saturated" condition corporelle de température et de pression pour un gaz saturé en vapeur d'eau). Ceci pour tenir compte des facteurs de correction thermodynamique entre l'air inspiré et expiré (5).

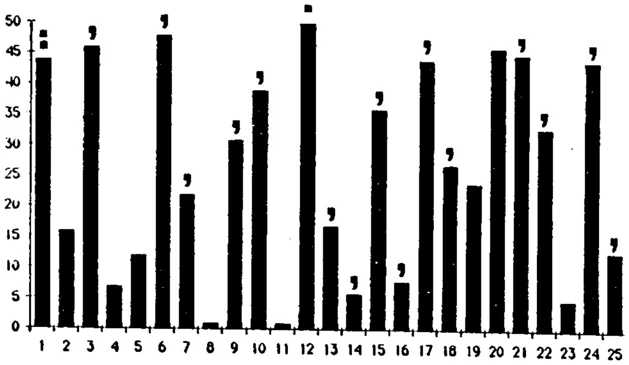
Les débits sont :
- le débit d'air buccal inspiré (DABI) et expiré (DABE).
- le débit d'air nasal inspiré (DANI) et expiré (DANE).

Les volumes sont :
- en fin de phase d'inspiration; volume d'air total inspiré (VATI) somme du volume d'air buccal inspiré (VABI) et volume d'air nasal inspiré (VANI).
- en fin de phase d'expiration; volume total expiré (VATE) somme du volume d'air buccal expiré (VABE) et du volume d'air nasal expiré (VANE).
- durant la phonation; volume total d'air expiré (VATEP) somme de VABEP et de VANEP.

ANALYSE DES RESULTATS

La répartition des pauses avec prises de souffle pour les 5 réalisations des 10 sujets se distribue comme suit :

"Ils regardèrent le cheval toujours étendu sur le flanc au fond de l'écurie (1), on avait jeté une couverture dessus (2) et seuls dépassaient ses membres raides (3), son cou terriblement long (4) au bout duquel pendait la tête (5) qu'il n'avait plus la force de soulever (6), osseuse (7), trop grosse (8) avec ses méplats (9), son poil mouillé (10), ses longues dents jaunes (11) que découvraient les lèvres retroussées (12). Il n'y avait que l'œil qui semblait vivre encore (13), énorme (14), triste (15), et dedans (16), sur la surface luisante et bombée (17), ils pouvaient se voir (18), leurs silhouettes déformées comme des parenthèses (19) se détachant sur le fond clair de la porte (20) comme une sorte de brouillard légèrement bleuté (21), comme un voile (22), une taie (23) qui déjà semblait se former (24), embuer le doux regard de cyclone (25), accusateur et humide".



La répartition des prises de souffle varie d'un sujet à l'autre mais également entre les cinq réalisations d'un même lecteur. Certaines pauses silencieuses ne deviennent jamais respiratoires, d'autres en revanche, se transforment en prises de souffle lors d'une nouvelle lecture. Les prises de souffle les plus fréquentes sont synchrones avec les marques de ponctuation. Lorsque celles-ci font défaut, les prises de souffle sont organisées sur des frontières syntaxiques. Il existe une hiérarchisation des prises de souffle, en fonction de leur durée, de leur volume, et de leur débit (rapport fort débit / durée brève). Les groupes de souffle suivis d'une importante prise d'air comportent à leurs frontières des phases silencieuses d'expiration buccale et nasale.

L'étude des données aérodynamiques ne peut être menée qu'en comparaison avec la respiration calme. Chez tous les sujets elle fait apparaître une durée plus longue à l'expiration qu'à l'inspiration, ce qui caractérise une expiration freinée (5). La lecture fait apparaître les différences suivantes :
- La durée de l'inspiration se raccourcit dans un rapport de 1/2 à 1/20 de la RC.
- Le volume d'air inspiré VATI varie dans un rapport de 1 à 30, sa moyenne étant de la moitié de celui de la RC.
- La répartition VABI, VANI est dans un rapport moyen de 4 fortement variable d'un sujet à un autre.
- Les valeurs les plus fortes de débit sont corrélées avec la brièveté des prises de souffle.
- Toutes ces mesures laissent apparaître une très grande plasticité des volumes et durées des prises de souffle en relation avec les groupes de souffle des différentes réalisations, relations atténuées sur les durées restreintes. Nous nous proposons d'approfondir tout ceci dans un travail futur.

En conclusion, nous pouvons affirmer que les échanges respiratoires en cours de lecture sont fortement corrélés par la ponctuation et les marqueurs syntaxiques. En revanche on remarque une grande variabilité d'amplitude et de durée des prises de souffle, tant entre différents sujets, qu'entre les différentes réalisations d'un même sujet. Tout semble se passer comme si, le tempo étant fixé, les coordinations pneumo-phoniques se réalisent d'une manière très souple au gré de l'"humeur" de l'individu.

(1) TESTON, B. et AUTESSERRE, D., "Réalisation d'une unité d'analyse polyphonométrique", **CLOS**, 5-6, Hommage à Georges Mounin, 415-437, 1975.

(2) TESTON, B., "A system for the analysis of the aerodynamic parameters of speech : The Polyphonometer model III", **Abstract of the 10th International Congress of Phonetic Sciences**, Foris, Dordrecht, Holland, 457, 1983.

(3) HORII, Y. and COOKE, P., "Some airflow volume and duration characteristics of oral reading", **JSHR**, 21, 470-481, 1978.

(4) ANTHONY, J., "Breathing and speaking. The modifications of respiration for speech". **Ph.D. Thesis**, University of Edimburg, 1982, 386 p.

(5) DEJOURS, P., **Physiologie de la respiration**, Flammarion, Paris, 3e éd. 1982, 315 p.

MEASUREMENT OF THE GLOTTAL IMPEDANCE WITH A MECHANICAL MODEL

HANS WERNER STRUBE STEFAN RÖSLER

Drittes Physikalisches Institut, Universität Göttingen,
Bürgerstr. 42-44, D-3400 Göttingen, Fed. Rep. of Germany

ABSTRACT

The glottal impedance is measured at acoustic frequencies, using a mechanical model with adjustable slit width and air flow. The glottis is inserted in a measuring tube with subglottal absorber, supraglottally excited by periodic wideband pulses. The complex reflectance of the glottis as function of frequency is directly computed from the incident and reflected waves, which are separated by a two-microphone directional coupler. The measured curves are compared to theory and are expressed as functions of frequency, slit width, and air flow.

INTRODUCTION

The knowledge of the glottal impedance is essential for the understanding of the source-tract coupling, e.g., the variation of formant frequencies and damping during the glottal cycle, and of the oscillation mechanism itself. The resistive part of the impedance consists of a linear, viscous component R_V and a nonlinear, flow-dependent component R_k due to kinetic effects (turbulence, beam formation, etc.). These components were measured for DC flow by the pressure drop across a glottal model [1]. For nonstationary flow, the air mass causes an additional, reactive part of the impedance, so that the electrical analogue (pressure = voltage, volume velocity = current) is an RL series circuit. This form was also used in a simulated self-oscillating glottal model [2].

However, it can not be theoretically expected that the values for R measured at DC still hold at acoustic frequencies, since the viscous boundary layer and the turbulence formation are frequency-dependent. Further, the inductance should be somewhat larger because of the approximately radial flow close to the glottal slit and also slightly frequency-dependent (see below), and turbulence effects on the inductance are unknown. As the theoretical treatment of all these effects, including nonzero DC flow and turbulence, is highly difficult, an experimental determination of the impedance as function of frequency, air flow, and glottal slit width appears desirable. Such measurements have previously been performed by means of the resonances of a tube attached to a glottal model [3]. Our approach is more direct, immediately yielding the complex reflectance as function of frequency.

THEORY

The impedance measured by us is a differential (AC) impedance. As the acoustic amplitudes are small, all terms of the Navier-Stokes equations nonlinear in AC quantities are neglected. Especially, if the total kinetic part of the pressure drop is KU^2 (U = instantaneous total volume velocity), the kinetic part of the AC pressure drop is $2KU_{DC}U_{AC}$, so that $R_k = 2KU_{DC}$ (vanishing for no DC flow!). According to [1], $K = 0.44\rho/A^2$ (ρ, A see below). At higher frequencies possibly this might not hold.

The linear (viscosity and mass) parts of the impedance, Z_{vi}, can be theoretically derived in good approximation. The glottis is assumed as a rectangular slit of length l, width w, area $A = lw$, and depth d. If $w \ll l$, and w and $d \ll$ wavelength, the impedance is

$$Z_{vi} = (i\omega\rho d/A)g/(g - \tanh g), \qquad g = (w/2)\sqrt{i\omega\rho/\eta},$$

ρ = air density, η = dynamic viscosity. For $\omega \rightarrow 0$,

$$Z_{vi} = 12d\eta l^2/A^3 + i\omega(6/5)\rho d/A,$$

which is the classic expression except for the factor 6/5 in the inductance. For $\omega \rightarrow \infty$, on the other hand, $Z_{vi} = i\omega\rho d/A$. Thus the inductance is also slightly frequency-dependent.

As the slit is contained in a partition across a tube, the impedance has to be supplemented by an end correction due to the approximately two-dimensional radial flow near the slit. This yields an additional inductance of roughly

$$L_{rad} \approx (\rho/l\alpha)\ln(D/w),$$

D ≈ tube diameter perpendicular to the slit, α = sum of opening angles of sub- and supraglottal baffles. Also some additional damping will result which we will not derive.

METHOD OF MEASUREMENT

(The same principle, suggested by M.R. Schroeder, was used earlier at this Institute for measuring the lip radiation impedance with a model head [4].)

Apparatus

A fairly realistic larynx model was formed of metal (Fig. 1). The glottis itself is a slit between two adjustable parallel plates tightly inserted in the larynx model. The slit measures are: l = 18 mm, d = 3 mm, w = 0 to 3 mm. The larynx model is extended on both sides by thick-walled uniform brass tubes of 10 mm inner diameter. Subglottally, a funnel with sound absorbing material is attached through which a DC air flow can be supplied. Supraglottally, the tube ends at a pressure-chamber loudspeaker and an air outlet with a plastic hose filled with cotton wool (Fig. 2).

The loudspeaker emits periodic wide-band pulses (68-5000 Hz), chirp-like with Schroeder phases [5] for a low peak factor. They are generated by a TMS 32010 signal processing system and D/A converted at 20 kHz sampling rate, with 2048 samples/period to facilitate FFT processing. By two ¼" condenser microphones (Brüel & Kjaer 4136) coupled to the tube some 22.5 cm "above" the glottis, the incident and reflected waves can be separated computationally and thus the complex reflectance be determined. The microphones are screwed into the tube walls without grid caps and coupled through holes of 1 mm diameter. Disturbance by the additional volume was estimated to be completely negligible. The signals are low-pass filtered at 5 kHz with 96 dB/octave and digitized at 20 kHz rate by two A/D converters in the TMS 32010 system. Sampling is period-synchronous with the excitation pulses. The blocks of 2048 sample pairs are transferred to a large laboratory computer (Gould 32/9705), where 100 periods are averaged for noise reduction and the further evaluation is performed. Channel crosstalk is less than -80 dB.

Evaluation

Let b be the distance from the centre between the microphones to the reference plane in which the reflectance is to be measured, 2a the microphone distance (Fig. 3), R = R(ω) the reflectance, and k_i, k_r the complex propagation "constants" for the incident and reflected waves. If c is the sound velocity and v the DC-flow velocity,

$$k_i = (i\omega + q\sqrt{\omega})/(c-v), \qquad k_r = (i\omega + q\sqrt{\omega})/(c+v),$$

where $q\sqrt{\omega}$ (small; $q \approx 0.8\ s^{-\frac{1}{2}}$) represents the combined viscous and heat-conduction losses. Then the microphone signals $F_1(\omega)$, $F_2(\omega)$ are proportional to $\exp(k_i(b\pm a)) + R\exp(-k_r(b\pm a))$, where +a and -a belong to F_1, F_2, respectively. Solving for R, we obtain

$$R = \exp((k_i+k_r)b)\cdot(F_2\exp(k_ia)-F_1\exp(-k_ia)) / (F_1\exp(k_ra)-F_2\exp(-k_ra)).$$

From R, the glottal impedance follows as

$$Z = Z_0(1+R)/(1-R) - Z_{SG},$$

where $Z_0 = \rho c/A_{tube}$ is the characteristic impedance of the tube and Z_{SG} the impedance of the subglottal system. The latter is measured before by replacing the larynx model with a uniform tube piece.

As the computation of R fails at the zeros of the denominator and becomes rather inexact at low frequencies, two different microphone distances 2a (24.6 and 120 mm) may be used.

Calibration

As the microphones and the connected amplifiers, filters and A/D converters are not identical for both channels and differences would cause detrimental errors, the channels must be calibrated relative to each other. For this purpose, the signals are recorded for each microphone screwed into the same fitting in the measuring tube. The complex

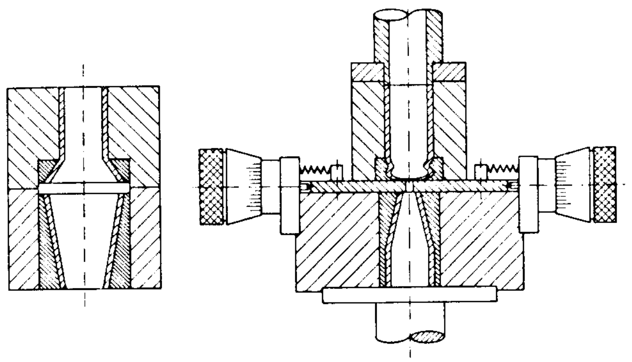


Fig. 1. Larynx model; two perpendicular sections.

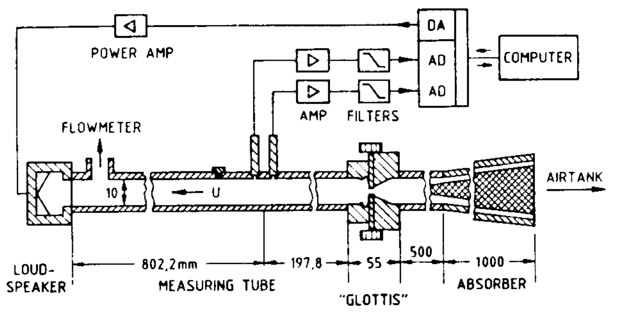


Fig. 2. Schematic of measuring apparatus.

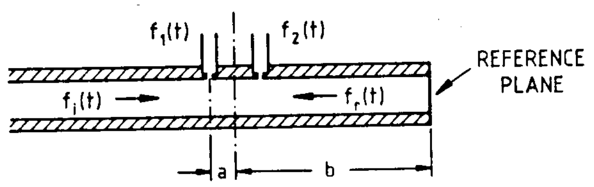


Fig. 3. Schematic of directional-coupler principle.

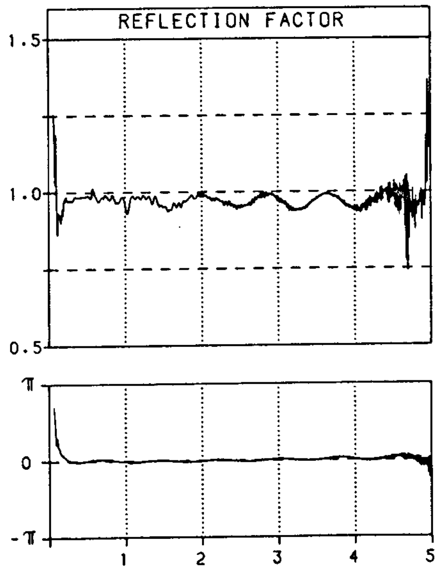


Fig. 4. Magnitude and phase of the closed-glottis reflectance.

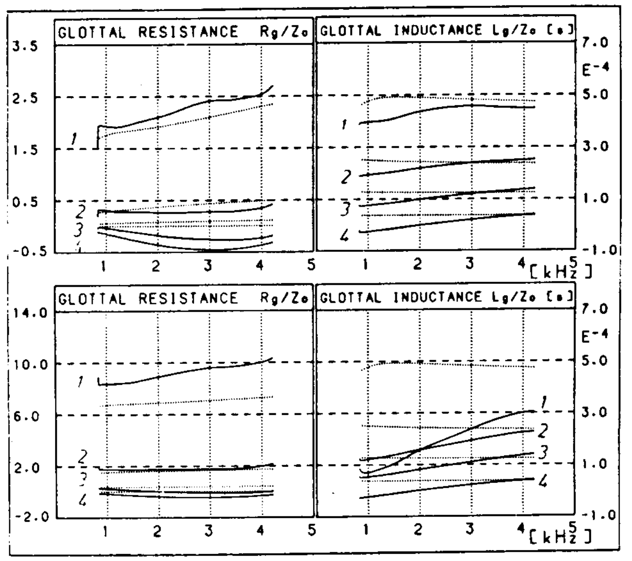


Fig. 5. Resistance and inductance for flows U = 0 (top) and 164 cm^3/s (bottom). Glottal width w = 0.2, 0.4, 0.8, 3.0 mm (1 to 4).

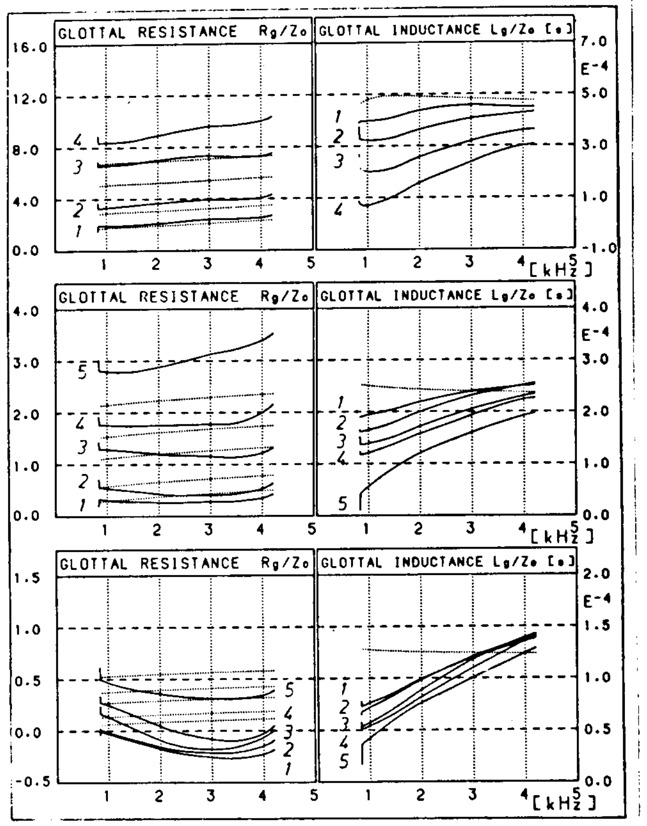


Fig. 6. Resistance and inductance for widths w = 0.2, 0.4 and 0.8 mm (top to bottom). Flow U = 0, 38, 109, 164, 245 cm^3/s (1 to 5).

quotients of the corresponding DFT values are taken as calibration factors for one microphone.

As a test, the result for completely closed glottis should yield $R(\omega) \equiv 1$, apart from some high-frequency deviations due to the nonuniformity of the tube close to the glottis. This allows to determine the exact reference distance b from the linear phase trend of R. Inexact assumptions of a and q will cause periodicities of R with frequency-period of c/2b; minimizing their amplitudes thus permits better adjustment of the a and q values.

RESULTS

All results shown here are preliminary and will hopefully have been improved at the time of the congress. So far, only one microphone distance 2a = 24.6 mm has been used.

Calibration

Fig. 4 displays magnitude and phase of the reflectance for closed glottis. The trends in the phase and the residual c/2b-periodicities show that we have not yet fully reached the required exactness; better calibration methods are under development. The average $|R|$ cannot be raised above 0.98 ($q = 0.88\ s^{-\frac{1}{2}}$) without distorting the curves.

The subglottal impedance Z_{SG} was found very close to Z_0 except at the lowest frequencies.

Measurements

The periodicities are presently smoothed out by a triangular moving average of the reflectance of length c/b in frequency. Figs. 5 and 6 show the glottal impedance (with Z_{SG} subtracted out) for various openings w and flows U. The dashed curves are the theoretical ones,

$$R_g = \text{Re}\ Z_{vi} + R_k(U), \quad L_g = (\text{Im}\ Z_{vi})/\omega + L_{rad},$$

see THEORY. For U = 0, the agreement is fairly good, except for a too low (even partly negative) resistance at large glottal openings and a too low inductance at low frequencies. The reason for these (unphysical) deviations is presently not yet clear but probably related with the calibration problems. For nonzero flow, the inductance is considerably decreased at low frequencies and the resistance is increased, especially for narrow width w where the velocity U/lw in the glottis is large. A similar effect for the inductance was also found by Laine and Karjalainen [3] around 1 kHz.

Discussion

A direct comparison of our results with [3] is not yet possible since our frequency range lies above that (≲ 1.5 kHz) considered in [3]. For useful results in the low-frequency range, we shall apply the microphone distance 2a = 120 mm and a lower sampling frequency.

The results at higher frequencies show a very strong dependence on the choice of the reference plane (distance b). Actually, as the "exact position" of the glottal impedance is somewhat arbitrary, so is the impedance itself. We define b so as to yield no linear phase trend for closed glottis, and the closeness between theoretical and measured curves seems to justify this procedure.

The flow effects on the inductance are presently not yet expressed by a theoretical or empirical formula. The relevant parameter appears to be the velocity in the glottis, U/lw, rather than the flow U. The kinetic resistance R_k at large U should be somewhat higher than according to [1]. The frequency dependence of R_k seems to be small.

As for the effect of the glottal impedance on the vocal-tract acoustics, the subglottal impedance must not be subtracted out. If the actual Z_{SG} is close to ours (our tube has roughly the diameter of the trachea), the real part for not too small openings w is entirely dominated by Z_{SG}.

REFERENCES

[1] J. van den Berg, J.T. Zantema, P. Doornenbal, jr.; J. Acoust. Soc. Am. **29**, 626-631 (1957).

[2] J.L. Flanagan, L.L. Landgraf; IEEE Trans. Audio Electroacoust. **AU-16**, 57-64 (1968).

[3] U. Laine, M. Karjalainen; Proc. ICASSP 86, 1621-1624 (1986).

[4] J. Kretschmar; Fortschritte der Akustik - DAGA'75, 429-432 (Weinheim: Physik Verlag, 1975).

[5] M.R. Schroeder; IEEE Trans. Inform. Th. IT-16, 85-89 (1970).

RESEARCH OF THE SPEECH DYNAMIC STRUCTURE

A.P.Belikov, V.D.Makhnanov, N.V.Mulyukin, K.V.Tunis

Maurice Thorez Institute of Foreign Languages

In the work dynamic properties of the speech signal are investigated. To describe speech dynamics a function is developed and calculated which integrally reflects the quality change of the speech signal. Algorithm of processing the acoustic speech signal is given and possibilities of an automatic segmentation of continuant speech are estimated.

At present linguistics and first of all phonetics, have got a social order from specialists in automatic speech recognition to study the speech signal structure. The fact of the existence of such a structure alongside with the language structure is originally set by the language and speech opposition, first well-founded by Ferdinand de Saussure. The urgency of the speech structure study may be explained by the fact that the practice of linguistic research in the first part of the 20th century did not stimulate intensive development of the problem and did not suggest any fundamental solutions of the speech segmentation problem and development of a well-based speech unit system.

Most researches consider syllable to be the minimal speech unit. In this case it is very important to avoid the mistake of using language notions in speech. From the point of linguistics the syllable is a linear combination of phonemes. Attempts to express the syllable with the help of parameters to extract its boundaries in the actual acoustic signal have not given reliable results.

In the decision of the principal task of speech segmentation psycho-physiological analysis of speech activity is often used. Two approaches are possible in that case: from the standpoint of speech production and speech perception. The approaches are not congruent between each other.

In /1/ the syllable is interpreted as an articulatory speech unit which is a realisation of a single articulatory act. As in the solution of the speech segmentation and speech recognition problems researchers first and foremost deal with the acoustic speech signal it is more reasonable to base oneself on the phychophysiological analysis of speech perception. It should be noted though that the peripheral mechanisms of perception are less studied than the effector mechanisms of articulation.

In /2/ an attempt was made to describe adequately the process of speech perception. It was suggested in the work that the speech signal should be presented as a flow of acoustic events detected in the signal by the auditory system. As an example of possible acoustic events increase or vice versa, decrease of energy in a certain part of the spectrum, the shift of the spectrum maximum in a certain direction, a short-time pulse or, vice versa, silence in the signal were pointed out. However, such a multidimensional and fuzzy description of an acoustic event cannot serve as a basis for the modelling of its automatic extruction procedure. Acoustic events are consideral real in the sense that without them it is difficult to model phonetic interpretation. At the same time they are unreal in the sense that it is not yet possible either to describe or enumerate them /2/.

A speech signal is given naturally in the acoustic form. In connection with this the following questions arise: in what way is that form organised? What shall we be guided by in the analysis and segmentation of the speech signal? The speech flow is first and foremost characterised by its changeability. Constant features in speech, in our opinion, are revealed only at the semantic level. At present it is not possible to say definitely what the substratum of constant speech features is. We consider that speech analysis should be based on different changes in the acoustic speech signal. Some researchers have studied speech signal dynamic features but the approach has not been consistent enough, as a rule.

Thus, A.A.Pirogov /3/ as far back as 1963 suggested a phonetic speech theory according to which phonetic speech units are entirely determined by the law of time spectrum alternations. It was suggested to consider typical sound combinations as typical phonetic speech units and transition between two adjacent phonemes as the principal characteristics of the signal. But even in that case the speech signal model was treated as a combination of different phonemes.

Judging by the experience phonemic analysis in automatic speech recognition does not give necessary results. Succession of phonemes or syllables which are in point of fact combinations of phonemes cannot serve an adequate description of speech as a dynamic process.

Speech activity analysis and acoustic speech signal analysis have lead to the idea that it is necessary to use dynamic speech signal features in the system speech analysis but at a different angle. The way we see it, all speech signal alterations are the realisation of speech dynamics which is inherent to speech and has its specific structure. In what way is it possible to characterise these changes? If we express speech in parameters, as it is done in most cases, all the parameters will change. In that case the mechanism of relation between separate parameters is not clear. It is not clear either which of the parameters should be the main one in the speech dynamic description. If we consider speech in a generalised way, that is as a process of communication, it is necessary to answer the following question: what makes the speech signal communicatively valuable? We can answer the question definitely enough: quality changes in the speech signal make it communicatively valuable. Considering speech as a certain movement form it is possible to assume that changes in the signal quality dynamics, constitute the base of speech dynamics. Time quality changes may be represented as a "quality function", the main dynamic characteristic of the speech signal. It should be noted that the realisation of the "quality function", i.e. a periodic change in the quality of the speech signal, needs a cyclic process which could provide the "filling" of time with single dynamic cycles. From the view point of speech production a dynamic cycle is single speech act at the automatism level for realisation of a speech element. (By a "single" speech act we understand a structurally formed complex of articulatory actions). Dynamic cycles should be regarded as peculiar technical means for the realisation of language programs.

Let us try to look at the dynamic cycle from the view point of diachrony. Supposing the role of depictive and imitation principles of forming the language at early stages of development was great the dynamic cycle then was used for the realisation of the elementary signal function. In accordance with this it could acquire the meaning of an image-bearing semantic unit. Later as a result of the language evolution considerable shifts took place in its sign system (including the semantic aspect of the language). As a result, elementary image-bearing semantic units could lose their independent meaning, where as the dynamic cycle having entered the class of automatisms, continued to improve itself at its level in the degree and reliability of the speech process automatism. Thus the developed systems of the language may be regarded as a superstructure over primary acoustic-physiological layer which developed in the process of phylogenesis and genetically consolidated itself in the form of an exclusive flexibility of the speech apparatus. In the process of ontogenesis this layer is formed under immediate influence of social aims.

A significant step on the way of the speech dynamics description is the search of a physicial correlate of the "quality function". Obviously, qualitative characteristics of the alternating speech signal are defined by the sum total of its physical components. In the procedure of the speech signal processing we chose the way of maximum integration and tried to develop and calculate a function which could, with all its generalised character, first of all reflect the changes in the quality of the speech signal. The speech signal spectrum gives practically complete information about its quality. The quality of the signal is in the first place determined by the amplitude-frequency structure of the instant spectrum. The instant spectrum of speech is a multiparametrical description each component of which is a time

function. It seems reasonable to us to use the time function of root-mean-square frequency of instant speech spectrum as a correlate of the "quality function",

$$W^*(t)=\frac{\sqrt{\int_0^\infty \omega^2 S^2(\omega,t)d\omega}}{\sqrt{\int_0^\infty S^2(\omega,t)d\omega}}$$

where the amplitude-frequency spectrum S (w,t) is regarded as a weight function. The function W*(t), integrally related with the frequency structure of the speech signal spectrum, in frequency unit shows its qualitative changes, which are conditioned by "pumpingover" the energy from certain frequency domains to others.

The culculation of the one-dimensional"quality function" may by expressed as a result of the maximum decrease in the dimension of the initial function which describes the process.

In the realisation of the device forming W*(t), a certain inconveniency is presented by the integration operations of spectrum functions in frequency. With the help Rayleigh theorem and mathematic definition of the instant spectrum we managed to pass over from frequency integration to time integration and to get a more convinient formula for W*(t):

$$W^*(t)=\frac{\sqrt{\int_{t-T}^{t}\left(\frac{df}{d\tau}\right)^2 d\tau}}{\sqrt{\int_{t-T}^{t} f^2(\tau)d\tau}}$$

where f(t) is the function of the acoustic pressure of the speech wave, T is an intergration interval in the calculation of the instant spectrum.

The use of the time dependence root-mean-square frequency of speech instant spectrum for the integral description of qualitative changes in the speech signal seems well-founded to us. It is known that devoid of relationship single parameters are characterised by a high entropy. In fact, they insert noise in the useful information if the inner structure of their relationships is not revealed. That is why striving for a more and more detailed description of the signal under research with the help of a number of single parameters often leads to the masking of the dynamic regularities.

The one-dimensional time function W*(t) can effectively characterise the general dynamics of the speech process at this function has got a number of useful properties: it is continuant and invariant in relation to the level of the speech signal and insensitive to stationary noises.

To test the speech signal dynamic structure, in general, and the segmentation of the continuous speech, in particular, a model of the system was made which realises the above described algorithm with the help of analog microschemes.

In the study of the properties of the function W*(t) complex from the segmentation point of view test phrases were used composed of the combinations of vowels and sonants. Developed on the base of the speech signal the function W*(t) consisted of repeated cycles with distinct extremums which we call dynamic cycles. In the study of the above-mentioned test phrases we got equal number of dynamic cycles and syllables. Meanwhile the synchronically registered speech signal intensity envelope had more extremums which were less explicit as compared with the W*(t) envelope. It enables us to conclude that the chosen system of speech signal operations minimises the number of extremums and makes them more explicit.

The equal number of dynamic cycles and syllables, as it was supposed, is not obligatory. It depends on the character of the speech material. In the general case differences were observed: hissing and hushing fricative sounds, as well as affricates, in test phrases formed separate extremums of the "quality function".

It should be noted that the dynamic cycle is not used instead of either the syllable, the phoneme or other units of the language system. The speech process is organised in a specific way - it cannot be treated as a language model. The major property of the dynamic cycle (unlike the units of the language system) consists in its possible quantative estimation and also in the comparison and analysis of its quantitative characteristics.

References

/I/ Н.И. Жинкин. Механизмы речи. М., 1958
/2/ Физиология речи. Восприятие речи человеком. Под ред. А.А. Чистович. Л-М., Наука, 1976
/3/ Вокодерная телефония. Методы и проблемы. Под ред. А.А. Пирогова. М.,Связь, 1974

WIGNER DISTRIBUTION - A NEW METHOD FOR HIGH-RESOLUTION TIME-FREQUENCY ANALYSIS OF SPEECH SIGNALS

W. Wokurek, G. Kubin, F. Hlawatsch

Institut für Nachrichten- und Hochfrequenztechnik, TU Wien
Gusshausstr. 25/389, A-1040 Vienna, Austria

ABSTRACT

Two methods for the time-frequency analysis of speech signals are compared: the tradionally used Spectrogram and the Smoothed Pseudo Wigner Distribution (SPWD). It is shown that the time and frequency resolutions of the Spectrogram are restricted by the uncertainty relation while SPWD allows arbitrarily high resolutions. If the analysis parameters are chosen carefully SPWD yields more accurate signal representations than the Spectrogram. This is exemplified by a "microscopic" analysis of vowels and unvoiced stop consonants.

1. INTRODUCTION

The Wigner Distribution (WD) is a method for the time-frequency analysis of signals. Along with the Spectrogram the WD is a member of a special class of bilinear, shift- invariant signal representations (Cohen class [1] p.376). Within this paper we compare a somewhat modified WD, i.e. the Smoothed Pseudo Wigner Distribution (SPWD) to the Spectrogram, first with respect to the basic features of time- and frequency resolutions (Sections 2-4). In sections 5 and 6 we compare the results of representing speech signals through SPWD and Spectrogram.

Observing the fact that SPWD enables high-resolution signal representation, we analyze short speech segments of a few pitch periods of length. Therefore it is pointless to compare the SPWD to Spectrograms of high frequency resolution (45 Hz) because they do not display the fine-structure in time that we will see in the SPWD. As a compromise between Spectrograms of high frequency resolution (which do not show the time- structure) and those of high time resolution (which smear out the formant structure etc.) we find the spectrogram of 300 Hz frequency resolution as a suitable partner for the comparison with the SPWD of vowels (see section 5). In the case of unvoiced stop consonants, equal frequency resolution of the Spectrogram and SPWD is chosen for the analysis of the whole explosion interval of several centiseconds duration because there is no significant time structure of the noise-like excitation observed (section 6).

2. DISTORTION OF TIME-FREQUENCY ANALYSIS DUE TO LIMITED RESOLUTION

The aim of a time-frequency representation of any signal is to show the stucture of the signal and not that of the analysis method. One of the basic distortions of any analysis method is its limited resolution. To study the nature of time resolution consider an impulse in the time domain as shown in Fig.1.

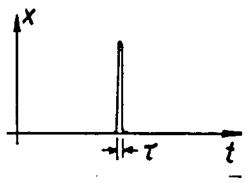


Fig.1: Impulse of length τ in the time domain.

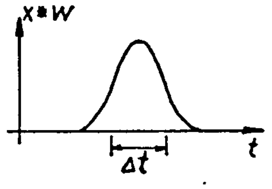


Fig.2: Representation of an impulse with a time resolution of Δt.

If we represent this impulse with a time resolution Δt ›› τ the impulse will be widened to the duration Δt (see Fig.2).

The mathematical model of this effect is the convolution of the signal x with a window function w of time width Δt:

$$[x \underset{t}{*} w](t) = \int_{\mathbb{R}} x(t-\tau)\, w(\tau)\, d\tau \tag{1}$$

A second interpretation of (1) is lowpass filtering. If the signal contains oscillations of periods less than the time resolution Δt, these components of the signal will be supressed in the representation.

A similar effect is caused by the frequency resolution Δf. All signal components will be widened by Δf in the frequency direction. On the other hand all signal changes within a frequency range of Δf will be canceled.

3. COUPLING OF THE RESOLUTIONS OF THE SPECTROGRAM

The Spectrogram is defined as the square magnitude of the Short-Time Fourier Transform (STFT). The signal is multipied by a window $w(\tau)$ that is shifted to the instant of analysis t. The Fourier Transform of this product is associated with the instant t.

$$S_x(t,f) = \left| \int_{\mathbb{R}} x(\tau)\, w(\tau-t)\, e^{-j2\pi f\tau}\, d\tau \right|^2 \tag{2}$$

Using elementary signal theory, we can recast eq. (2) in a form containing a convolution with the window w(t)

$$S_x(t,f) = \left| [e^{-j2\pi ft}\, x(t)] \underset{t}{*} w(-t) \right|^2 \tag{3}$$

or with its spectrum W(f),

$$S_x(t,f) = \left| [e^{j2\pi ft}\, X(f)] \underset{f}{*} W(f) \right|^2 \tag{4}$$

This shows us the simultaneous determination of both the time and the frequency resolution of the Spectrogram by a single window function. Like any other function, the window satisfies the uncertainty relation (5), where c is a constant that depends only on the definitions of Δt and Δf and is of the order 1.

$$\Delta t . \Delta f \geq c \tag{5}$$

The uncertainty relation (5) restricts the allowed values of the time and frequency resolutions of the Spectrogram to the region U shown in Fig. 3.

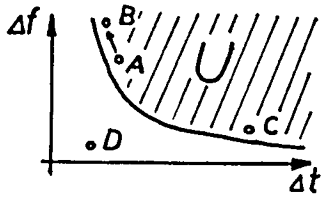


Fig.3: Restriction of the Spectrogram resolutions by the uncertainty relation

Because the product of the Spectrogram resolutions Δt.Δf cannot be less than the constant c, it is impossible to choose both resolutions arbitrarly high (i.e. Δt and Δf arbitrarily small) at the same time. This implies the necessity of trading-off between these two resolutions. If the time resolution is increased (smaller Δt), the Spectrogram must have a poorer frequency resolution (greater Δt, see Fig.3: movement from point A to B). The dual case is the choice of higher frequency resolution (Fig.3: point C), thus decreasing the time resolution.

4. INDEPENDENCE OF THE RESOLUTIONS OF THE SPWD

The Wigner Distribution (WD) of a signal x(t) is defined by (6)

$$WD\ (t,f) = \int_{\mathbb{R}} x(t+\tfrac{\tau}{2})\ x^*(t-\tfrac{\tau}{2})\ e^{-j2\pi f\tau}\ d\tau \tag{6}$$

and its features are described in [1] extensively. The WD does not show any effect of limited resolution, but in the case of fairly complex signals such as speech the result is quite unreadable owing to the occurence of *interference terms*, described in [2] (see section 5 also). Therefore we consider the SPWD of the signal which is defined as a WD with arbitrary smoothing:

$$SPWD_x = WD_x \underset{t}{*} u(t) \underset{f}{*} v(f) \tag{7}$$

Smoothing in both the time and frequency direction is performed by two independently chosen arbitrary windows u(t) and v(f), respectively. Because of the independence of the smoothing functions, the resolutions of the SPWD are not restricted by the uncertainty relation (5) (see Fig 3: point D).

Yet, from a practical point of view, the resolutions of the SPWD are restricted by the occurence of *interference terms* and depend on the structure of the signal in that way. The analysis of speech signals shows that with equal frequency resolution (e.g. 100 Hz), the SPWD allows a substantially higher time resolution than the Spectrogram (e.g. 1 ms instead of 10 ms).

An interesting insight into the relation between the Spectrogram and WD is obtained from the following equation:

$$S_x = WD_x \underset{t}{*} \underset{f}{*} WD_w \tag{8}$$

This equation proves that the Spectrogram is the WD of the signal smoothed in both directions with the WD of the Spectrogram window. In contrast to (7) the time and frequency smoothing is determined by one and the same window w(t) as we have seen already in (3) and (4) and this is why Spectrogram resolutions are bounded by the uncertainty relation (5) ([1] p.382).

5. ANALYSIS OF VOWELS BY SPWD AND SPECTROGRAM

Figure 4 shows a *contour plot* of the SPWD of three successive pitch periods extracted from the German vowel [a:] spoken by a male subject. This representation displays the following features:

(1) Quasi-periodic excitation of the vocal tract by wide-band narrow-time impulses every 10 msec. The time resolution of approx. 0.5 msec is sufficient to prove that these impulses have a time width of 1 msec or less.
(2) Exponential decay of three formants at the frequencies F1 = 0.7 kHz, F2 = 1.25 kHz, and F3 = 2.6 kHz. The frequency resolution of appprox. 100 Hz is sufficient to separate the individual formants and to measure their bandwidths during the intervals outside the excitation impulses.

(3) Besides these *signal terms* (formants, impulses), the SPWD contains *interference terms*. They are governed by a simple geometrical rule [2], i.e. they always lie half-way between two signal terms and oscillate in the direction perpendicular to the line connecting the two signal trems. These oscillations have a period in the time-frequency plane that is inverse proportional to the distance of the signal terms. The oscillatory nature of interference terms is the key to their suppression in any bilinear time-frequency representation. In SPWD, this is achieved by smoothing with the two independent window functions according to (7). The amount of smoothing must be matched to the signal structure:

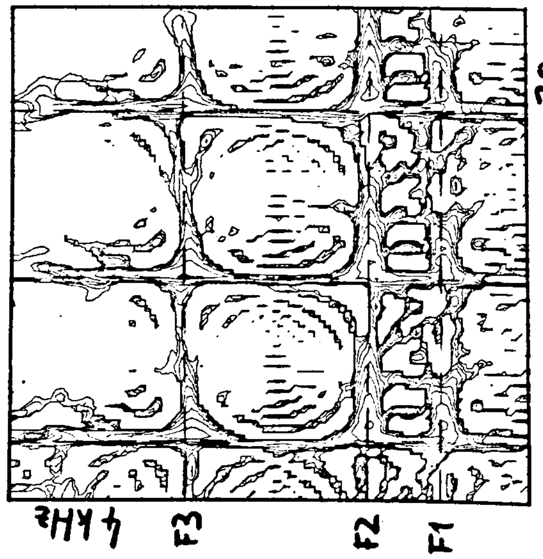


Fig. 4: SPWD [a:] from [ta:t]
Δf=100Hz, Δt=0.5ms, ΔtΔf=0.05

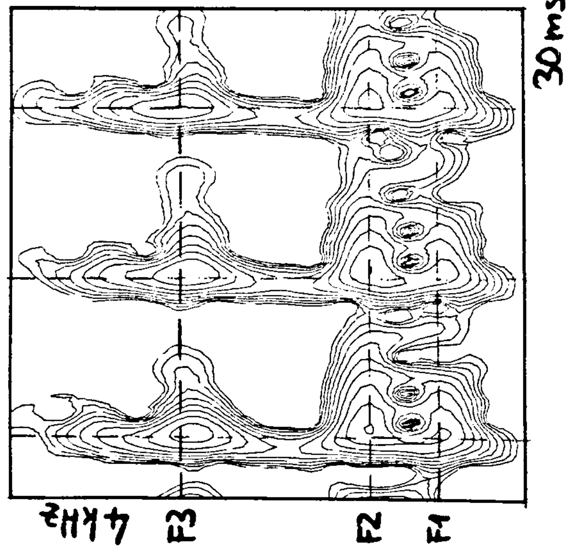


Fig. 5: Spectrogram [a:] from [ta:t]
Δf=300Hz, Δt=2ms, ΔtΔf=0.6

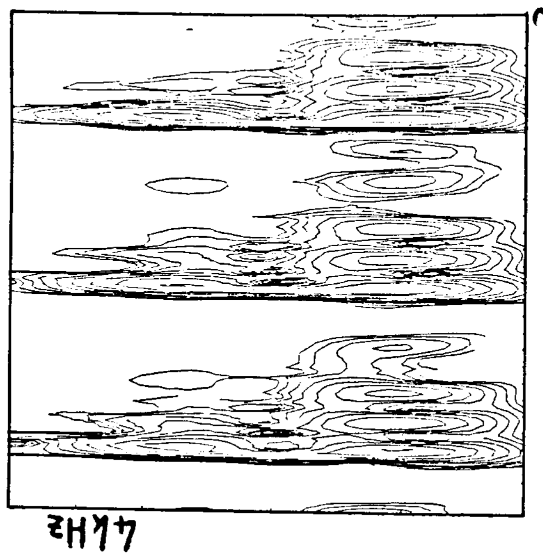


Fig. 6: Spectrogram [a:] from [ta:t]
Δf=600Hz, Δt=1ms, ΔtΔf=0.6

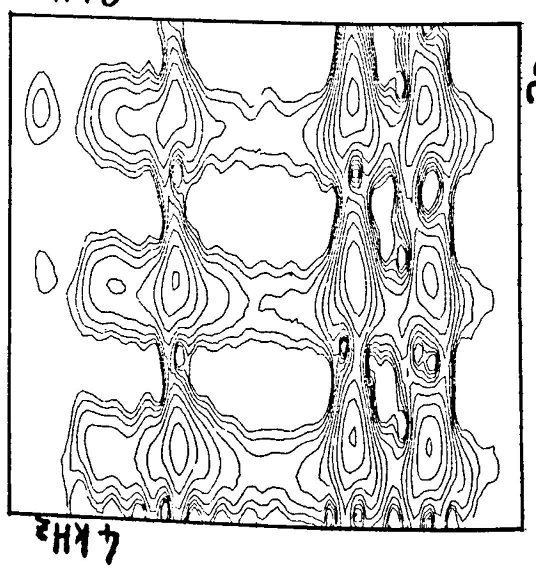


Fig. 7: Spectrogram [a:] from [ta:t]
Δf=100Hz, Δt=6ms, ΔtΔf=0.6

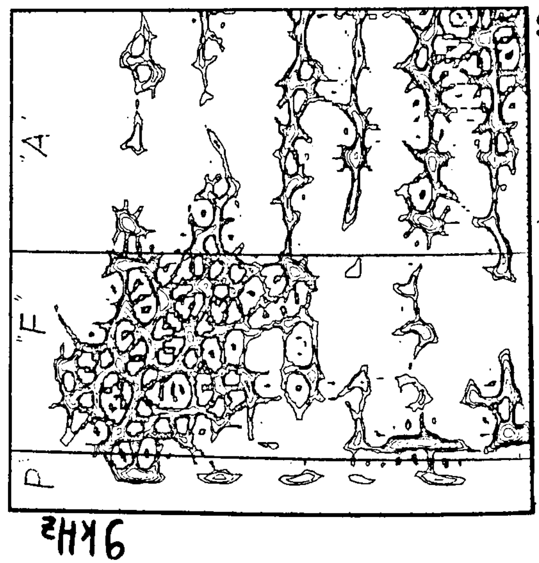


Fig. 8: SPWD of explosion interval in first [t] from [ta:t]
Δf=100Hz, Δt=1ms, ΔtΔf=0.1

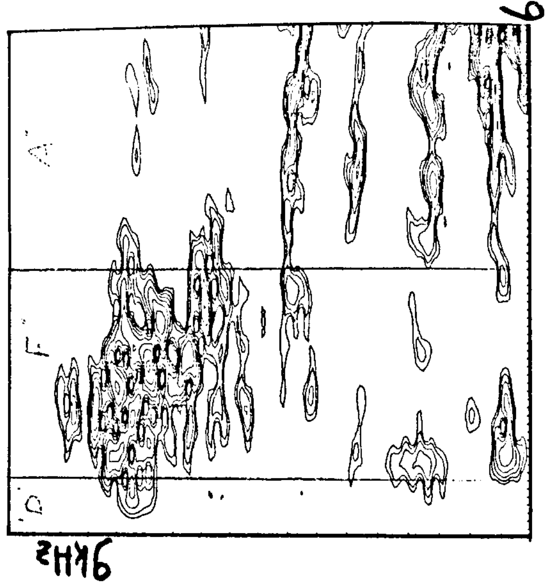


Fig. 9: Spectrogram of explosion interval in first [t] from [ta:t]
Δf=100Hz, Δt=6ms, ΔtΔf=0.6

The frequency resolution Δf = 100 Hz is just great enough to damp interferences between successive pitch periods (remember interference oscillations to occur perpendicular to the line from one excitation impuls to the next, i.e. parallel to the frequency axis!). The time resolution Δt = 0.5 msec is great enough to damp most of the interferences between neighbouring formants. Accordingly, oscillations in the time direction can only be observed between F1 and F2, the two formants closest to each other.

Figure 5 shows a Spectrogram of the same signal segment with resolutions Δf = 300 Hz and Δt = 2 msec. Note that the product of these resolutions equals 300 Hz · 2 msec = 0.6 which is more than ten times the product of resolutions of SPWD in Figure 4 (100 Hz · 0.5 msec = 0.05). The Spectrogram's resolutions are already chosen so as to achieve a signal representation as close as possible to SPWD. A simultaneous improvement of the Spectrogram's resolutions is impossible due to (5). Therefore the Spectrogram evidences much broader excitation impulses in the time dimension as well as much wider formants in the frequency dimension than SPWD. The inherently stronger smoothing of the Spectrogram renders better suppression of interference terms (though they are still perceivable between F1 and F2), yet worse fidelity in signal terms than SPWD. As interference terms are predictable from the above geometrical rule, SPWD is better suited to the analysis of vowels than the Spectrogram.

One may conjecture that a change of the Spectrogram window function w(t) in (2) may improve its resolutions. This has to be refuted when studying Figures 6 and 7. InFigure 6, the time resolution of the Spectrogram is improved to Δt = 1 msec, thus approaching the value of Δt for SPWD inFigure 4. Due to the uncertainty relation (5), the frequency resolution goes up to Δf = 600 Hz so that the two lower formants F1 and F2 are merged into a single unstructured lump stretching over several hundred Hz. In Figure 7 , the frequency resolution of the Spectrogram is improved to Δf = 100 Hz as is the case for SPWD in Figure 4. As time resolution has to go up to 6 msec, excitation impulses are broadened drastically and spilled over the formant structure even into the interval of the pitch period without glottal excitation. Therefore, formant bandwidth measurements are again more difficult than with SPWD, inspite of the high frequency resolution of Figure 7.

Summarizing we observe that the Spectrogram is not suited for *simultaneous* display of both the excitation and the formant structure of vowels whereas SPWD has this property notwithstanding its (easily controlled) interference terms.

6. ANALYSIS OF UNVOICED STOP CONSONANTS BY SPWD AND SPECTROGRAM

Figures 8 and 9 show the *explosion* interval (60 msec) of the first [t] in the German word [ta:t] making use of SPWD and Spectrogram, respectively. For the sake of comparison, both displays have a frequency resolution of 100 Hz and associated time resolutions of 1 msec (SPWD) and 6 msec (Spectrogram). The explosion interval consists of three more or less separable phases:

1. An impulse-like transient (about 4 msec) due to the release of the pressure built up behind the vocal-tract closure (*plosion* phase P).
2. A noise phase extending from approx. 4 kHz to 8 kHz (25 msec) due to the turbulent air flow at the opening constriction (*frication* phase F).
3. A noise phase with a formant structure (30 msec) due to the resonances of the open vocal tract excited by turbulent air flow at the glottis (*aspiration* phase A).

The adavantges of SPWD over the Spectrogram for the analysis of this type of sounds can be summarized as follows:

(1) The short impulse of the plosion phase P is readily seen in SPWD whereas the Spectrogram is not able to resolve this temporal fine structure (at the given frequency resolution).
(2) The boundary between frication phase F and aspiration phase A is more pronounced in SPWD than in the Spectrogram.
(3) Noise-like excitation of the vocal tract manifests itself as a very specific meshy *texture* in SPWD which is clearly distinguishable from deterministic excitation as seen inFigure 4. With the Spectrogram, noise-like excitation induces no significant changes in the texture if the contour plots when compared to deterministic excitation as seen in Figures 5, 6, and 7.

7. CONCLUSIONS

From the above discussion, it should be clear that SPWD is superior to the Spectrogram for the time-frequency analysis of speech signals as typified by the examples given in sections 5 and 6. It should be kept in mind, however, that the comparison was made on the basis of very short signal segments so as to emphasize SPWD's character as a time-frequency "microscope". If the analysis interval is extended to 1 second or more both the resolutions of video displays and the human eye become insufficient to realize the differences of the two methods. Anyway, these long-time displays are only useful for the compressed visualization of slowly time-varying and global features characterizing whole syllables or words. For the detailed high-resolution study of rapidly time-varying speech phenomena, preference is to be given to the new method.

8. REFERENCES

[1] T. Claasen, W. Mecklenbräuker, 1980; The Wigner Distribution - a Tool for the Time-Frequency Signal Analysis; Philips J. Res. Vol. 35 pp. 217-250, 276-300, 372-389

[2] F. Hlawatsch, 1984; Interference Terms in the Wigner Distribution; International Conference on Digital Signal Processing; Florence, Italy

METHODS OF SPEECH SIGNAL PARAMETRIZATION BASED ON GENERALIZING OF LINEAR PREDICTION

A.N. Sobakin

Moscow, USSR

ABSTRACT

The generalization of speech analysis method on the basis of linear prediction reveals unused potential possibilities of this method and permits to develope new algorithms of evaluating speech signal parameters.

INTRODUCTION

Modern achievements in the sphere of speech analysis and synthesis are mainly connected with the use of algorithms of speech signal parametrization, that take into consideration in some degree the nature of speech production.

According to Fant's model [1], the speech production consists of excitation signal transformation by the linear dynamic system (LDS), which parameters correspond to the state of vocal tract at the moment of articulation.

The change in the vocal tract state during articulation leads to the LDS parameters modification.

The tracing of these changes is usually carried out by shifting analysis window within which the LDS parameters may be considered to be sufficiently stable. The transfer function of such LDS at the analysis interval has the form of fraction-rational function with zeroes and poles.

The signal at the LDS input is looked upon as a sequence of alternating intervals, corresponding to voice or noise excitation. The whole excitation signal in that case is modulated by the time envelope of the speech signal.

Linear prediction [2-6] as a method of speech signal analysis was worked out on the basis of much more simplified pattern of speech formation, than one described above.The method is based on deriving the LDS parameters according to the speech signal estimates, ignoring transfer zeroes within the analysis interval. The most simple calculation formulas are obtained in the metrical space.

The quality of obtained LDS parameters estimates will essentially depend on the location of the analysis window at the time axis.

If the interval of analysis corresponds either to an interval of noise excitation or to an interval of free LDS oscillations (for example, the interval of vocal cords closure) then it is possible to show, that in that case the estimates will be unbiassed.

But in case when the analysis interval contains one or several pitch impulses, LDS parameters estimates will be biassed. It is explained by the misagreement between the analysis method and the speech signal structure, for example at the voiced intervals of speech.

Thus. the problem of more complete agreement between the analysis method and the speech formation pattern is an urgent issue.

According to the said above, it seems perspective to examine possible linear prediction generalizations, introducing additional parameters and characteristics of the method. Additional degrees of freedom may be used for more complete agreement between the method of analysis and the speech signal structure.

The generalization of linear prediction leads to the algorithm modifications of speech signal parameters estimates, and, in the long run to obtaining new parametrical spaces for analysis and speech recognition.

GENERALISATION OF LINEAR PREDICTION

The essence of linear prediction is nonrecursive p-order filter that transforms the speech signal counts [x] into residual signal e[n], using weight coefficients $\{A_k\}$:

$$e[n]=\sum_{k=0}^{p} A_k * y_k[n], \qquad (1)$$

$$A_0=1, \qquad (2)$$

where $y_k[n]=T^k(x[n])=x[n-k], \quad k=0,1,\ldots,p; \qquad (3)$

$T^k(\ )$ k-power of the delay operator.

When analysing speech optimal coefficients of filter (I) $\bar{a}_{opt}=\{A_0,\ldots,A_p\}$ are determined from condition of minimum residual signal deviation $\{e[n]\}_0^M$ from the coordinate beginning in the metric space L_2 within the analyses interval [0,M]:

$$\bar{a}_{opt}=\arg\min F(\bar{a}), \qquad (4)$$

where $F(\bar{a})=\sum_{n=0}^{M} e^2[n] \qquad (5)$

-is a squared quality criterion.

Suggested linear prediction generalization concerns two filter components : extension of operator class , on the basis of which it is formed (3) and generalization of constraint imposed on it's coefficients (2).

Quality functional is also generalized (5), that allows to choose different metric spaces for estimation of analysis parameters.

As seen from (3),the original transformation (1) was formed on linear delay operators,that represent the class of physically realizable linear systems with constant parameters.Principally,it is possible to substitute the original delay operators for a set of any stable operators $U_0,U_1,\ldots,U_p$ from the class indicated.

Then proportion (3) is transformed into corresponding cascade form as follows :

$$y_k[n]=U_k(y_{k-1}[n]), \qquad k=0,1,\ldots,P; \qquad (6)$$

where $y_{-1}[n]=x[n]$.

Each linear operator (6) is determined in the frequency sphere by the transfer function of fraction - rational type.

According to the speech signal physical characteristics,the choice of transfer function parameters allows to change in necessary direction the structure and features of linear transformation (1).

Thanks to that,the agreement between algorithm analyses and dynamic speech characteristics will be achieved.

The cascade form (6) of transformation (1) also allows examine the corresponding generalized structures of lattice filters [7] on the basis of linear operators specifically chosen.

It is worth-while to note ,that besides cascade form (6) ,the parallel form of the speech signal preliminary transformations can be easily formed on the basis of indicated set of linear operators. Each of the output signals $y_k[n]$ is obtained as the result of application the corresponding operator directly to the input signal x[n].

The condition (2) influences the structure and features of filter (1) not to a lesser degree.

This limitation for parameters of the filter was introduced to eliminate zero solution during the search for quality functional minimum. In essence it can be considered as the constraint on vector $\bar{a}$ coordinate magnitude in (p+1)-dimentional space of parametres.

In general,this constraint may be written down as follows:

$$f(\bar{a})=f(A_0,A_1,\ldots,A_p)=0 , \qquad (7)$$

where f()is an arbitrary function of (P+1) variable.

The only condition of choosing the function is zero solution elimination in the problem under consideration.Thus,the equation (7) in (P+1) space of parameters determines a surface,not passing through the coordinate beginning.

The search for an optimal vector of coefficients $\bar{a}_{opt}$ with constraint (7) may be realised on the basis of generalized quality functional $F_r(\bar{a},b)$:

$$F_r(\bar{a},b)=\sum_{n=0}^{M} |e(n)|^r + b * f(\bar{a}), \qquad (8)$$

where b - the Lagrange factor from the set of real numbers, R,r - an integer.

The value of "r" in (8) determines the choice of the metric space L_2, where the search for optimal vector of parameters $\bar{a}_{opt}$ is carried out.

Lagrange factor b increases by one the amount of target unknown values and reduces the problem of conditional extremum searching to the search for unconditional extremum for quality functional (8).

As before,condition (4) in which functional (8) was used instead of functional (5) ,determines vector $\bar{a}_{opt}$ and factor b in expanded (P+2)-dimentional metric space $L_{r,1} = L_r * R$.

Proceeding from condition (4) of the quality functional minimum (8) ,the task of searching filter (1) parameters may be presented as generalization of linear prediction method.

The particulare choice of basic operators (6) of limiting function (7) and characteristical constant determines in each case different algorithms of speech signal analysis and different parametric spaces for their description.

ON THE CHOICE OF BASIC LINEAR OPERATORS, LIMITING FUNCTION AND METRIC SPACE.

Among three components,that determine the particulare form of analysis algorithm in the formulated task,the most promissing and the most difficult at the same time is the problem of the best choice of basic linear operators (6).

As in classical methods of digital filters design [8], the complexity of this problem for the class of linear systems with infinite impulse responce (IIR-filters) is increasing as compared to the choise of linear operators from the class of linear systems with finite impulse response (FIR-filters).

Let's confine to setting a mathematical problem of choosing operators (6) from the FIR-system class with impulse responses of p-length. In that case the set of transformations (6) is represented by a linear equation system,that is formed with the help of square B matrix of (P+1)*(P+1) size.

B matrix lines are the impulse responses of the basic operators,derived from the above mentioned class of the FIR-systems.

In that case,the set of operators (6) in parallel form is expressed by delay operators (3),and the corresponding vectors of coefficients $\bar{a}$ and $\bar{c}$ for both variants are related to each other by the following linear equation system:

$$\bar{c}' = B' * \bar{a}' \qquad (9)$$

(an accent means transposition).

In case of B matrix inversion,parameters $\bar{a}$ and $\bar{c}$ are equivalent according to the information theory.

However,the latter doesn't mean their equivalence from the viewpoint of their optimum coding for speech transmission and recognition. Thus,the problem of the best choice of basic FIR-systems is formulated as the problem of transformation search (9) (i.e. B-matrix),that brings about the improvement of estimated parameters in the systems of speech transmission and speech recognition.The B matrix choice allows to take into account more completely the speech signal structure and features.

Using the linear operator theory in Gilbert spaces [9] it is possible to approximate any linear operator from the IIR -system class by a linear operator from the FIR-system class.The problem of the optimum choice of basic operators from the class of IIR-systems may be reduced to the above formulated task of FIR-systems .

It doesn't seem possible to examine different variants of condition (7) fuly enought. Let's confine ourselves to 2 types of function f(.).

For predicting methods,the choice of function f(.) in the form of scalar product of weight coefficients $\bar{q}$ by parameter vector $\bar{a}$ is a natural generalization of constraint (2):

$$f(\bar{a})=(\bar{q}\ ,\ \bar{a}) - 1 = 0 . \qquad (10)$$

Equation (10) with minus one in the left part determines a hyperplane in the space of parameters ,that doesn't pass through the coordinate beginning.

Interesting results are obtained if a square form of the parameter vector is taken as the second limiting function:

$$f(\bar{a})=(D\bar{a}'\ ,\ \bar{a}) - 1 = 0. \qquad (11)$$

Equation (11) determines the second order plane in the parameter space with the help of D matrix of (P+1)*(P+1) size.

In both cases, the choice of either particulare vector $\bar{q}$ for condition (10) or D matrix for condition (11) gives aditional degrees of freedom, helping to determine the structure and features of the corresponding estimation algorithm of the speech signal parameters.

The choice of metric space L_r ,i.e. the choice of characteristical number r ,also determines the structure and features of the obtained algorithms.

The most developed and examined algorithms are the estimation algorithms for squared quality criterion in metric space L_2 (r=2).

However, the results of theoretical calculations and experimental[10] researches show that modular criterion (r=1) has the advantages in the speech signal analysis .For example ,single excitation pulses don't distore the target values of the LDS parameters and the obtained parameter estimations are nonbiassed.

It seems interesting to examine the minimax quality criterion for r= ∞ and the obtained results of the speech signal investigation, though there arises the necessity to use complex Remez algorithm [11] for estimating parameters.

THE EXAMPLES OF ANALYSIS ALGORITHMS

In practice the determinating of functional extremum may be carried out in two ways: either on the basis of the equation system that is derived when the quality functional gradient is equal to zero or by adaptive methods [12] in the form of consecutive approximations to target parameters.

The adaptive methods are of the most interest in the sphere of aplied researches.

The system of adaptive equations for determining the LDS parameter estimates in case when m-th coordinate of vector $\bar{q}$ is equal to one and other coordinates are equal to zero,will look as follows:

$$A_k[n+1]=A_k[n]-g(n)*e_1(n)*y_k[n],\ k=0,\ldots,m-1,m+1,\ldots,P; \qquad (12)$$

where g(n) is normalizing multiplier.

The equation system (12) reminds of the system of adaptive equations for linear prediction based on the method of the least squares [6].In fact when the first coefficient is equal to one [a_0=1],the forward linear prediction is obtained ,when the last coefficient is equal to one (a_p=1) the backward linear prediction is obtained.

Thus,there exists a principal possibility to work out filters [7] on the basis of generalized linear operators.

The adaptive algorithm obtained for a unitary D matrix will differ from other known methods of estimating in most degree.Condition (11) in that case will mean that the norm of coefficient vector is equal to one:

$$\|\bar{a}\| = 1. \qquad (13)$$

Equation (13) in parametrical space determines a spheric surface of an unitary radius,within which the search for quality functional extremum is carried out.

The corresponding adaptation equations look as follows:

$$A_k[n+1]=A_k[n]-g(n)\{e_1(n)*y_k[n]-b[n]*A_k[n-1]\}$$

$$b[n+1]=b[n]-g_1(n)*(\sum_{k=0}^{p} A_k^2[n-1] - 1); \quad k=0,1,\ldots,P. \qquad (14)$$

The estimation of coefficient vector, obtained on the basis of equations (13) and (14) is an approximated latent vector value of covariation signal matrix y[0],y[1],...,y[n] that correspond to maximum latent value of this matrix. This algorithm differs from the classical method of linear prediction.

Function e (n), used in equations (12) and (14), is identically equal to residudl signal for squared quality criterion (r=2) and is of the same sign as the residual signal for modular quality critetion (r=1). In these equations normalizing multipliers g(n) and $g_1(n)$ secure the convergence of successive iterations $\bar{a}[n]$ to the LDS parameters optimal value, determined by condition (4).

The initial value of target parameters in adaptive algorithms (12) and (14) may be equal zero.

CONCLUSIONS

Suggested generalization of linear prediction allows to develope algorithms of the speech signal parameters estimation, that differ from traditional ones.

Introduced constants of generalized method, at the stage of joint constraint of coefficients and at the stage of preliminary transformations as well, provide additional degrees of freedom, that allow more completely take into consideration the current speech signal characteristics.

The given examples of the adaptive algorithms show the potential abilities of the examined above generalization of linear prediction method, but it is evident, that the problem of the speech signal parametrization is not solved yet.

REFERENCES

[1] Г. Фант.
Акустическая теория речеобразования.
М. Наука, 1964.

[2] B.S. Atal and M.R. Schroeder.
Adaptive predictive coding of speech signals.
Bell. Syst. Thechn. J.,v.49,pp.1973-1986, oct. 1970.

[3] B.S. Atal and J.R. Hanauer.
Speech analusis by linear prediction of the speech wave.
J.Acoust.Soc.Amer.,v.50,n.2,pp.637-655,Aug.1971.

[4] А.Н. Собакин.
Об определении формантных параметров голосового тракта по речевому сигналу с помощью ЭВМ.
Акустический журнал АН СССР, т.XVIII, вып. 1, стр. 106-114, 1972.

[5] H. Vakita.
Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveform.
IEEE Trans. on Audio Electroacoust.,v.AU-21,n.5.pp.417-427,1973.

[6] Дж. Маккол.
Линейное предсказание. Обзор.
ТИИЭР, т.63, вып.4,стр.561-580. Апр. 1975.

[7] Б. Фридландер.
Решетчатые фильтры для адаптивной обработки данных.
ТИИЭР, т.70,вып.8,стр. 54-98. "Мир",М., 1982.

[8] Л. Рабинер, Б. Голд.
Теория и применение цифровой обработки сигналов.
"Мир", М., 1978.

[9] Н.И.Ахиезер, И.М.Глазман.
Теория линейных операторов в гильбертовом пространстве.
"Наука", М., 1966.

[10] E.Denoel,Silvay J.P.
Lineat prediction of speech with a least absolute error criterion.
IEEE Trans.on Acoust.,Speech,Signal Processing,v.ASSP-33,n.6, pp.1397-1403, 1985.

[11] Е.Я. Ремез.
Общие вычислительные методы чебышевского приближения.
Изд. АН УССР, Киев, 1957.

[12] Я.З. Цыпкин.
Адаптация, обучение и самообучение в автоматических системах.
"Наука", М., 1968.

ABOUT ONE CLASS OF THE PHONETIC UNITS USED FOR SPEECH RECOGNITION

L.L.Besednaya, V.I.Bogino

Institute of Cybernetics
Kiew, Ukraine, USSR 252207

ABSTRACT

A formal approach to attributing of the phonetic units is offered: the limits of the elements are rigidly connected with the behavior of the structural characteristics of the signal regardless of its phonetic essence. The segments obtained form a class of phonetic units of limited capasity. Their concrete linguistic characteristics are useful for speech recognition.

The choice of the unit of the phonetic analysis is of great significance for automatic speech recognition: it influences the means and extent as it is one of the basic stages of speech signal processing. Usually the procedure of speech segmentation is orientated towards the concrete speech units such as phonemes and their variants in speech, diphones, syllables and pseudo-syllables, moreover the advantages of choice either of the units are not obvious. Segmentation is carried out according to the changes of time-parametres with the help of threshold methods. The analysis devoted to this problem shows that this procedure is essentially combined with the process of verification. Segmentation is carried out while speech recognizing being realized as a hierarchical prosedure simultaneously with attributing of the groups to which the phonemes belong.

However the process of segmentation may be also considered as a preliminary stage of speech analysis. In this case simplified ways of cutting speech are possible that are more rigidly connected with behaviour of the structural characteristics of the signal. This supposes more free determining of the limits of the segments in regard to their phonetic essence.

As an example we may consider the possibility of the phonetic units use limited in the speech flow by means of the characteristic behaviour of stationary or non stationary time function of speech signal for recognition of speech communication. The method of automatic segmentation was tested on the feedback educating system model. It showed that the choice of this feature as a segmentating function makes possible to determine the limit of the segment within speech flow taking into account the end of the vowel or the contact before the plosive phonemes with accuracy of 0.93.

Thus we get speech segments including one or more phonemes and building the following phonetic structure: the separate phonemes (C, V), the sequence of the consonant or vowel phonemes (C...C,V...V); the open type pseudo-syllable sequences (C...CV). The number of the like linguistic elements in each language is limited, so it is possible to express any vocabulary by means of alphabet composed of original elementary phonetic segments (EPhoS), posessing identical contents and differing from each other by a phoneme, number of phonemes or their sequence order. The usage of the EPhoS as elementary recognition units makes it possible to distinguish their most distinctive characteristics in comparison with phonemes, because the structure of the E PhoS being more informative, ensures "effectiveness and independance" from variations. Besides, it is possible to use the law of construction of words through the EPhoS within a certain vocabulary. For example, in case of lack or definiteness of information while recognizing a selected element, it may not be identified, and recognition may proceed from the structure of the word on the whole at the following stages.

This approach to determination, of the EPhoS makes possible to get a wide spectrum of the variants of segmentation depending on the choice of the corresponding system of the indications. The use of larger number of indications naturally enables to get such class of the EPhoS which is not so numerous but its elements contain less information. If the number of indications is smaller, a greater number of the original EPhoS is segmentated, though they are more informative representing a larger fragment of the data. Besides segmentation is more effective thank to the choice as segmentating function of the indications revealed at the first stage of the process with a sufficient stability.

There are some 3000 EPhoS in the Russian Language. They are the result of segmentation according to the signs of the stationary structure of the signal within a given time-interval. For processing the authors used: the frequency Russian dictionaries, scientific vocabularies and articles, extracts from newspapers and fiction. On the whole the texts comprised 20000 words. A special complex of algorythms was developed and brought to the programme realization to process the printed texts. The complex comprised the algorythm of automatic transcribing, segmentation, selecting the set of the EPhoS, their statistic processing and coding.

Table

The quantitative components of the EPhoS for different groups of texts

The group of texts	The problem-orientated vocabulary of SAPR	The frequent vocabulary of scientific lexics	The frequent Russian dictionary	Various texts	Texts and frequent Russian dictionary
The total number of the words	1001	2084	8647	7913	16560
The number of the original EPhoS	730	1013	2070	1945	2845

In the table the results on the quantitative components of the EPhoS for the various groups of the texts are summed up. The dynamics of appearing of the original EPhoS (N_{eph}) depending on the capacity of the processed texts (N_w) is shown in Fig. 1 (dependence 1). Here are shown: dependences of the accumulated frequence of appearing F_a (curve 2) and the accumulated time of existance T_a (curve 3) of original EPhoS from the total number of the EPhoS with regulated frequence (Σ_{eph}) : they indicate uneveness of distribution of informative stress of the EPhoS - the most active element, amounting 20% of the EPhoS, covering some 90% of the texts being analyzed and more than 80% of the duration of their pronounciation. It is interesting that the composition of the EPhoS is practically independent on the ana lyzed texts, especially for the active (the most frequent) EPhoS. It enables using one set of standard EPhoS or its main part for automatic recognition of different concrete vocabularies. A special type of the EPhoS is of some interest. It was selected during the following experiment: when a group of free texts was being analyzed it was supposed that the end of the word did not limit the EPhoS. The amount of the elements exceeded the previous figure of the quantitative contents of the EPhoS on account of speech segments, appearing at the border of two words but not appearing inside any word. This class of the EPhoS named disjunctive appeared to make 30% of their total number and determined about 30% of the words from the analyzed free texts. I.e. the disjunctive EPhoS enable to formal speech segmentating into words without drawing the results of sence analysis for that purpose.

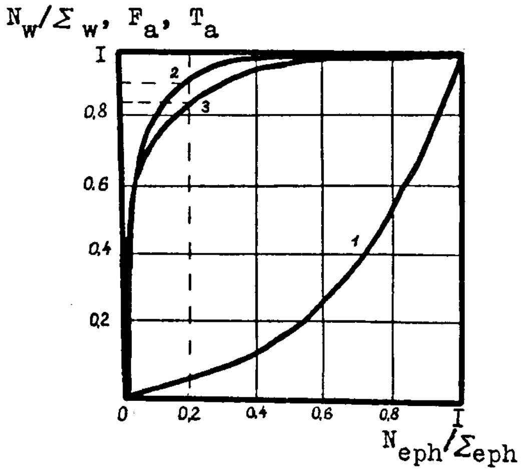


Fig. 1. The dynamics of appearing (1),accumulating of frequence (2) and duration (3) of the original EPhoS.

The study of the results of statistical processing of the EPhoS and their distribution in words, especially for small vocabularies, enabled selecting a few main types of the EPhoS such as: key (appearing in one word), forecasting (selecting a group of words), specifying (defines one from the group of selected words), disjunctive -- their characteristics enable achieving higher parametres of speech recognition procedure.

A full set of the EPhoS containing some 3000 elements was formed as the result of usage of phonetical system of 60 phonemes. However such number of the EPhoS excess. Let us name a sub-multitude of the phonemes taken from their full set a phoneme group united by a stable indication and build a dependence of the number of the phoneme groups ($N_{ph.gr}$) which we get using a definite system of indications (curve 1 in Fig.2). Then let us examine if it is possible to recognize a concrete vocabulary with the help of different sets of the EPhoS differing from each other by the numbers of the phoneme groups used for their identification. It turns out (curve 2 in Fig. 2) that usage of 10-12 phoneme groups (that is 12 - 20% of the total number of the EPhoS) ensures recognition of 80 - 90% of the words of the given vocabulary (vocabularies of 2000 words were examined).

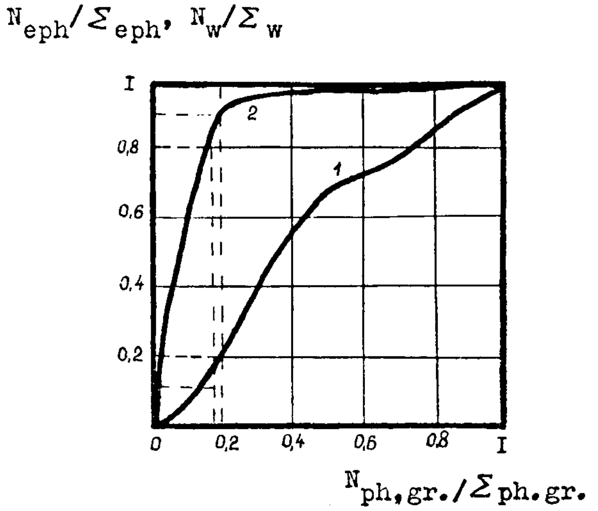


Fig. 2. Dependence of the number of the EPhoS (1) and the number of the recognized words (2) on the number of the phoneme groups.

Analyzing of small vocabularies (60-250 words) used in the systems of various functions shows that it is possible to recognize 95 - 96% of the words of each vocabulary using 30 - 50 EPhoS formed on the basis of 10 or 12 phoneme groups.

The phonetic elements under analysis can be used for speech recognition as well as for speech synthesis. Moreover it is possible to describe separate words as well as continuous utterances.

A SYLLABLE APPROACH TO THE SPEECH INFORMATICS

A.KNIPPER

Institute for the Problems of Information
Transmission Academy of Sciences
101447, Moscow, GSP-4, USSR

ABSTRACT

A problem devoted to trends of syllable development for the usage in speech informatic systems is under consideration. It is noted that simple open CV syllables (C is a consonant, V is a vowel) are the most stable discrete phonetic units of continuous speech with respect to the context , speaker variability and noise. Problems of that type syllable classification and statistics for the Russian speech and their relations with letter records of the speech information are discussed. Some experiments on compilation syllable synthesis of the Russian speech of free contents and on analysis of the speech signal using CV fragments are briefed.

INTRODUCTION

The main problem of the speech informatics is development of man-machine communication systems on the base continuous speech. In that case speech communication between a user and a system is ensured with the best conditions. In continuous speech recognition / understanding the most promising approach is representation of the speech flow with the help of symbol sequences similar to the speech transcription with afterwards decoding at the word or phrase lewel / 1, 2, 3 /. The main requirment of that approach is transformation of a continuous signal into a discrete sequence of speech elements, phonetically stable to speaker variability, context , noise and other facts which influence the speech signal features. I that case in the process of speech recognition / understanding system operation and its new vocabulary training the conviniency for users is ensured / 4 /.

In the process of synthesis of any piece of speech information an iverse problem is solved, i.e. a letter sequence is transcribed by phonetic symbols and then is transformed into the corresponding acoustical signal, and besides for comfortable usage it is desirable to synthesise any voice and any speaker sterotype according to a user choice.

The choice of a phoneme as a phonetical symbol for a speech communication system is the most reasonable and convinient, as it permits relatively easily to pass to the conventional letter representation of any data, accessible and intelligible by broad circles of users. However, numerous researches on phonetics and speech informatics show that there is no direct relationship between speech segments and phonemes. The same sounds match speech segments with essentially different spectral and temporal characteristics, that is determined by context, positional and speavariability of the speech. Simple open CV syllables have more stable characteristics especially those that are cut off from left and right from a transition line, named CV fragments / 5, 6 /. It is considered that CV syllables are base speech elements for Russian, Italian, Japanese and other languages / 7, 8, 9, 10, 11, 12/ and is more widely used in different speech informatic system. In the following sections it is shown that usage of CV syllables as base units of Russian permits to perform a rather distinct classification of context depended pairs of sounds and to choose the minimal alphabet of discrete phonetic elements which describe the continuous speech.

BASE ELEMENTS OF THE RUSSIAN SPEECH

It is considered that open syllable is a speech universal unit for the majority of languages / 8, 9 /. In the speech informatics open syllables may be also prefered, due to the fact that there is a distinct transition from the corresponding consonant to the vowel in the interval of that open syllable that makes easier to label the continuous speech visually and with the help of technical means /5, 6, 13 /. The number of open syllables for Russian is great, i.e. about 2500 / 14 /, however any open syllables may be represented as a concatenation of the base CV syllable and separate consonants and vowels. Thus compound open syllables such as CCV, CCCV, CVV may be expressed by C+CV, CC+CV, CV+V correspondingly those constructions the most strong coarticulation can be observed in CV combinations / 8, 10 /, that determines the necessity to examine that speech element as a whole. In that case CV syllables cover about 80% of any text of the Russian speech. The stressed and the first prestressed syllables with stabile phonetic quality have frequency occurrence in the text equal to 50% /6/.

For the needs of analysis and synthesis it is useful to represent the base CV syllables in the table form. In that table consonants and vowels are written not in the phonetic symbols, but in traditional Russian letters, that allows to transform a written sequence of letter symbols into the corresponding syllable one. Twenty consonants are written in the vertical column of the table according to the manner of production, and in the horizontal rous according to the plase of articulation (labels L, D, A, P, Vl,V, Lq and N correspond to labial, dental, alveolar, palatal, voiceless, voiced, liquid and nasal consonants). Ten vowels are divided into two groups which form hard or soft variants of

Table. Classification of the base elements of the Russian speech

Cosonants							Vowels									
Place of articulation Manner of production			L	D	A	P	Hard					Soft				
							А	О	У	Ы	Э	Я	Ё	Ю	И	Е
							I	2	3	4	5	6	7	8	9	IO
I	Fricatives	Vl				Х			/	///	///	///	///	/	/	/
2					Ш				/		/	Х	Х	Х	Х	Х
3					Щ		Х	Х	Х	Х	Х	/	/	/		/
4				С				/		/	/	/	/	/		
5			Ф				/	/	/	/	///	///	/	///	/	/
6		V			Ж		/	/	/		/	Х	Х	Х	Х	Х
7				З				/	/	/	/		/	/	/	/
8			В								/	/	/	///		
9	Affricates				Ч		Х	Х	Х	Х	Х		/	/		/
IO				Ц				/	/			Х	Х	Х	Х	Х
II	Plosives	Vl				К		/		///	///	///	///	/	/	/
I2				Т					/		/	/	/	/		
I3			П							/	/	/	/	/		
I4		V				Г				///	///		/	/		
I5				Д							/	/	/	/		
I6			Б				/		/	/	/	/	/	/		
I7	Sonants	Lq		Л						/	/		/	/		
I8					Р			/		/	/	/	/	/		
I9		N		Н				/		/	/	/	/	/		
20			М								/	/	/			

consonants in the base CV syllables. Diphthongs Я,Ё , Ю belong to the soft vowel as well. Vowels pairs А - Я ,О - Ё ,У - Ю , Ы - И ,Э - Е have similar properties in ranking according to their typical duration and positional variability, however , they essentially differ in spectral and temporal characteristics of the transition segment of sounds. At the same time the place of articulation of a consonant influence on the transition segment of a vowel, therefore CV syllables including consonants with the same place of articulation have similary characteristics for the initial part of each vowel. Consonants of CV syllables have the colour of the following wovel due to the effect of coarticulation and that effect is more associated with the place of articulation than the manner of production of consonants. Thus the characteristics of consonants and vowels determinates their context (allophonic) variability. The table is made for the strssed syllables and besides in the cells, which are formed at the intersection of consonant rows and vowel columns, the rough frequency of occurence of the base syllables from / 6, 14 / is given. From 200 possible CV combinations of Russian 25 are not used and 14 combinations occur vary seldom, those syllables are labeled by X and /// in the table. Seventy syllables corresponding to the table empty cells are used more often and cover about 50% any Russian speech and 91 syllables marked by / occur less frequetly. Thus the common number of Russian base elements is relatively not great. For unstressed CV syllables all consonants have realisations with rather good phonetic quality, and the number of vowels decreases up to three, including only sounds А , У and И / 8, 9 /. The duration of unstressed CV syllables shortens 1.5 or 2 times as much as the duration of the stressed syllables both at the expense of consonants and vowels.

SYNTHESIS OF ANY SPEECH ON THE BASE OF CV SYLLABLES

In the process of synthesis on the base of CV syllables two aims were pursued first, a practic one, to develop a speech synthesator which could synthesize any Russian text any speakers voice, including a female one. The second aim was to make clear if it is possible to synthesize a speech signal perceived as qualitative continuous iformation, concatenated from a minimal alphabet of discretely pronounced speech elements. For that purpose a group of speakers pronounced (in according with the table) 175 syllables and 10 vowels afterwards stored in the computer "Eklipse - 330". For each CV syllable places of transitions from consonants to vawels were marked using a graphic display with the following audition and correction, those data and syllable duration data were recorded into the computer memory. Unstressed, reduced syllables and separate consonants can be formed due to shortening of vowel duration of any stressed CV syllable. The perception of hardness or softness of Russian consonants was achieved due to the maximum vowel reduction of any syllable. The effect of coarticulation between cosonants in compound open syllables was produced by concatenating syllables reduced to minimum, those syllables having the same vowel as the base CV syllable. Coarticulation in the words consisted of concatenations of open syllables with different vowels was simulated with the help of addition of a short segment of the succeeding vowel to the end of the preceeding one. The duration of that segment depended on the contrast F-picture of the adjacent vowels and was increasing proportionally to that contrast increase. A more detailed description of CV syllable compilation speech synthesis is given in / 15, 16, 17 /. Algorithms and computer programmes of the syllable synthesis including phonetic transcription of any Russian text were developed by I.Orlov /18/, using the syllable interpretation of a letter record in accordance with the table. Concatenation of speech elements into a continuous piece of information was produced without any additional transformations exept the preliminary amplitude compression of the sygnal. The speech compiled of discretely pronounced syllables sounded as continuous and rather naturally with high percentage of word intelligibility equal to 97 - 99%. That experiment besides having practical significance proves that CV syllables are really the base elements of the speech.

CV ANALYSIS OF CONTINUOUS SPEECH

The syllable analysis of continuous speech pursueing an aim of automatic transcription of a speech signal is much more complicated than the problem of the speech synthesis. Difficulties of the speech analysis mainly depend on the variability of a speech signal and were briefed in Introduction. However, the choice of an analysis unit is of great importance since in addition the number and the type of the base speech elements are determined and their spectral and temporal characteristics become preliminary known as well. The continuous speech analysis as well as the speech synthesis is reasonable to carry out on the base of CV syllables. That approach is discussed in details in / 5, 6, 19, 20 /. That is why we brief here only some conclusions.

1. The number of base CV syllables as well as in the speech synthesis is equal to about 200.

2. A current analysis of the continuous speech should be performed using fragments with duration of about 100-120 ms, in that case the dependence of spectral and temporal characteristics of CV syllables on the context and the position decreases and besides the analysed segment of a vowel should be 2o-25 ms longer than of consonant. In addition to CV syllables it is necessary to extract separate consonants and vowels which form compound open syllables as CCV, CCCV, CVV etc. approximately at the same time window as CV segments. Naturally in that case very short sound wouldn't be extracted, but their number in the Russian continuous speech is insignificant / 6 /.

3. It is useful to perform linear time normalisation of the CV fragment duration dependent on the speech rate typical for a definite speaker / 20 /.

4. A base problem in determination of rules for fragment extraction from continuous speech is parametrical representation of a signal. A lot of experiments show that the best speech representation is a formant one using pitch synchronisation / 6 , 19, 21 /.

CONCLUSION

The syllable approach has good prospects for usage in speech informatics, since it establishes sufficiently adequate correlation between physical and phonetic properties of continuous speech. However, that and higher levels of speech processing are specific for each national language and therefore they should be thoroughly studied for any language.

REFERENCES

/I/Кринов С.Н.,Савельев В.П.,Цемель Г.И. Классификация сегментов при распознавании устных команд.Труды АРСО-I3,ч.I, Новосибирск, I984, IОI-IО3.

/2/ Зигангиров К.Ш., Сорокин В.Н. Применение последовательного декодирования к распознаванию слитной речи."Проблемы передачи информации",№ 4,I977, 8I-88.

/3/Кринов С.Н.,Слуцкер Г.С. Многоуровневая речевая диалоговая система "САПФИР", Труды АРСО-I4,ч.I, Каунас, I986, 92-94.

/4/ Кринов С.Н.,Слуцкер Г.С. Автоматическоэ формирование звуковых эталонов слов по их орфографической записи.Труды АРСО-I4, ч.I, Каунас, I986, стр.86.

/5/ Книппер А.В.,Мирошников В.С. Текущее выделение фрагментов в речи. Труды АРСО-6, Таллин, I972, I07-IIО.

/6/ Книппер А.В.,Махонин В.А. К описанию речевого сигнала."Речевое общение в автоматизированных системах". М.,"Наука", I975, 46-59.

/7/ Чистович Л.А. и др. Речь.Артикуляция и восприятие. М.-Л.,"Наука", I965.

/8/ Бондарко Л.В. Звуковой строй современного русского языка."Просвещение",I977.

/9/ Златоустова Л.В., Потапова Р.К., Трунин-Донской В.Н. Общая и прикладная фонетика. Изд. МГУ. М., I986.

/IО/Francini G.L.,Debiasi G.B. and Spinabelli. Study of a system of minimal speech repr. units for Italian speech. JASA,v.43,N6,June 1968, 1282-1286.

/II/Tanaka A.,Togava F., et all. A study of the syllable oriented recognition of continuous speech. Speech Communication, 1983,v.2,N2-3, 207-210.

/I2/Потапова Р.К. Слоговая фонетика германских языков.М.,"Высшая школа", I986.

/I3/Книппер А.В. СГ-представление русской речи.Тезисы докл. конф."Просодия речи", (7-9 дек.), МГПИИЯ,М.,I982, II7-II9.

/I4/Елкина В.Н., Юдина Л.С.,Статистика слогов русской речи.Сб. трудов ИМ СО АН СССР "Выч. системы",№IО, I964,58-78.

/I5/Вайншток А.П.,Книппер А.В.,Орлов И.А., Потапов В.Г. Фрагментная компиляция речи.Труды АРСО-I2,Киев-Одесса,I982, 387-389.

/I6/Вайншток А.П.,Книппер А.В.,Орлов И.А., Потапов В.Г. Способ слоговой компиляции речи. А/с №IО75300,Б.И.№7,I984.

/I7/Книппер А.В.,Орлов И.А. Компиляция речи с учетом коартикуляции.Труды АРСО--I3, Новосибирск, I984,I5О, I5I.

/I8/Орлов И.А. СГ-слоговое преобразование текста для компиляционного синтеза речи. Труды АРСО-I4, ч.2,Каунас,I986, стр. 64.

/I9/ Knipper A. Acoustic events in CV syllables with liquid and nasal sounds. Signal Processing , N 3, 1981, 389-396.

/20/Книппер А.В. О нормировке СГ-фрагментов по темпу. Труды АРСО-I3,Новосибирск, ч.I, I984, стр.IОО,IОI.

/2I/Книппер А.В.,Махонин В.А.,Орлов И.А. Элементы формантного анализатора. В кн."Распознавание образов. Теория и приложения", м.,"Наука",I977,I2I-I25.

AN EVENT-BASED APPROACH TO AUDITORY MODELING OF SPEECH PERCEPTION

Toomas Altosaar and **Matti Karjalainen**

Helsinki University of Technology, Acoustics Lab.
Otakaari 5A, 02150 Espoo, Finland
Tel. (358-0) 451-2794

ABSTRACT

Auditory modeling is usually based on peripheral physiological phenomena. It is found, however, that this basis is not sufficient in all applications, e.g. in successful speech recognition. Our opinion is that more important than the details of periphery is to include higher-level functional processing in the models. This paper describes an experimental system that uses several spectral and temporal representations to create a hierarchical description of speech. The front-end processing is performed by an auditory model which is based on psychoacoustical principles. Several temporal and spectral representations are extracted from the resulting auditory spectra and are viewed under multiple time resolutions to yield reliable and flexible descriptions of the speech. Based on these spectral and temporal resolutions prominent extrema are located and are classified as objects called events. These objects are organized into event lists according to masking criteria and measures of prominence.

INTRODUCTION

The usual basis for auditory modeling is peripheral physiological phenomena. Transmission-line or filter-bank models are used for basilar membrane and neural models for the next stage, e.g. [1], [2]. This may give a detailed picture of the periphery but the models tend to become overly complicated and there is a certain lack of knowledge of how the higher levels work.

Another approach to auditory modeling is to apply psychoacoustical theory and knowledge. Here we can concentrate on wider functional properties of hearing that are not always directly related to physiological details. Surprisingly few models are explicitly based on psychoacoustics.

The limited success of auditory modeling in speech recognition shows that an auditory front end does not necessarily solve existing problems. We have to pick up the most essential peripheral features and combine them with higher-level symbolic processing. With this approach we are immediately faced with several problems, some of which we hope to solve by formalisms proposed in this paper. There is not much hope to find principles with evidence and support from concrete hearing research. Instead we, have to use hypothetical models that could be possible in the human auditory system.

The central problem for us appears to be in the transformation from a continuous-time speech signal to a discrete and symbolic representation without loosing any key information. The traditional pattern matching and decision process isolates the continuous and discrete domains in a way that makes it very hard to pay attention to the most essential features in a given context.

There are several concepts that we have found to be important. Retaining redundancy with multiple feature representation at each level of the auditory process and even multiple resolution analysis of each feature is needed. This presumes parallel processing to a large extent if such a system is to be implemented in real-time.

Other key concepts in our approach are events and event stuctures. Instead of segments with time boundaries we analyze events (time objects) with rich internal structures: time moment, effective time span, type according to several criteria, amplitude or prominence, link to a feature it is supported by, etc. The list-like data structures consisting of events form the basis for flexible representations that can be applied to rule-based processing at several levels of auditory modeling.

The prototype system to be presented in the next section reflects our approach in a preliminary form. It should be considered as a collection of examples to be developed towards a future speech recognizer that includes all phases from a peripheral auditory model to natural language processing.

SYSTEM DESCRIPTION

The system contains many different levels of processing ranging from auditory modeling of the speech input to symbol and event processing. Figure 1 shows an overview of the current and proposed system. The following sections explain how the system functions.

Auditory Front End

The system obtains auditory information from a filter bank that closely matches the human auditory system in terms of sound perception. The model [3] is based on the most important features of peripheral hearing known from the theory of psychoacoustics [4] and simulates the human's frequency selectivity and sensitivity as well as its temporal and masking properties. By the use of this model only relevant auditory spectral information is retained. Irrelevant information is efficiently removed during the early stages reducing the computation rate in later processing.

The auditory model is implemented as a filter bank and its output is represented by a 48 element spectral vector for each point in time. The vector's elements indicate approximately the amount of energy falling in 1 Bark (1 critical band) regions of the auditory spectrum and are scaled in loudness [4]. Each channel of the filter bank is separated by 0.5 Barks and this provides adequate frequency resolution over the entire 24 Bark auditory spectrum. A spectrum is calculated every 10 ms.

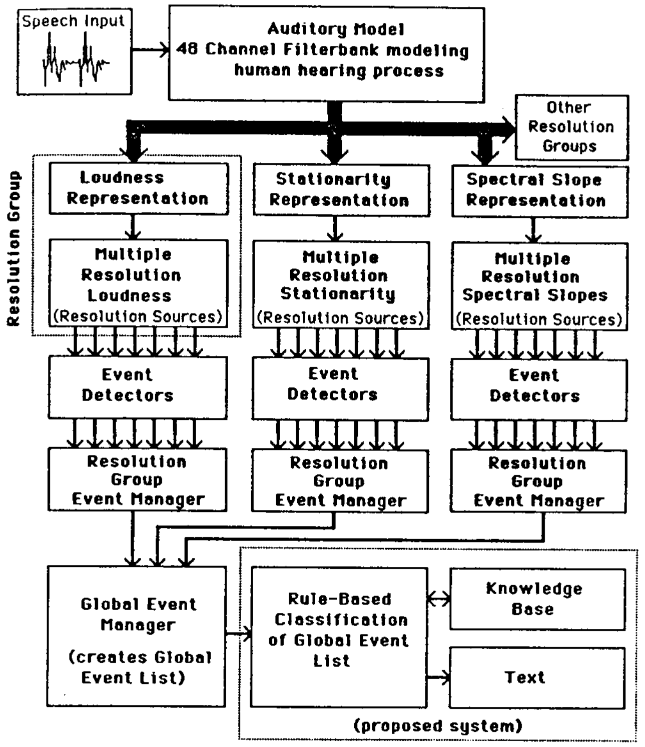


Figure 1. *Signal-to-Symbol Speech Analysis System.*

Multiple Representation Analysis

The loudness scaled auditory spectra are transformed into several parallel representations which help to identify the different speech features and events. These representations can be separated into two major groups: the frequency domain, and the time domain. These groups are described in the following sections.

Frequency Domain Processing

The frequency resolution of the hearing system to broadband signals is at best 1 Bark. For phonetic classification of speech signals different studies have shown that this can vary from 1 to 3.5 Barks. We can simulate this effect by bandpass filtering the spectrum in the *frequency* domain to emphasize the desired resolutions. This bandpass filtered spectrum representation is called the **formant spectrum.** Adequate resolution has been achieved for this system with both 1 and 2 Bark bandwidth filters. The basis for use of multiple resolutions for a single representation is explained later on. The formant spectrum can be used to identify the existence and locations of formants and formant pairs. Formant lists are created by searching for local maxima and indicate where likely formants exist as well as what their amplitudes are but no attempt is made to classify them. The Formant lists are used in an auditory spectrogram display which is shown in figure 2.

Time Domain Processing

The other category of representations are based upon information that the front end supplies in the time domain. One such representation is **total loudness** and is calculated by summing the elements of a loudness spectrum. Total loudness as a function of time reveals the temporal energy structure of the speech while being independent of the individual spectral components.

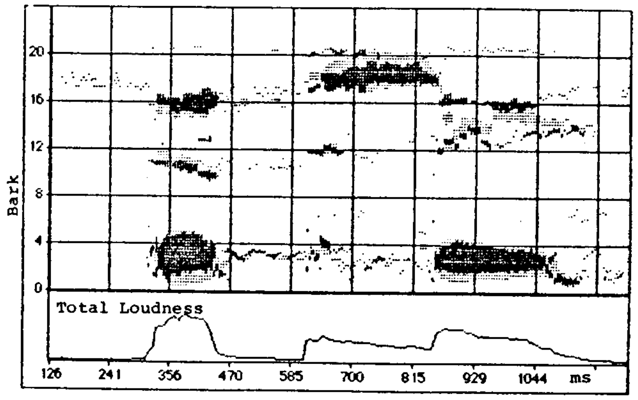


Figure 2. Auditory Spectrogram of the Finnish word /yksi/.

Stationarity is a representation that measures changes in the spectra by comparing the similarity between neighbouring spectra. Stationarity is calculated for time t_i by first finding the average spectra at time t_{i-j} and at time t_{i+j} (average computed over several spectra) and then summing the absolute difference between these averages to yield a scalar measure of distance. This representation is used to identify locations where spectral changes occur and indicates most phonemic boundaries with good reliability. Stationarity is sensitive to both spectral and amplitude changes in speech.

Another representation used in the system is **spectral slope** which indicates where the majority of the energy lies in the spectrum. Four different representations of spectral slope are used: global, formant 1, formant 2, and formant 3 slope. Global slope is a wideband locator of spectral energy while the remaining three analyze the regions of the spectrum where each formant is generally found. These functions are robust indicators of certain features such as fricatives and plosives and can also be used to detect spectral centers of gravity [5].

Time domain multiple representation analysis views the speech signal with several different but parallel perspectives. Figure 3 shows the responses of three representations to the Finnish word /yksi/.

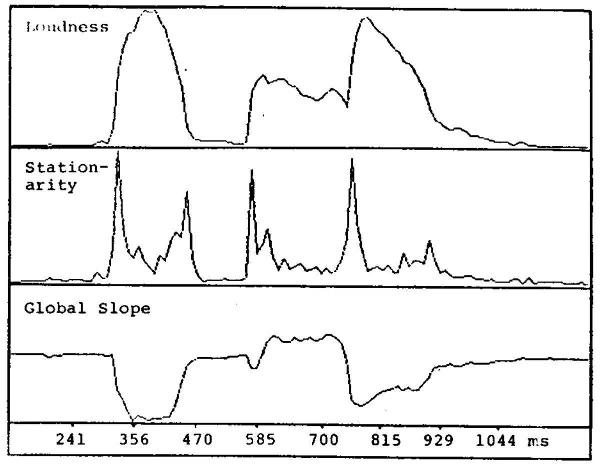


Figure 3. Multiple Representational Analysis of the word /yksi/.

Multiple Resolution Analysis

To obtain a more flexible description of each frequency and time domain representation, all representations are analyzed under several resolutions. This is performed by bandpass filtering a representation with filters having different resolutions. The impulse responses for some of these filters are shown in figure 4. For the frequency domain representation the loudness spectrum is filtered with 1 and 2 Bark resolutions as was mentioned earlier. In this case the filters are scaled in frequency. In the time domain representations the filters are scaled in time, and resolutions of the loudness, stationarity, and spectral slope representations are calculated in a similar way. This method is similar to *scale-space filtering* [6] and is used to generate qualitative descriptions of signals.

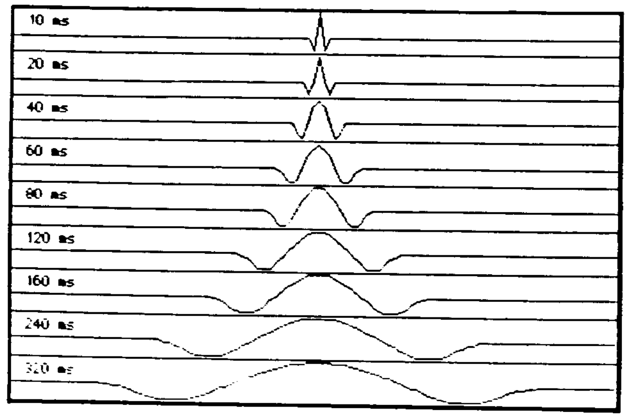


Figure 4. Impulse responses of some of the filters used in Multiple Resolution Analysis.

Each resolution of a representation is defined as a **resolution source** while the representation along with its resolutions is defined as a **resolution group**, as indicated in figure 1. Speech analysis with multiple resolutions facilitates determining event locations and their respective properties with greater ease and accuracy than would be possible with the original representations alone. The curves in figure 5 show the response of the loudness resolution group to the word /yksi/. The multiple/parallel representations and their resolutions allow for a reliable description to be created of the speech. New resolution groups may be added to the system such as pitch detection and a voiced/unvoiced indicator as is found necessary.

Event Detection and Analysis

The next phase of processing transforms a signal, in this case a resolution source, into a discrete and symbolic representation. The resolution groups are operated upon by event detectors which find local extrema and zero-crossings, depending upon which resolution group is being analyzed, and yield symbols as their outputs. Symbols are more flexible to manipulate during later stages of processing than signals since partial classification has already taken place. These symbols may contain information regarding their type, time, amplitude and formant structure. The symbols are ordered chronologically and are placed in a list for later processing.

The resolution group event manager is responsible for analyzing a resolution group and finding the most prominent areas of interest. One measure of prominence is determined by searching for the event with the largest absolute amplitude. It uses as its input the lists of symbols presented to it by the event detector. The resolution group event manager operates on these lists to produce a single list called the resolution group event list that contains the most significant events from a representation.

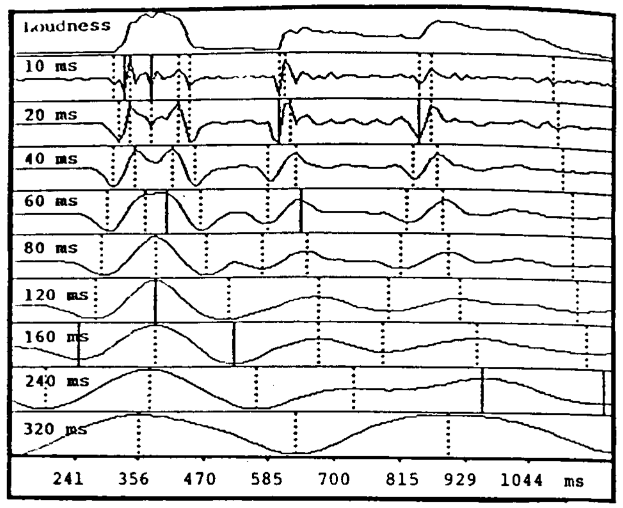


Figure 5. Multiple Resolution Analysis of the Loudness Representation for the word /yksi/. Solid lines indicate events, dashed lines indicate related events.

To avoid multiple entries of the same event in the list, all related events from different resolutions are marked as belonging to the most prominent event. Figure 5 also shows the events (solid lines) found for the multiple resolution loudness representation as well as the related events (dashed lines). Another measure of prominence that could be used is to choose the event with maximum span over the sigma axis when using scale-space filtering techniques [6] to yield a top-level descriptor.

An integrated description of the speech is constructed by the gobal event manager and it considers all the resolution group event lists created by the resolution group event managers and builds a global event list that contains the most prominent events.

The final set of symbols created by the global event manager have been proposed to be used as a primary representation of the speech in a rule-based recognition system. These symbols would describe speech in similar terms as a human would when reading a spectrogram or by listening to speech. The rule-based system would analyze these symbols until enough evidence existed to fully support a hypothesis for final classification. By deferring classification to this final stage, diverse sources of information may be viewed in a global perspective making high rates of recognition possible.

IMPLEMENTATION

The preliminary version of the model is currently implemented on a two processor system. The auditory model filterbank is realized on a TMS 320 signal processor and the remainder on an Apple Macintosh. The Macintosh is the host for the TMS and executes NEON which is an object oriented language [7]. NEON is a hybrid language with many of its features derived from Forth and Smalltalk. The next extended version of the program is being currently implemented on a Symbolics 3670 Lisp Machine including a small-scale speech recognition system.

To efficiently represent and manipulate the different resolutions, representations and symbols, object oriented programming methods are used. Object orientation is a powerful data and knowledge representation principle since knowledge regarding the object is contained within the object itself thus exhibiting object-centered control [8],[9]. Objects can communicate with each other by message passing methods. They also belong to classes and can inherit properties from other classes. This approach allows for building rule and frame-based systems.

Since each representation's analysis can be processed independantly, parallel-processing of the representations, resolutions, and events is a natural topology for the implementation of such a system. Such a concurrent system could be implemented e.g. using Transputers [10] and is one of our long-range goals.

CONCLUSION

Higher-level functional processing must be included in auditory models if the information they supply is to be of greater use. This is because peripheral physiological phenomena often does not offer a sufficient basis for applications such as speech recognition. In this paper we have described an approach to implant the higher-level processing activities into an auditory model. The conversion of speech into a loudness spectrum, the derivation of some representations, and the analysis of these under multiple scale resolutions was explained. Finally, the transformation of signals into discrete frequency/time events was described.

ACKNOWLEDGEMENTS

The authors are grateful to the Finnish Academy which has financed this research.

REFERENCES

[1] Lyon, R.F., A Computational Model of Filtering, Detection, and Compression in the Cochlea. Proc. of IEEE ICASSP-82, Paris.

[2] Lyon, R.F., Computational Models of Neural Auditory Processing. Proc. of IEEE ICASSP-84, San Diego.

[3] Karjalainen, M.A., A New Auditory Model for the Evaluation of Sound Quality of Audio Systems. Proc. of IEEE ICASSP-85, Tampa.

[4] Zwicker E., Feldtkeller R., Das Ohr als Nachrictenempfänger. S. Hirzel Verlag, Stuttgart, 1967.

[5] Chistovich L.A. et al., "Centres of Gravity" and Spectral Peaks as the Determinants of Vowel Quality, *Frontiers in Speech Communication Research,* Academic Press, 1979.

[6] Witkin, A.P., Scale-Space Filtering: A New Approach To Multi-Scale Description. Proc. of IEEE ICASSP-84, San Diego.

[7] NEON Programming Manual, Kriya Systems, Inc., Sterling, 1986.

[8] Winston, P.H. *Artificial Intelligence* (second edition), Addison-Wesley, 1984.

[9] Waterman, D.A. *A Guide to Expert Systems,* Addison-Wesley, 1986.

[10] Transputer Reference Manual, INMOS Limited, October 1986.

PERCEPTION AND MEASUREMENT OF DISTORTION IN SPEECH SIGNALS – AN AUDITORY MODELLING APPROACH

Seppo Helle Matti Karjalainen

Helsinki University of Technology
Acoustics Lab., Otakaari 5 A, Espoo
Finland

ABSTRACT

The perception of nonlinear distortion in speech signals was studied. Subjective listening tests were carried out using Finnish speech sounds as test material. A computational model was used to obtain auditory spectra from the undistorted and distorted sounds, and the spectral difference was compared to subjective sound quality evaluation.

Our studies showed the so-called 2-dB deviation rule to be a useful measure for the just noticeable level of nonlinear distortion. This rule implies that if the changes in auditory spectrum exceed 2 dB, the difference between the original and distorted sound can be perceived. This result also verifies the applicability of the psychoacoustic approach to distortion perception. For distortions exceeding the perception threshold, a more sophisticated objective measure than the maximum spectral deviation is needed. A distortion measurement system based on an auditory model has also been constructed.

INTRODUCTION

The work with auditory models has been active in our laboratory since 1981 /1/ - /4/. One aim of the research has been a psychoacoustical model imitating the human hearing process. A mathematical model that performs this is not a physical simulation of the hearing system. Instead, it attempts to imitate the functional properties of subjective perception of the sound, no matter what kind of physical processes there exist. This is our approach to auditory modelling.

Auditory models can help us, for example, to create better measuring techniques of nonlinear distortion. Conventional techniques, like harmonic distortion measurement, don't take into account how we actually perceive the distortion. This might lead to incorrect results and not to what we want – the sound quality in terms of perceived distortion. If the important properties of the auditory system are built into the measurement method, results can be improved.

Application areas include speech recognition and speech analysis for phonetic speech research. These auditory models can provide some new insights to how we perceive speech.

Some important phenomena of the human auditory system that should be implemented in auditory models are:

- Frequency selectivity of about 1 Bark and masking effect in frequency domain (excitation spreading function).
- Frequency sensitivity of the human ear according to the loudness curves (60 dB-level, e.g.).
- Temporal integration; time response of any 1 Bark channel should be its power lowpass-filtered by a time constant of 100-200 ms.
- Temporal masking; pre-and postmasking effects.

FILTERBANK MODEL

The filterbank principle is well suited to auditory spectrum analysis because the human auditory system — basilar membrane and hair cells — also consists of a multi-channel analyzer /6/. The bandwidth of the overlapping channels is about one critical band or one Bark. Instead of thousands of hair cells it is enough to have 1-4 channels per one Bark in a computational model. This means 24-96 channels covering the 24 Bark audio range. With 0.5 Bark spacing our model has 48 channels, which seems to be a practical compromise between good resolution of spectral representation and low amount of computation.

Each channel consists of a bandpass filter, a square-law rectifier, a fast linear and a slower nonlinear lowpass filter, and a dB-scaling stage (fig.1).

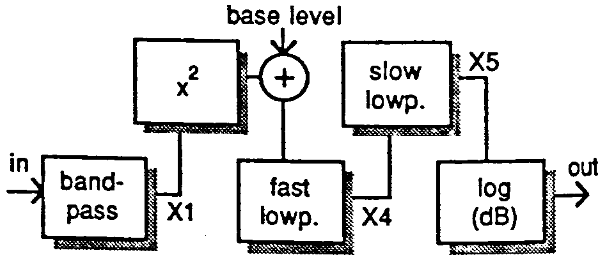


Fig. 1. One channel of the 48-channel filterbank used for auditory spectrum configuration. x^2 = square law detection, log = dB-scaling.

Bandpass filters with 0.5 Bark spacing and a little more than 1 Bark bandwidth give the desired frequency selectivity to the model. Each bandpass is a 256-order FIR-filter designed to have a frequency response which is the mirror image of the spreading function B(x) given by Schröder et al /5/.

Not only frequency selectivity but also frequency response (sensitivity) of the ear must be built into the filterbank. The simple way we used is to let the relative gains of the channels vary according to the inverse of the equal loudness curve (60-dB level).

The rectification effect in hair cells of the inner ear is primarily of half-way type. Our model did not have a half-wave rectifier, because a square-law element was included. We found out that in auditory spectrum analysis of speech this makes no remarkable difference. A constant level is added after the rectification to simulate the threshold of hearing.

The remaining two filters are for smoothing the outputs of the selective channels. The faster one is a first-order low-pass with time constant of about 3 ms. Its role is not important here. The second one is more fundamental. Its purpose is to implement many effects: temporal integration as well as pre- and postmasking.

Temporal integration is realized by linear first-order filtering (time constant about 100 ms) applied to the output of square-law rectification. Premasking is not a very important and critical phenomenon, and this simple solution was quite sufficient.

Postmasking was more difficult to be implemented. A linear lowpass filter with a 100 ms time constant gives an overall masking that is several times too long. To make a better match we used nonlinear (logarithmically linear) behaviour of the filter for masking situations /3/.

PERCEPTION THRESHOLD OF NONLINEAR DISTORTION

One of the most useful rules of the psychoacoustic theory is the 2-dB rule of just perceivable difference. This means that any variation in a sound, resulting at least in about 2 dB level change in any Bark channel, will be noticeable in subjective listening tests. The hypothesis was tested by distorting three Finnish speech sounds /a/, /i/ and /s/ with three nonlinear distortions (square-law, crossover and clipping). Duration of the distorted sound was the third variable. Three persons were asked to find the just noticeable levels of distortions (JND). The test was made by direct comparison of distorted and undistorted signals from a loudspeaker in an anechoic chamber. The corresponding maximal distances in auditory spectra were then computed. The results are shown in fig. 2.

It was found that the types of distortion and speech sound have no essential effect on the auditory spectrum distance of JND-threshold. Duration also has only a minor effect. The 2-dB rule is valid or, more exactly, distortion is just perceivable when the maximum value of auditory spectrum distance is about 1.5 - 2.5 dB (undistorted reference was available to the listener).

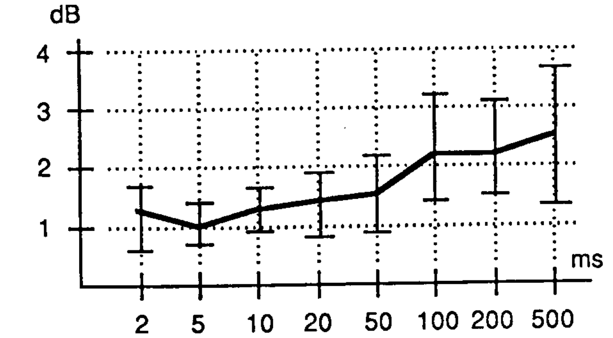


Fig. 2. Auditory spectrum distances corresponding to the JND-thresholds of different distortions applied to three speech sounds (see text) as a function of distortion duration.

An interesting detail is that the the temporal integration must really be present in the model. This also means that if the duration of distortion is less than 100 ms, the physical level of distortion must be higher for short durations to get the same threshold of perception.

In another experiment we found that the perception threshold of distortion without pure reference correponds to 1.5 - 13 dB distances depending on types of distortion and speech sound. We can conclude that if the distance is less than 1.5 dB, the distortion is practically never perceivable.

SUBJECTIVE DISTORTION EVALUATION VS. AUDITORY SPECTRUM DEVIATION

Another series of experiments was carried out later to investigate further the correlation between maximum auditory spectrum distance and subjective distortion evaluations, this time especially for higher than JND levels. Test sounds were Finnish vowels /a/, /i/ and /u/ spoken by two male speakers. Test samples were about 200 ms long and they were distorted artificially with four types of distortions: zerocrossing, clipping, square-law and angle distortions (angle distortion: a piecewise linear input-output relation having an angle discontinuity at the origin). In each test, one of the test vowels was played to the listeners with different distortions in a random order. A test series contained 6 - 8 distortion levels for each distortion type plus clean signals. The undistorted reference could be listened to before the series, but not between the test signals. Each test signal could be repeated as many times as required before making the evaluation using a scale from 0 to 10. Definitions for the values on the scale were:

0 No audible distortion.
1 The listener supposes to have heard something like distortion but is not sure.
2 Distortion is on the just noticeable threshold.
3 Distortion is always perceived when concentrating on listening.
4 Distortion can be heard easily as "soft" distortion.
5 Distortion is not "soft" anymore, but not yet disturbing.
6 Distortion is now disturbing.
7 Listener feels some uncomfort because of distortion but the sound is still easily recognized.
8 Distortion is increased to the level where some problems of correct recognition exist.
9 Recognition of sounds is like guessing.
10 Recognition of the sounds is impossible.

There were three test subjects, all of which listened to each series five times. Figures 3 - 5 show the results from three vowels (/a/, /i/ and /u/) of one speaker. The figures present subjective evaluations of distortion as a function of maximal auditory spectrum distance over time and full 24 Bark range. On the y-axis is the evaluation scale that was used in the test. (Presented are only three of the six test sounds, but the results from the other speaker's sounds were roughly of the same type.)

The plots show immediately that the vowel /i/ is the most sensitive of the sounds: that is, distortion is easiest to detect. The other sounds /a/ and /u/ are less sensitive to distortion.

From the plots it is seen that the vowel /i/ exhibits the least variation between the four types of distortion while /u/ exhibits the most. If we look at fig. 5, we see that for the vowel /u/ the spectral difference corresponding to the "disturbing threshold" (value 6) is over 20 dB for square-law distortion, but only about

10 dB for crossover distortion. For the other speaker's /u/, however, the characteristics of the four distortion type curves were different (variations were again large, but the order was different).

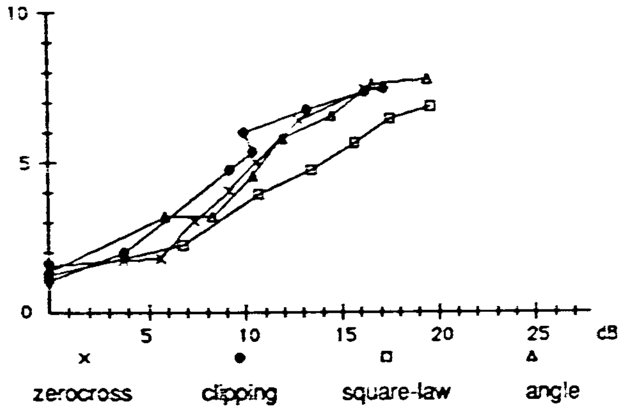


Fig. 3. Subjective distortion evaluation vs. maximal auditory spectral deviation. Vowel: / a /. Average from 15 evaluations for each point.

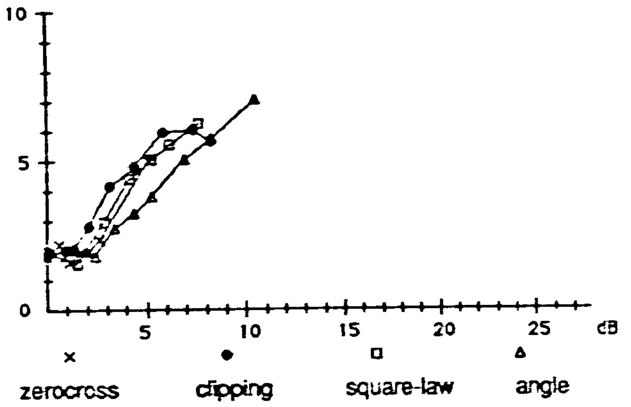


Fig. 4. Subjective distortion evaluation vs. maximal auditory spectral deviation. Vowel: / i /. Average from 15 evaluations for each point.

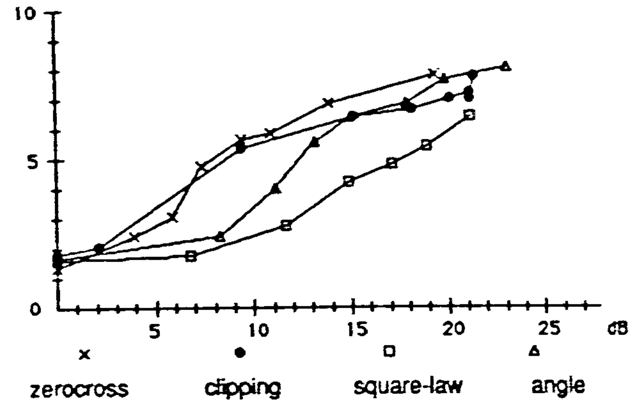


Fig. 5. Subjective distortion evaluation vs. maximal auditory spectral deviation. Vowel: / u /. Average from 15 evaluations for each point.

Considering the results we can say that although the auditory spectrum method is good at JND threshold, it has only moderately good correlation to subjective distortion evaluation at higher distortion levels. Therefore the method needs further refinements. Possible ways of doing this are: (1) to define a better distortion measure than maximal spectral deviation, and, (2) to improve the auditory model itself.

Improving the distortion measure

Some possible ways of changing the distortion measure are:

- Frequency weighting. The current measure handles all the 48 channels in the model equally, but it could be advantageous to give more weight to the highest channels, since high-frequency components are usually more disturbing than lower ones.
- Area and level weighting. The distortion measure could be made a function of the geometrical area of the spectral deviation, which would give a measure related to the total amount of distortion.

Changing the auditory model

Our model does not take into account what happens inside one pitch period of speech sound but rather only the long-term perception phenomena are considered. However, it is known that the temporal fine structure of sound has some effect on the perception. If the time constants of the model were shortened so that the fine structure of the signal would have an effect on the auditory spectra, this could give some extra information about the signal. In the case of distortion perception this information could be important: for example, if one distortion mechanism distorts only the peaks of the signal (say, clipping), it may have a different subjective effect than another type which has more effect on the low-level parts (crossover).

AUDITORY MODELLING APPROACH IN DISTORTION MEASUREMENT

Since the 2-dB rule is found to correlate well with distortion perception threshold, the auditory spectrum analysis can be used to measure distortion in audio and speech transmission equipment. This method enables the use of actual speech (or other sounds) as measurement signals. The results correspond to subjective sound quality better than results obtained with traditional methods like total harmonic distortion measurement.

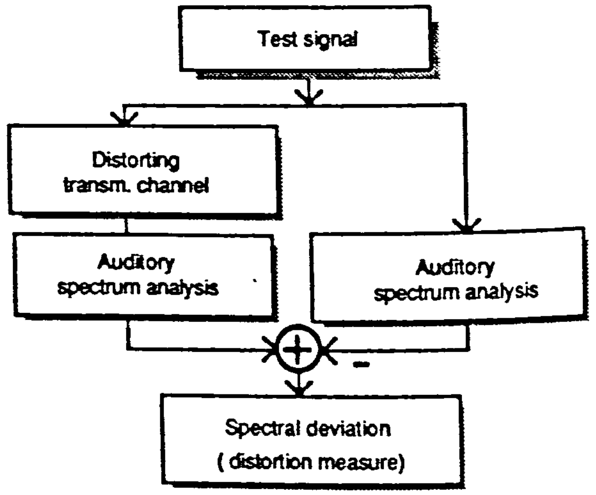


Fig. 6. Block diagram of auditory distortion measurement. By subtracting the auditory spectrum of the original test signal from the distorted signal we obtain the auditory spectral deviation, from which the distortion measure can be derived.

We have realized an auditory model based measuring system. The auditory model is implemented in a Texas Instruments TMS 32010 signal processor. An Apple Macintosh personal computer is used for system control and user interface, and a slightly modified Sony PCM-F1 pulse code modulator acts as the DA- and AD-converter. Figure 6 presents the nonlinear distortion measurement principle as a block diagram. Our system can handle the entire audio range (20Hz - 20 kHz) with a dynamic range of over 90 dB. The Posts and Telecommunications of Finland is testing the applicability of the method in telephone equipment measurements.

CONCLUSIONS

The auditory models have proven to be a useful means of determining perceived nonlinear distortion in speech. Already the relatively simple method of maximal spectral deviation is a good measure for the JND threshold (2-dB rule). More severe distortion levels need a more sophisticated measure. Practical applications of auditory methods are under development – possible areas are the evaluation of telephones and audio equipment as well as research systems for phonetic science.

ACKNOWLEDGEMENTS

The auditory model was developed in a project sponsored by the Academy of Finland. The study of applicability of the model to the measurement of distortion in speech transmission was financed by the Posts and Telecommunications of Finland. The implementation of the auditory model for the TMS 320 signal processor was realized by Vocom Ky.

REFERENCES

/1/ **Karjalainen M.**, Objective Measurements of Distortion in Speech Signal Channels by Computational Models of Speech Perception. Proc. of 11th ICA, Paris 1983.

/2/ **Karjalainen M.**, Sound Quality Measurements of Audio Systems Based on Models Of Auditory Perception. Proc. of IEEE ICASSP-84, San Diego 1984.

/3/ **Karjalainen M.**, A New Auditory Model for the Evaluation of Sound Quality of Audio Systems. Proc. of ICASSP-85, Tampa 1985.

/4/ **Karjalainen M., Helle S. & Altosaar T.**, Spectral Representations Based on an Auditory Model: Experiments and Applications. Proc. of Nordic Conference on Speech Processing, Trondheim, Norway 1986.

/5/ **Schröder M. et al.**, Objective Measure of Certain Speech Signal Deteriorations Based on Masking Properties of Human Auditory Perception. In the book: Frontiers of Speech Communication Research (ed. Lindblom & Öhman), Academic Press 1979.

/6/ **Zwicker E. & Feldtkeller R.**, Das Ohr als Nachrichten- empfänger. S. Hirzel Verlag, Stuttgart 1967.

A MODEL FOR THE PHONETIC MENTAL REPRESENTATION OF WORDS

WALTER F. SENDLMEIER

Max-Planck-Institut für Psycholinguistik
Nijmegen, The Netherlands

ABSTRACT

Seven psychological models of word recognition are analysed as to their explicit and implicit assumptions on the phonetic mental representation of words, and are then considered in the light of experimental results concerning the concept of the primary perceptual unit and findings from first language acquisition research. On the basis of these considerations a model for the phonetic mental representation of words is proposed which assumes simultaneous representation of differently sized units in the form of prototypes. The implications of this model for models of word recognition are discussed.

INTRODUCTION

Hardly any of the leading word recognition models contains explicit information on the phonetic mental representation of words. This may be seen as a serious drawback of these models considering that (phonetic) mental representation may not only be regarded as a result of the perception process, but that it functions at the same time as a monitor for perception. Almost all models, however, make more or less clear statements on primary perceptual units to which - at least implicitly - the status of mental representation is ascribed.

- Klatt /1/ assumes in his 'LAFS' (lexical-access-from-spectra) model that the listener is able to distinguish words directly by spectral analysis of the speech signal without having to segment it into smaller units. However, he also assumes that words have an internal structure which can best be described by units of diphone size. An important part of the word recognition process according to Klatt's model is the recognition of the internal diphone structure of a word by a listener. In this model words must thus be mentally represented as diphone sequences in the listener.
- In describing his 'logogen model' Morton /2/ gives the impression that he does not regard any segmentation within word boundaries necessary for the recognition process. Words are held to be represented as holistic entities.
- In the 'cohort model' /3/ it is assumed that words are represented as sequences of discrete units in the listener. The size of these units equals approximately that of single sounds, although statements on the linguistic status of the units and thus on their degree of abstractness (phoneme, allophone or phone) are avoided.
- Forster /4/ was the first to include specifications on the phonetic mental representation of words in his 'search model'. This model is based on the assumption that words in the lexicon are represented as sequences of phonological segements (phonemes).
- Pisoni, Nusbaum, Luce and Slowiaczek /5/ also make explicit statements on the mental representation of words in their 'phonetic refinement theory'. They believe that words are represented in the mental lexicon as sequences of discrete phonetic segments equalling single sounds which are defined in a multi-dimensional space /6/.
- Elman and McClelland /7/ assume that there are processing units of different sizes on different levels. These processing units are acoustic phonetic features, phonemes (allophones) and words. Even though Elman and McClelland assume interactions between these different units during the word recognition process, on closer examination of their 'trace model' these units appear to be hierarchically organized. Thus the question remains, whether the different units are simultaneously present in the sense of a mental representation or whether they have to be deduced one from another in a given sequence.
- Grosjean and Gee /8/ distinguish between units of processing and units of representation, but only make specific statements on the former. In their view, units of processing are the stressed syllable and the phonological word consisting of a stressed syllable and a number of unstressed syllables linked with the stressed syllable. Unfortunaltely, Grosjean and Gee do not specify how these units are related to potential units of mental representation. Considering the importance the authors ascribe to the function of prosodic features in the word recognition process, it seems feasible to deduce that they do not tend to assume that words are phonetically represented in form of sequences of discrete single sounds.

PRIMARY PERCEPTUAL UNITS

As mentioned above, the problem of phonetic mental representation of words is closely linked with the question of the basic (natural) units of speech perception. When, in the early fifties, experimental phoneticians and psychologists started to investigate the relation between the linguistic unit and its processing by the human listener, they were guided by the concept of minimal pairs and the ensuing distinctive feature theory developed by phonologists. Thus they focussed on the smallest isolated and reduced units - presented in form of synthesyzed signals to listeners in the laboratory who were asked to identify and discriminate them. Notwithstanding the valuable results obtained by such studies, one should be aware of the fact that the experiments were based on artificial acoustic phenomena which were as far distant as possible from their natural manifestations.

In criticizing the assumption of distinctive features as being psychologically real, in the beginning of the seventies an explicit discussion on the nature of the primary perceptual unit began. It was believed that in reaction time experiments, especially by target monitoring tasks, one could determine linguistic, taxonomically structured units according to their relevance as units in the speech perception process. One of the important results of these experiments is that the reaction times for short sentences, words, syllables and sounds are the same, if the search list consists of units of the same size as the target unit /9, 10, 11/. On the condition that reaction time experiments are an adequate means to reveal information on the primary perceptual unit, it can be deduced that units of different sizes may serve as primary perceptual units. In spite of such results a number of authors still argue for certain units to be the exclusive representatives of primary perception and try to prove their hypotheses by experimental studies /12, 13/.

RESULTS FROM FIRST LANGUAGE ACQUISITION RESEARCH

Another possibility of gaining insight into the phonetic mental representation of words lies in looking at the early stages of the child's language acquisition process. In first language acquisition research it has become quite an unquestioned fact that the child learns a word as bearing meaning corresponding to a certain object or class of objects. It seems plausible to assume that in this learning process the phonetic characteristics are globally perceived; in other words, the child learns the word 'ball', for example, as a phonetic unit and not as a combination of the single sounds /b/+/ɔ:/+/l/ or even as a matrix of 3x9 distinctive features.

Empirical results support this view: For example, Bruce /14/ found in investigations with 5- to 7 1/2-year-old children that during this stage in development holistic processing of words changes to more analytic processing. Liberman, Shankweiler, Fischer and Carter /15/ carried out experiments with 4- and 5-year-olds and found that these children could segment words much more easily into syllables than into single sounds. In using rhyming tests Magnusson, Naucler and Söderpalm /16/ found that preschool children were not able to give metalinguistic judgments on the basis of the phonetic-phonological structure of the words they heard. School children, however, were well able to do this, which may be accounted for by their ability to read and write. These findings, among others, point to the fact that at first the child perceives words phonetically in a global, non-analytic manner.

The prerequisites of a more analytic way of perceiving speech elements, in other words, the insight into the existence of certain recurring features, is only possible on the grounds of a substantial vocabulary. The possibility that an analytic recognition of words may occur in a more advanced stage in the process of cognitive development and that it may be furthered by special training is not questioned. But such perception of speech which analyses different speech signals within word boundaries may only follow global perception in the developmental sequence, and it cannot extinguish the earlier developed global way of perception.

To summarize, in this approach it is assumed that the child begins by recognizing words as global units. More analytic ways of speech perception may be used in later stages of language acquisition with interindividually varying degrees.

A MODEL OF THE MENTAL PHONETIC REPRESENTATION

These considerations lead to the following model of phonetic mental representation. The grown-up speaker/listener has stored a variety of mental representations on the phonetic level, the most important being: words, syllables, single sounds and phonetic features. Figure 1 illustrates the outlines of the model.

It should be noted that the different units are not localised on different levels of representation, but that they are different kinds of representation within one level, i.e. the phonetic level. These different kinds of representation are simultaneously

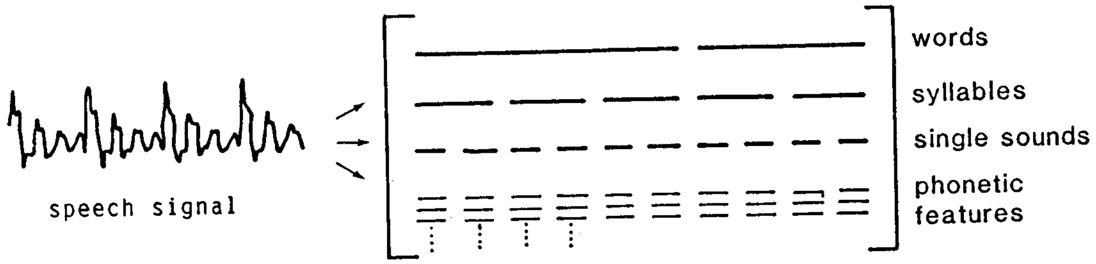


Fig. 1 : Different kinds of mental representation of words on the phonetic level which are simultaneously at the disposal of the listener; the listener focusses that kind of representation first which seems most efficient for word recognition.

at the disposal of the listener/speaker once he has established them. From which kind of representation the listener primarily takes the relevant information for solving a perception task is determined, for example, by the type of task, the context of perception, the speed and/or the complexity of the incoming stimuli etc.. Besides, it seems to make sense to assume that the perceptual activities of a listener vary not only with varying tasks, but that he may also interchangeably focus on different kinds of representation while solving one particular task, for example by recognizing a phrase or a sentence. Thus a listener can switch to single sounds or even phonetic features when discriminating difficult words such as proper names or words of a foreign language, and then he can switch back to words later.

Such a type of model in which a simultaneous representation of stimuli within different systems of similarity and contexts is postulated, is successfully being used in other psychological fields, as for example in the cognitive psychological research on problem solving; it has amply been shown that the flexibility in problem solving is based on the ability to change perspective /17/.

Since different listeners make different experiences in their perceptual surroundings, the degree of their ability to differentiate, i.e. the number of types of representation of a given word they have at their disposal, may differ from one individual to another. This is why the kind of representation on which listeners rely in a successful recognition process may also vary according to properties of the listeners themselves. For example, the knowledge of a phonetically oriented writing system (such as is acquired when learning to read and write an alphabetical writing system) may lead to a more differentiated organization of the mental representation of words. Morais, Cary, Alegria and Bertelson /18/ could in fact show that adult illiterates had much more difficulties in solving certain linguistic tasks involving detailed phonetic analyses than literate adults. What Morais et al. showed for speakers of Portugese, Sendlmeier /19/ could confirm also for native speakers of German. Within the scope of the introduced model these results may be explained in such a way that the adult illiterates have no concept of the single sound the way literates have. This, however, should not lead to the misinterpretation that the one group could listen better than the other. As a matter of fact, illiterates are just as able as literates to distinguish minimal phonetic differences in discrimination tasks, which, however, gives no clue as to the primarily focussed type of representation in the process of word recognition.

Closely related to the question in which size the phonetic perceptual units are represented is the problem of how these representations are present. Here Wertheimer's concept of 'ideal types' /20/ or Rosch's related concept of 'prototypes' /21/ seem to be adequate alternatives to abstract feature matrices.

The representation in form of prototypes is postulated for all kinds of representation of the phonetic level in the model. It seems plausible to assume that a listener generates a prototype from all the ever heard representatives of a category in the sense of a statistical mean during the course of language acquisition. If one supposes that phonetic units of different sizes (up to words) are represented analogously in form of typical prototypes, but not in the sense of a first degree isomorphy, this implies an enormous capacity of the long term memory. Objections by scientists who by referring to - up to now uncertain - principles of economy argue against such a supposition of storage-consuming representation can be rejected in view of an almost unlimited capacity of the human brain /22/. The material basis of an analogous representation in form of prototypes may be seen in neurophysiological correlates of spectral patterns, since it may be taken for certain that the incoming soundwave is subjected to a frequency analysis by the peripheral hearing system.

CONSEQUENCES FOR WORD RECOGNITION MODELS

The presented model of mental representation contains a number of constraints on the process of word recognition. This is due to the fact that structure and process mutually depend on each other. It is up to word recognition models to delineate the rules and mechanisms that characterize the different types of strategies in speech perception. However, in doing so the following facts should not be ignored:

- Word stress patterns are normally used in word retrieval; words seem to be organized in the lexicon according to stress contours /23, 24, 25/.
- Linguistic differences can cause listeners with different languages to develop different perceptual strategies /26/.
- Configurational (prosodic) features of words often hinder the listener from focussing on single sounds in recognizing words /27/.
- Unstressed function words usually are recognized some time after their off-set, in most cases only after taking into account the following stressed syllable /8/.
- The size of the phonetic units used by listeners varies with the complexity of the words in similarity judgments /28/.
- The size of the primary perceptual unit varies with the size of the respective context /29/.

Word recognition models which assume only one kind of primary perceptual unit - phonetic features, single sounds, syllables or words - are confronted with a number of problems when trying to explain findings like the ones listed above. It seems that only such models will be of lasting importance which start from the assumption that the listener has active control over the process of auditory word recognition and that he can focus at will on any kind of representation that seems useful for successful word recognition.

REFERENCES

/1/ D.H. Klatt, "Speech understanding systems and speech perception theory", Journal of Phonetics 7, 1979, 279-312.

/2/ J. Morton, "Word recognition", in J. Morton, J. C.Marshall (eds.), Psycholinguistics 2: Structures and processes.Cambridge, 1979, 107-156.

/3/ W. Marslen-Wilson, A. Welsh, "Processing interactions and lexical access during word recognition in continuous speech", Cognitive Pschology 10, 1978, 29-63.

/4/ K.I. Forster, "Levels of processing and the structure of the language processor", in W.E. Cooper, E.C.T. Walker (eds), Sentence processing: psycholinguistic studies presented to Merrill Garrett. Hillsdale 1979, 27-86.

/5/ D. Pisoni, H. Nusbaum, P. Luce, L.M. Slowiaczek, "Speech perception,word recognition and the structure of the lexicon", Speech Communication 4, 1985, 75-95.

/6/ M. Treisman, "Space or lexicon? The word frequency effect and the error response frequency effect", Journal of Verbal Learning and Verbal Behavior 17, 1978, 37-59.

/7/ J.L. Elman, J.L. McClelland, "An architecture for parallel processing in speech recognition: the trace model", in M.R. Schroeder (ed.), Speech and Speaker Recognition, Basel, 1985, 6-35.

/8/ F. Grosjean, J.P. Gee, "Another view of spoken word recognition", Cognition 15, 1987, in press.

/9/ H.B. Savin, T.G. Bever, "The nonperceptual reality of the phoneme", Journal of Verbal Learning and Verbal Behavior 9, 1970, 295-302.

/10/ D.J. Foss, P.A. Swinney, "On the psychological reality of the phoneme: perception, identification and consciousness", Journal of Verbal Learning and Verbal Behavior 12, 1973, 246-257.

/11/ D. McNeill, L. Lindig, "The perceptual reality of phonemes, syllables, words and senteces", Journal of Verbal Learning and Verbal Behavior 12, 1973, 419-430.

/12/ D. Norris, A. Cutler, "Juncture detection", Linguistics 23, 1985, 689-705.

/13/ J. Mehler, J.Y. Dommergues, U. Frauenfelder, J. Segui, "The syllables role in speech segmentation", Journal of Verbal Learning and Verbal Behavior 20, 1981, 298-305.

/14/ D. J. Bruce, "The analysis of word sounds by young children", Journal of Educational Psychology 34, 1964, 158-159.

/15/ I.Y. Liberman, D. Shankweiler, F.W. Fischer, B. Carter, "Explicit syllable and phoneme segmentation in the young child", Journal of Experimental Child Psychology 18, 1974, 201-212.

/16/ E. Magnusson, K. Naucler, E.Söderpalm, "Form or substance? The linguistic awareness of preschool children and school children investigated by means of a rhyming test", Abstracts of the 10th International Congress of Phonetic Sciences, Dordrecht, 1983, 633.

/17/ W. Prinz, "Wahrnehmung und Tätigkeitssteuerung", Berlin, 1983

/18/ J. Morais, L. Cary, J. Alegria, P. Bertelson, "Does awareness of speech as a sequence of phones arises spontaneously?" Cognition 7, 1979, 323-331.

/19/ W.F. Sendlmeier, "Psychophonetische Aspekte der Wortwahrnehmung, Hamburg, 1985.

/20/ M. Wertheimer,"Untersuchungen zur Lehre von der Gestalt, II", Psychologische Forschung 4, 1923, 301-350.

/21/ E. Rosch, "Cognitive Reference points", Cognitive Psychology 7, 1975, 532-547.

/22/ W. Penfield, "Consciousness, memory and man's conditioned reflexes", in K.H. Pribram (ed.), On the biology of learning, New York, 1969.

/23/ R. Brown, D.McNeill, "The 'tip of the tongue' phenomenon", Journal of Verbal Learning and Verbal Behavior 5, 1966, 325-337.

/24/ D. Fay, A. Cutler, "Malapropisms and the structure of the mental lexicon", Linguistic Inquiry 8, 1977, 505-520

/25/ E. Engdahl, "Word stress as an organizing principle of the lexicon", in D. Farkas (ed.), Papers from the parasession on the lexicon. Chicago, 1978, 138-147

/26/ A. Cutler, J. Mehler, D. Norris, J. Segui, "A language specific comprehension strategy", Nature 304, 1983, 159-160.

/27/ A.G. Samuel, W.H. Ressler, "Attention within auditory word perception: insights from the phonemic restoration illusion", Journal of Experimental Psychology: Human Perception and Performance 12,1986, 70-79.

/28/ W. F. Sendlmeier, "Auditive judgements of word similarities",Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 40, 1987, in press.

/29/ H. Fujisaki, K. Hirose, H. Udagawa, N. Kanedera, "A new approach to continuous speech recognition based on considerations on human processes of speech perception", ICASSP, IEEE 86, 1986, 1959-1962.

ACKNOWLEDGEMENT

I thank Prof. H. Fujisaki for fruitful discussions on the topic of mental phonetic representation.

LINGUISTIC FACTORS IN SPEECH PERCEPTION

A.S. STERN

Dept. of Phonetics
Leningrad State Univ.
Leningrad, USSR, 199034

ABSTRACT

The hypothesis tested in this research is that certain linguistic characteristics have a material influence on speech perception. A statistical model based on analysis of variance in perceptual data is proposed, where significant factors are assumed to be the perception cues and their levels to be decision making units. The investigation of the model has enabled us to elucidate a number of psycholinguistic features of the speech perception process, the typological properties of a given language as well as some characteristics of perceptive ability development in both native language acquisition and second-language learning.

HYPOTHESIS, METHODS, MATERIAL

In the present work the perception of cardinal psycholinguistic units, i.e., syllables, words, sentences and texts, was studied. Listening to speech stimuli was chosen as an experimental procedure, since it seems to be a perceptual activity that is mainly dependent on the processing of sound sequences and is not closely related to the higher levels of speech comprehension. A group of 7-10 subjects was asked to listen to sets of speech stimuli presented against the background of some distortion and to write them down. The texts were presented several times, while other stimuli only once. Different kinds of distortions or their combinations were used: a) objective distortions (white noise, distant reception, synthetic speech stimuli, accented speech), and b) subjective (poor hearing, poor knowledge of the language, aphasia). The quantitative aspect of distortion namely, the signal/noise ratio (S/N), the degree of hearing loss, the level of performance in the second language, etc. was also varied.
Each speech segment can be described on the basis of its correct perception frequency. Besides, one may obtain a number of ratings for various linguistic features. For example, the word "ruka" (hand) is a noun (a level of the factor Parts of Speech), with the highest possible frequency of occurrence (a level of factor F_{ob}), containing the stressed "a", bisyllabic, etc. Correct recognition of the word "ruka" is assumed to be determined by these factors, or more precisely, by their levels. Hence, it is quite natural to use analysis of variance to discover the significant linguistic features (factors) and to establish a hierarchy among them. Results of this analysis have yielded a statistical descriptive model of speech segment perception.
Let us consider a fragment of such a model, giving the correlation ratio η_x^2 of some factors in word recognition: 1 - against the background of white noise at S/N = -6dB; 2 a,b - in hard of hearing adults with different degrees of hearing loss; 3 - for German students who perceived Russian words in white noise at S/N = = -2dB. The significant factors are underlined (see the Table).

Table

Experiments / Factors	1	2a	2b	3
Stressed Vowel	0.052	0.020	0.020	0.006
Voiced/Voiceless	0.000	0.007	0.002	0.005
Soft/hard	0.017	0.015	0.004	0.009
Length in Syllables	0.073	0.010	0.006	0.013
Parts of Speech	0.018	0.040	0.030	0.094
F_{ob}	0.012	0.003	0.002	0.043

For correct use of analysis of variance, the factors being investigated in the experimental material should be orthogonal. In most cases balanced articulatory tables were used /2/.
Our conclusions are based on the analysis of about 50 experiments, giving approximately 70,000 responses. These experiments were conducted, in part, in collaboration with my collegues. The study of the models obtained has made it possible to discuss three groups of problems.

I. THE PSYCHOLINGUISTIC FACTORS IN SPEECH PERCEPTION

A. Isomorphism of Models for Speech Unit Perception at Different Linguistic Levels in Auditory Listening Tests.

This is confirmed, first, by the fact that the models for speech unit perception at all linguistic levels are shown to be analogous, and, secondly, the same factors hold for units of different levels. For example, the factors Stressed Vowel and some Distinctive features of consonants are significant for both syllable and word recognition. Thus, a certain isomorphism of linguistic levels in the process of listening may be postulated. It should be noted that the obtained factors act simultaneously in every instance and no "input" can be found into a set of this type.

B. Similarity in Mechanisms of Perception Irrespective of the Distortion Type.

Each type of distortion is characterized by an individual set of factors or a hierarchy of these factors. There are, however, factors which turn out to be significant in the majority of cases. Among them we find the following: relative frequency of occurrence and length in syllables for words, the stressed vowel, parts of speech. To conclude, it should be mentioned that there is an evident similarity in the mechanisms of speech perception under different conditions of distortion, which not only justifies the accepted approach towards speech pathology, insufficient knowledge of the language and noise as a distortion, no matter what its nature may be, but also helps to understand every single case on the basis of distortions of other types.

C. Differences in Mechanism Depending on the Degree of Distortion.

When the type of distortion is constant but the degree is altered not only common but specific factors as well are revealed, besides, their ranks may vary. For example, F_{ob} of speech units (syllables or words) is found to be one of the most important factors in poor reception conditions and to decrease in significance as the reception conditions improve. The factor Parts of Speech is insignificant under poor reception conditions whereas under superior conditions it becomes a factor of great value. Thus, it can be said that the analysis revealed both common and specific features. The first of these two findings, i.e. the existence of common features, was not unexpected. The second one, on the other hand, is difficult to predict and, therefore, is mostly ignored by researchers. In order to sum up the results of this section and of the preceding one, let us underline that the common features in mechanisms of perception are at work in all types of distortion, whereas specific features depend on the degree of distortion.

D. An Extension of Jakobson's Regression Hypothesis.

Let us now look at the data from a different angle. R. Jakobson proposed a hypothesis according to which aphasic speech disorders mirror the process of language acquisition in children. The data on the factor levels indicate the following: vovels are better recognized than consonants, /a/ is much more easily recognized than /ɛ/; choreic words are easier than iambic ones; nominative case is better perceived than other cases; the direct object is superior to the indirect object in the number of correct responses. The active construction is recognized more easily than the passive one. The dialogue is easier to perceive than the monologue, words of frequent occurrence are recognized correctly more often than rare words. Is is clear that the first members of the oppositions are acquired earlier in the ontogenesis. We can therefore attempt to extend Jacobson's hypothesis in the following way: the linguistic features which are the earliest to have been acquired are the most stable in all types of distortion.

E. The Existence of Simple and Complex Factors Functioning as One Whole.

Some of the factors are simple and cannot be further disintegrated into other features (i.e. distinctive features of the phonemes or parts of speech). Other factors, such as syllabic contrast or communicative type of text, may be conceived as a combination of more elementary features. But in the process of speech perception these complex features may become crucial, that is, they function as a whole. An increase in the weight of such complex features is often caused by an improvement in reception conditions. This fact is in agreement with some recent psychological investigations.

F. Differences in the Perception of Isolated Units and Units in Context.

Comparison of sets of significant factors for isolated words and words included in a text indicates that some of them are present in both test conditions. For most factors, however, a decrease in significance or a complete loss of significance is observed. Thus, the mechanism of perception is different for isolated words and words in context.

G. Simultaneous Perception of Speech Unit as a Whole and in Elements.

Some factors are related to elements into which the speech units can be subdivided (e.g. stressed vowels), whereas the others describe the unit as a whole (e.g. the rhythmic structure, frequency of occurrence). Since both types are significant simultaneously, one may suppose that the recognition of the whole unit and that of its parts occurs parallelly. Let us consider some additional facts. If we compare the hierarchies of all factors for words and syllables under similar conditions we can clearly see that for S/N = = -6dB rank test ρ is +0.86, for OdB it is +0.60, and at +4dB it is +0.09. These data indicate that under poor reception conditions the mechanism of phonetic processing of a word is highly efficient which is not the case under good reception conditions. In another experiment listeners were given words spoken by non-native speakers of Russian (the Agul) and parts of these words pronounced with a strong accent. It was found that ρ (rank test) for the correct recognition of words and their parts in 4 different groups of listeners varied from -0.10 to +0.17, that is, there was actually no correlation at all. This signifies that words were perceived regardless of the presence of some distorted segments, i.e. as whole units.
Moreover, when German students recognized Russian words both masked and not masked by noise, correct recognition scores in the latter case were twice as high as in the former case. This improvement was due to perception of both familiar and infamiliar words. Thus, a possibility of phonemic decoding has been demonstrated. Now we can amend the rule as follows: speech units are perceived simultaneously as sequences of elements and as integral units (Gestalt), the strategy depending on the perceptual situation.

H. Simultaneous Involvement of All Linguistic Levels Regardless of the Type of the Unit to be Perceived.

To make this item clear, let us take our data on words. Word perception is determined by the following factors: certain distinctive features of consonants and vowels (the sound level), length of words in syllables (the syllabic level), part of speech and length in morphemes (morphemic level), the number of quasiomonymes (word level) and F_{ob} (the text level). This indicates that various linguistic levels are involved in the perception of speech units at the same time.

I. Speech Perception as an Action.

It is generally considered that the probability prediction is based on the fact that the listener is an active recipient of speech. Our experiments have confirmed the significance of the probability factor. Thus, the greater the probability of a word or syllable, the higher the correct recognition scores. An additional experiment has shown, however, that this mechanism is closely related to the frequency distribution in a sample,i.e. when frequencies of elements correspond to their linguistic probabilities this dependence is the lowest. Conversely, when the elements are equally distributed the direct dependence is higher.When the distribution is reverse, i.e., when elements with high probabilities occur rarely and vice versa, the dependence is also higher, but the correlation will have an opposite sign ("-") indicating that high probability elements are harder to recognize than low probability ones. Thus, the active character of perceptual processes is revealed in an interplay of the listener's sociolinguistic experience and the current analysis of frequency distributions in a given sample. The listener's activity is also revealed in series of choices he has to make: of a perceptual (phonetic) base from those he has at his disposal; of a morpheme from a corresponding morphemic class; of a word from a set of similar words, etc. All this applies only to speech units (from sounds to words) presented in isolation. In a text,however, the role of this factor considerably decreases. On the other hand, a key word prediction factor emerges, whose activity is linked with the work of association mechanism.

II. THE PSYCHOLINGUISTIC TYPOLOGY OF LANGUAGES.

Comparison of significant factors for a number of languages, namely,Russian, German, English and French enabled us to obtain both universal and language specific factors. F_{ob} and Parts of Speech are examples of universal factors.Specific factors for the Russian language are the location of the word stress and word order. The former is non-existent in French while the latter in German. The word-length factor may serve as another example. In Russian, the word length in syllables is quite significant whereas word length in morphemes is of less value (η^2_x is 2 times less). In German the situation is the reverse,word length in syllables being completely insignificant and word length in morphemes is in the forefront of significant factors. This latter fact is evidently connected with the greater "syntaxicality" of the German word. A projected analysis of other languages will help to establish a typology of languages at the perceptual level.

III. THE FORMATION OF THE PERCEPTUAL MECHANISM IN SPEECH ACQUISITION AND IN SECOND-LANGUAGE LEARNING.

A. A comparison of speech perception mechanisms in normal adults against the background of white noise,in hard-of-hearing adults, in normal children listening to speech in white noise and in hard-of-hearing children has shown that there was a +0.11 and a +0.14 rank correlation between adults and children for the same distortion type, and ρ = +0.50 between the two groups of children as well as the two groups of adults. This indicates that speech perception is determined by the age of the listener. It is especially important for children.

B. The Sets of factors and their hierarchy change in the course of second language learning, the degree of similarity with the native language mechanism decreases as that of the second language increases. For example, in the group of German students that participated in recognition tests of Russian words in white noise in their 1st, 3rd and 5th years at the university, ρ varied as follows:
0.40 → 0.28 → 0.18 as compared to the mechanism in German and 0.45→0.42→0.71 as compared to that in Russian.
On the basis of the above presented data it may be concluded that significant linguistic factors are perceptual cues (in the sense of the word introduced by S.Vygotsky and A.A.Leontyev), reflecting the elementary psychological operations of the speech perception processes.Moreover, the investigation suggests that the significance (and the maximum η^2_x) cannot be obtained unless an adequate way is found of determining factor levels (see the example on word length in Russian and German given above). The listener is assumed to make use of linguistic factors "keeping in mind" a particular level of factors. Hence, levels of linguistic factors are decision-making units.

Acknowledgements

The author wishes to thank Prof.L.R.Zinder for helpful comments on earlier versions of this report and his interset in all stages of the research. Thanks are due to my colleagues who took part in various experiments of the study. I also thank I.Panasjuk for looking through the English version of the text and making useful suggestions.

REFERENCES

1. L. Zinder, A. Stern. Factors Affecting Word Recognition. - Recent Trends in Soviet Psycholinguistics. New York, 1977-1978, p. 123-130.
2. A. Stern. Articulatory Tables for the Development of Perceptual Skills and Testing the Auditory Function. Leningrad State University Press, 1984 (in Russian).

PERCEPTION OF TONAL PATTERNS IN SPEECH: IMPLICATIONS FOR MODELS OF SPEECH PERCEPTION

DAVID HOUSE

Department of Linguistics and Phonetics
Lund University
Sweden

ABSTRACT

This paper advances a model of pitch perception in speech in which spectral changes influence the analysis of the tonal contour. This interrelationship is examined in view of certain linguistic requirements of tonal contours in the perception of spoken language. It is concluded that the perception of tonal movements is optimized when these movements occur in regions of spectral stability, that movement at the syllable level can be perceived directly as linguistic categories and that movement at the phrase level can be reconstructed from tonal levels stored in short-term memory.

INTRODUCTION

Intonation provides listeners with important information which facilitates the perception of spoken language (1). In this paper the word intonation will be used in a wide sense, that of perceptually significant changes in fundamental frequency which have a linguistic function. The purpose of this paper is to examine how these changes and their relationships to spectral changes can be represented in the peripheral auditory system and in short-term memory, and how this representation can be used to aid and guide the speech perception process.

Information obtained from Fo movement can be greatly varied and can function on several different levels simultaneously. The type of information dealt with here concerns linguistic categories such as relative syllable importance (stress), relative word importance (focus), language specific information at the word level (word accents and tones), phrase boundaries (juncture) and connective patterns over a longer time domain (grouping). Some of the principles involved in Fo-movement perception might, however, also be applicable to other types of information such as emotions, involvement, etc.

Raw Fo movement must be transformed by the perceptual mechanism into relevant tonal categories. This transformation presupposes an analysis of frequency (pitch), direction of movement (rising, falling) and range of movement. Current psychoacoustic and physiological models of pitch perception are generally in agreement that some degree of central processing is involved, but it is still unclear as to what extent pitch analysis interacts with spectral resolution (2,3). Pitch perception in spoken language involves the additional problem of coping with rapidly changing spectral cues and a pitch contour broken up by voiceless segments. This leads to a key question. Is pitch analysis continuous, following Fo without being influenced by breaks and spectral events, or is it more selective and economical using critical portions of movement which are then stored in short-term memory and retained for decisions involving larger time domains? On the basis of two perception experiments, this paper advances a model which takes the latter view.

PERCEPTION OF TONAL MOVEMENT AT THE SYLLABLE LEVEL

The first experiment was designed to test the influence of rapid spectral changes on the categorization of simple rise-fall and fall-rise tonal patterns at the syllable level. In this experiment, the categories were not linguistic ones but rather were presented to the listeners in the form of an ABX test design (4).

A Klatt software synthesizer and a VAX digital computer were used to synthesize a Swedish /ɑ/ vowel with formant frequencies of 600, 925, 2540 and 3320 Hz. (5,6). Vowel duration was 300 ms including 30 ms intensity onset and offset. Fundamental frequency was systematically varied to create 18 different stimuli. The Fo contour for stimulus A, designed to elicit rise-fall categories, rose from 120 Hz to a turning point of 180 Hz and then fell to an end point of 100 Hz. The Fo contour for stimulus B, designed to elicit fall-rise categories, began at 120 Hz falling to 80 Hz and then rose to 160 Hz. The difference in end-point frequency was designed to test the effect of end-point variation on the rise-fall, fall-rise categories, i.e. movement pattern versus discrete frequency analysis. The 18 stimuli were constructed by systematically varying the turning point in steps of 20 Hz from 80 Hz to 180 Hz with three different end-point configurations: 100 Hz, 160 Hz and 120 Hz. The beginning point was always 120 Hz. Listeners consistently categorized these stimuli on the basis of movement pattern and did not use end-point frequency.

To test the effects of rapid spectral changes on the categorization, three more versions of the test were made by introducing a gap, consisting of an intensity drop preceded and followed by formant transitions for /b/, into the first part, the middle part, and the final part of the vowel respectively. Figure 1 illustrates the Fo contours of the stimuli with the gap in the first part of the vowel.

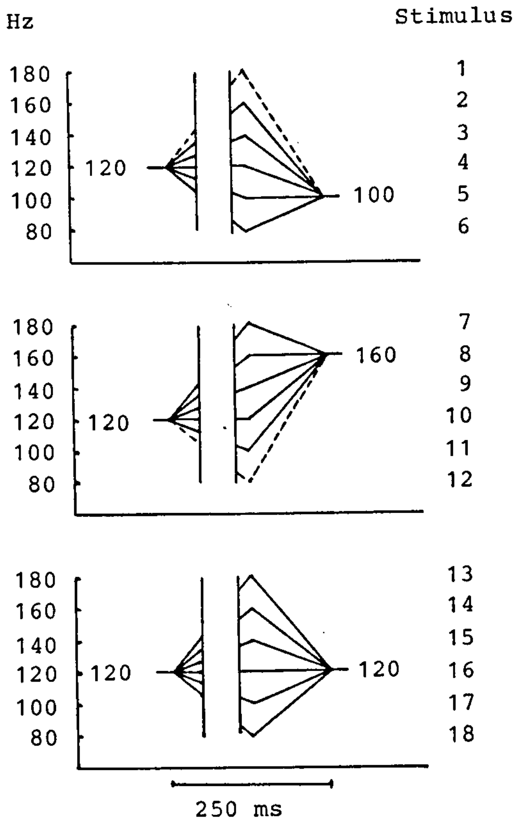


Figure 1.
Stylized tonal contours of one version of the ABX test. The dashed lines (stimuli 1 and 12) were also stimuli A and B.

Although a few listeners continued to categorize the new stimuli on the basis of tonal movement, most of the listeners´ responses were altered by the intrusion of the spectral changes. When the intrusions were placed in the middle and in the last part of the vowel, categorization was more strongly based on end-point frequency. When the intrusions were placed in the beginning of the vowel, the categorizations were reversed vis-a-vis the end-point frequency but corresponded to the average frequency 40-80 ms after the intrusion.

These results seem to indicate that tonal movement is optimally perceived during portions of high spectral stability. If the perceptual load is increased by rapid spectral changes, and the duration of spectral stability is decreased, tonal movement will then be perceived and stored as tone levels. This interpretation also complies with the results obtained by Gårding, et al. (7) where perception of tone 4 (falling) in Standard Chinese was altered to tone 3 (dipping) by moving the fall backwards in time toward the CV boundary and also by increasing the steepness of the fall. These manipulations were done by means of LPC synthesis.

Languages, then, which need to manifest rising and falling Fo at the syllable level should optimally place these movements in places of spectral stability. This corresponds to Bruce´s (8) production and perception data for Swedish concerning the timing of the word accent fall in non-focal position, where accent II is marked by a strong falling Fo well within the stressed vowel. This interpretation also has explanatory power concerning production data reported by Lindau (9) for Hausa (a two-tone language) where tonal turning points occur at the end of the vowel, a high being manifested as a rise and a low being manifested as a fall.

PERCEPTION OF TONAL MOVEMENT AT THE PHRASE LEVEL

The second experiment concerns perception of phrase boundary markers and connective patterns (10,11,12). Listeners were presented with sequences of five fives (55555) and asked to judge whether the sequence was grouped 55-555 or 555-55. The fundamental frequency of a natural Scanian fem (five) was manipulated in various ways using LPC synthesis. Variations comprised fall-rise and rise-fall patterns at different frequency levels as well as rising and falling patterns having different ranges. These variations were then joined together to create the sequences. Duration was not variable as each syllable was equal in length as were the intervals between them. 36 different sequences were used as stimuli.

The results clearly showed that listeners can use a rising or a falling Fo movement having a greater range than in the surrounding syllables as a demarcative cue signalling the end of a group. The results also indicated that listeners can rely on connective Fo movement patterns encompassing the entire group. Examples of such patterns are the "hat-like" and "trough-like" intonation patterns (13). The perception of such patterns implies the use of some type of short-term memory where Fo movement is stored (either as movement patterns or as frequency levels) to be retrieved when the entire group has been heard.

Another example from the material where the use of memory seems to be important is found where listeners interpret precisely the same falling syllable in the same position (the second "five") in two different ways depending on the surrounding Fo movement. In one case the falling Fo movement of the syllable is interpreted as the end of a "hat-like" pattern signalling the end of a two-syllable group. In the other instance, the same falling Fo movement is followed by a greater fall to a lower frequency. This causes the second syllable to be interpreted as the middle "five" of a three syllable group (Figure 2).

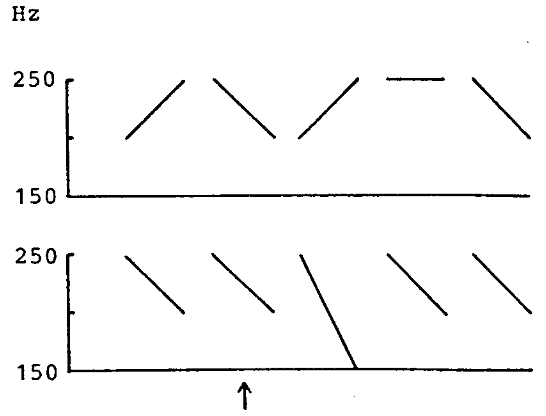


Figure 2.
Stylized tonal contours of two 55555 stimuli showing how the same falling syllable was interpreted in two ways. The top stimulus was interpreted as 55-555 and the bottom one as 555-55.

IMPLICATIONS FOR SPEECH PERCEPTION MODELS

When constructing a model of speech perception which takes into consideration fundamental frequency movement, pitch analysis is generally viewed as presupposing a first-order frequency analysis of the speech wave based on the the mechanical properties of the basalar membrane and characteristic frequencies and temporal responses of auditory-nerve fibers. This analysis provides the raw materials for a second-order analysis of pitch and timbre (14). On the basis of the data reported here, I would like to tentatively propose two different mechanisms of second-order pitch perception. The first is a direct conversion of Fo movement into linguistic categories. The second is a reconstuction of tonal movements or levels from short-term memory.

The categories of stress, word accents and tones, and in certain cases focus are likely candidates for the direct conversion of Fo movement. This movement, optimally located in the vocalic segments, is not then stored as movement, but rather as the corresponding linguistic category. This type of direct perception can be seen as corresponding to an event approach to segmental perception as proposed by Fowler (15). The rapidly perceived stressed syllables, for example, marked by tonal movement, can serve to guide perception to important areas of meaning (16).

Candidates for short-term memory based pitch analysis are juncture cues for boundaries, connective patterns for grouping and in certain cases focus. In this type of analysis, pitch could be stored first as tonal levels and then transformed into linguistic categories. Figure 3 presents a schematic diagram of the two different perceptual mechanisms.

Where the perception of intonation is seen as an important part of speech perception, the proposed division of movement perception into two mechanisms could have implications for more general models of speech perception. Although this division is tentative and speculative, it is an attempt to understand pitch perception in a linguistic frame of reference.

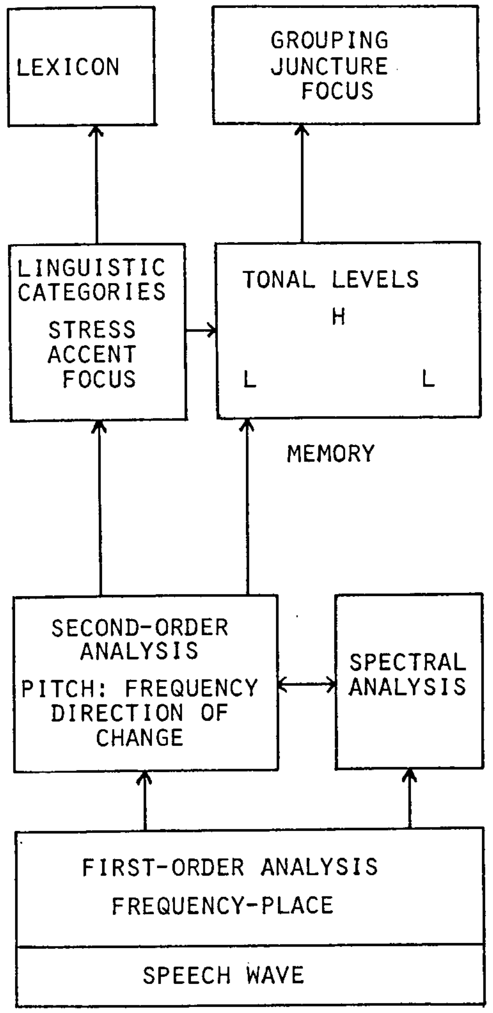


Figure 3.
Diagram illustrating two different perceptual mechanisms for pitch movement perception.

REFERENCES

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

(2) Plomp, R. 1976. Aspects of Tone Sensation, A Psychophysical Study. Academic Press, London.

(3) Itoh, K. 1986. A neuro-synaptic model of auditory memory and pitch perception. Annual Bulletin 20, Research Institute of Logopedics and Phoniatrics, University of Tokyo.

(4) House, D. 1985. Implications of rapid spectral changes on the categorization of tonal patterns in speech perception. Working Papers 28, Department of Linguistics and Phonetics, Lund University.

(5) Klatt, D.H. 1980. Software for a formant synthesizer. J. Acoust. Soc. Am. 67, 971-995.

(6) Fant, G. 1973. Speech sounds and features. The MIT Press. Cambridge, Mass.

(7) Gårding, E., Kratochvil, P., Svantesson, J.O., & Zhang, 1985. Tone 4 and Tone 3. Discrimination in Modern Standard Chinese. Working Papers 28, Department of Linguistics and Phonetics, Lund University

(8) Bruce, G. 1977. Swedish word accents in sentence perspective. Travaux de l'Institut de Linguistique de Lund XII. Gleerups, Lund.

(9) Lindau, M. 1986. Testing a model of intonation in a tone language. J. Acoust. Soc. Am. 80, 757-764.

(10) Gårding, E. & House, D. 1985. Frasintonation, särskilt i svenska In Svenskans beskrivning 15, eds. S. Allén et al. Göteborgs universitet, Göteborg: 205-221.

(11) Gårding, E., & House, D. 1986. Production and perception of phrases in some Nordic dialects. Working Papers 29, Department of Linguistics and Phonetics, Lund University.

(12) House, D. & Gårding, E. 1986. Phrasing in some Nordic Dialects. Paper presented at the fourth Nordic Prosody Conference in Middlefart, Denmark. In preparation.

(13) Collier, R. & t´Hart, J. 1975. The role of intonation in speech perception. In Structure and Process in Speech Perception. (Eds) Cohen, A. & Nooteboom, S.G. Berlin.

(14) Gelfand, S.A. 1981. Hearing, an introduction to psychological and physiological acoustics. Marcel Dekker, Inc. New York and Basel, Butterworths, London.

(15) Fowler, C.A. 1986. An event approach to the study of speech perception from a direct-realist perspective, Journal of Phonetics, 14, 3-28.

(16) Bannert, R. 1986. From prominent syllables to a skeleton of meaning: a model of prosodically guided speech recognition. Working Papers 29, Department of Linguistics and Phonetics, Lund University.

MICROPROSODY IN SEGMENT PERCEPTION

KLAUS J. KOHLER

Institut für Phonetik und
digitale Sprachverarbeitung
Universität Kiel
2300 Kiel, FRG

ABSTRACT

For German it has been demonstrated in a number of experiments that in production as well as in perception a level and a level + falling F0 contour on a prestop vowel are cues for fortis and lenis stop, respectively. This paper reports on perception experiments that replicate the German findings for English, and relates the results to an interaction of three factors: (a) prestop microprosody, (b) poststop microporosody, (c) global utterance macroprosody.

INTRODUCTION

The importance of F0 after stop release as an acoustic cue for the lenis/fortis categorization of stop consonants has been known for a long time /1/. F0 preceding the stop closure, on the other hand, has not been attributed a similar cue value. For German it has been demonstrated in a number of experiments with the utterances "Diese Gruppe kann ich nicht leiden/leiten." ("I cannot stand/lead this group.") that in production as well as in perception a level and a level + falling F0 contour on the prestop vowel are cues for /t/ and /d/, respectively /2/. These results have been only partially replicated for English in the utterances "I am telling you I said widen/whiten." with very much smaller effects /3/. This difference was related to the fuzziness of the segment boundary in /w/ + /ae/ as against /l/ + /ae/ and to the fact that long initial formant transitions have been found to increase the perceived duration of a following vowel. To test this hypothesis, three perception experiments were carried out. In the first one, the previous German test was repeated (a) with another German group in order to demonstrate the generalizability of the discovered signal/perception link for German, (b) with a group of British English speakers in order to show up any perceptual differences due to language background, and to establish a base-line for the other two experiments, which (1) replicated the segmental chain and the F0 patterns of the German test items (/'laedn/ - /'laetn/) in an English sentence frame, and (2) compared its results with those for /'waedn/ - /'waetn/.

EXPERIMENT 1

Procedure.
The test tape of experiment 2 of /2/ was presented to a group of 16 native speakers of German (students of phonetics and languages), in several subgroups, via a loudspeaker in a sound-treated room of the Kiel Phonetics Institute. They classified the stimulus utterances as "leiden" or "leiten" sentences by ticking the appropriate boxes on prepared answer sheets. Two groups of 6 and 7 British English speakers performed the same test under the same conditions, but they gave their answers by pressing one of two buttons at the recording stations of a reaction-time measurement system. They were students of German spending 6 months in Kiel to improve their proficiency in the language.

Results.
The German group replicates the results of the previous test (cf. /2/, pp. 24ff) in every respect (see figure 1). The two English groups, which do not differ from each other and are, therefore, combined in the data presentation of figure 2, also show clearly separate identification functions for level and falling F0. But they have a higher percentage of /d/ responses in the middle of the duration ratio range for both level and continuously falling F0, and the response curves for falling and level + falling F0, which are already close together in the data of the German group, coalesce in this upward shift of two of the identification functions. This means that the English subjects show the same perceptual effects with regard to level F0 as against the other two F0 patterns, but that they nevertheless locate the duration ratio boundary at a lower value than the German listeners. The reason fo this difference may be that because English speakers generally devoice the nasal plosion after fortis stops, the absence of this feature in the German test stimuli biases English listeners towards /d/ in the middle of the duration ratio range.

EXPERIMENT 2

Procedure.
Two English sentences were constructed that replicate the focal and utterance-final position as well as the segmental structure and the phonetic context of the German test words in Experiment 1. The two family names "Lyden" and "Lighton", which are of equal (low) frequency in Britain, were inserted in the sentence frame "I think you'd have to ask ..." They contain the same phoneme sequences as the German words and can also be realised with nasal plosion. They, too, occur after a voiceless consonant cluster that interrupts the F0 glide from a low value on "ask" to a high one in the contrastively stressed name so that F0 has practically reached its peak value when it sets in again at voiced /l/ onset.

These sentences were pronounced several times by a native speaker of Southern British, with focus stress on the name, elicited by the context "Who do you think would know about this, Lyden or Lighton?" The F0 contours across the names were very similar to those found in the German sentences of Experiment 1 (cf. /2/, p. 24): before the lenis stop F0 drops much further in the stressed vowel than before fortis. One token of a "Lyden" sentence was selected for the test stimulus generation, which followed the principles laid down in /2/. The stressed vowel measured 289 ms, its closure duration 46 ms and its stop release 24 ms.

Three F0 patterns were generated across the stressed vowel: (a) Level + falling (122-120-75 Hz) with the fall beginning at the vowel center, (b) level (122-120), (c) linearly falling throughout (122-75 Hz). These F0 contours were combined with 7 rate-manipulated vowel durations, from 260 ms down to 200 ms in 10-ms steps. The closure voicing and release were excised and replaced by silence, which was increased from 70 ms up to 160 ms in 6 equal steps, complementary to the vowel shortening. The 21 vowels produced in this manner, together with the complementary closure pauses, were spliced into the carrier utterance. Thus the durations and F0 patterns of the resulting 21 "Lyden/Lighton" stimuli were fully comparable to those generated in the German test, the only difference being that after the silence F0 set in at 70 Hz (instead of 66 Hz) and that the periodicity of the nasal was more regular and of much greater amplitude than in the German "leiden/leiten" stimuli, i.e. there was proper and strong voicing instead of creak.

Since the frame was not synthesized, the stimuli sounded completely natural, and no "synthetic" quality was detectable in the synthesized vowel sections either. The 21 stimuli were copied ten times and randomized to give a test of 210 stimuli, following the same procedure as in the German test. The same two groups of native British English speakers as in Experiment 1 acted as informants under the same listening conditions in separate sessions. They classified the stimulus utterances as "Lyden" or "Lighton".

Results and discussion..
The two groups differ in their responses to the level F0 stimuli, one giving more /d/ judgements. Figure 3 presents the combined group results. They are basically congruent with the English group results of Experiment 1: the identification curves occupy more or less the same positions along the duration ratio axis, the functions for the two falling F0 sets are again not differentiated from each other, but are clearly separate from the function for level F0, which yields significantly more /t/ responses. The differences between the two experiments are (a) somewhat more /d/ judgements in the lower half of the duration ratio scale for Experiment 2, and (b) different as against identical behaviour of the two groups in the two experiments. So there must be some essential acoustic difference between the English "Lyden/Lighton" and the German "leiden/leiten" stimuli. The obvious candidate is the strong voicing instead of creak in the final nasal of the English utterances. It provides a more promiment release cue for /d/, which may enter into conflict with the fortis cues and weaken their effects, i.e. the effect of flat F0 generally and the effect of duration in the lower range. This conflict can be solved differently, according to whether the release is weighted more highly, especially than flat F0. The two groups differ in this respect.

EXPERIMENT 3

Procedure.
The sentences "I am telling you I said widen/whiten." were pronounced several times with focus stress on the final word and with nasal plosion by the same native Southern British speaker that produced the utterances for Experiment 2. One "widen" token was selected for constructing 21 test stimuli according to the same principles as in Experiments 1 and 2. The vowel durations ranged from 265 ms to 205 ms, the silence durations from 70 to 160 ms. Again 3 F0 patterns were generated with each vowel duration. In the level + falling F0 pattern the level section was represented by the naturally produced fluctuation between 119 and 123 Hz over the first 100 ms of the original vowel,

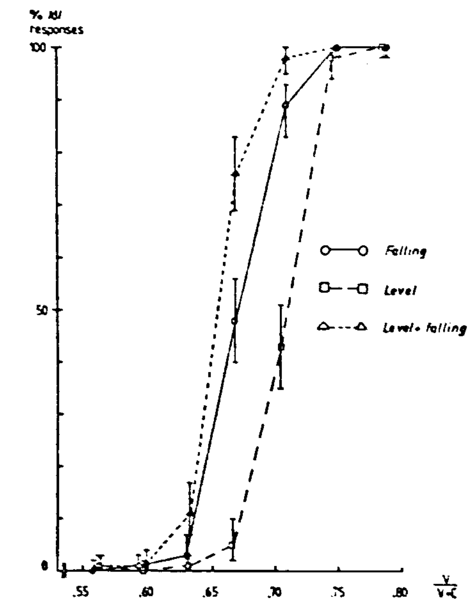


Fig. 1. Percentage /d/ responses as a function of vowel/(vowel + closure) duration ratio for the 3 F0 conditions in Experiment 1 ("leiden/leiten", German group), and binomial confidence ranges at the 5 % level; 16 listeners. At each data point N = 160.

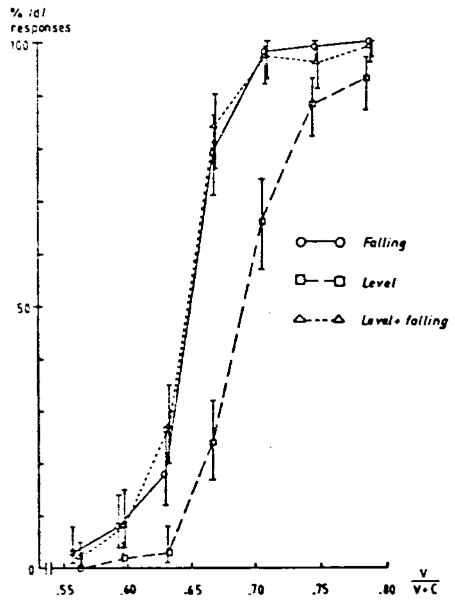


Fig. 2. Responses of the combined British English groups in Experiment 1. At each data point N = 130.

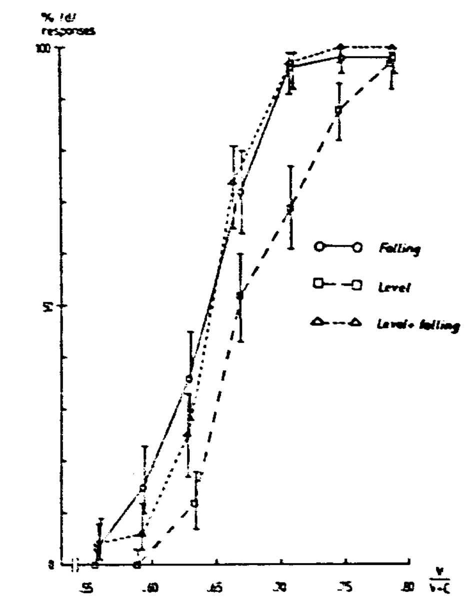


Fig.3. Responses of the combined British English groups in Experiment 2 ("Lyden/ Lighton"). At each data point N = 130.

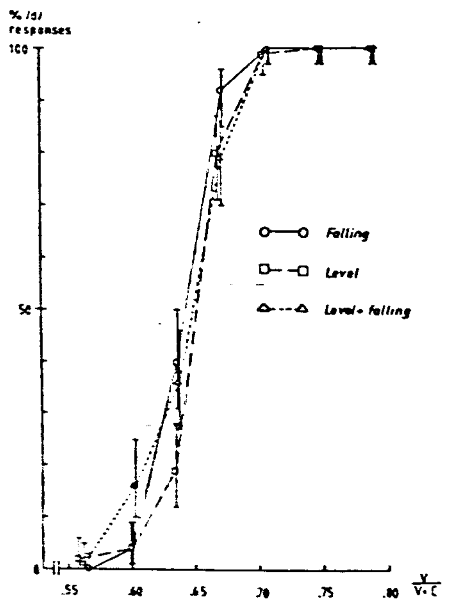


Fig. 4. Responses of the combined British English groups in Experiment 3 ("widen/ whiten"). At each data point N = 110.

followed by a linear fall to 85 Hz, the proportion of level and slope sections staying the same in all 7 stimuli. The first 100 ms of the level F0 were identical with the level section of the level + falling pattern in the longest vowel and changed proportionally with the vowel duration; the remainder descended to 122 Hz. In the third pattern, F0 fell linearly throughout from 119 to 85 Hz.

The original /d/ release was again eliminated, and the 21 synthesized vowels + closure pauses were spliced into the sentence frame. F0 at voice onset of the final nasal was 89 Hz, descending to 69 Hz. The very large amplitude of the regular periodicity in /n/ was adjusted to the one found in "Lyden" by applying the reduction factor .35. The durations and the F0 patterns were comparable to the ones in the test stimuli of Experiments 1 and 2, but with important differences in the height of the pre- and postconsonantal F0 ending and starting points.

The test tape construction and the running of the experiment followed the same lines as in Experiment 2. A previous run of the test was reported in /3/. It was repeated here by the same two British English groups as in Experiments 1 and 2. In a pretest, each of the 13 subjects was examined as to whether they distinguished "wh" from "w". Two informants did and were, therefore, excluded from the test because their expectations for "whiten" would have been different.

Results and discussion.

Figure 4 provides the data for the combined group. There are no inter-group divergencies: The differences between the three F0 patterns have practically disappeared. The effect of flat F0, which was still slightly present in the previous run of the same test, has been levelled out. Otherwise the two test runs provide corresponding locations of the identification functions. Since it is only the response curve for flat F0 that is positioned differently in the "Lyden/Lighton" and the "widen/whiten" data, the initial consonant /w/ cannot be responsible for the increase of /d/ judgements. It must be an acoustic feature difference that is peculiar to the flat F0 stimuli. In "Lyden/Lighton", F0 is flat across the stressed syllable, and a rise from the preceding syllable is masked by voicelessness; after the closure silence, F0 resumes at its low utterance-final value. The flat F0 contour is thus bounded by voiceless stretches on both sides, with low F0 preceding and following. In this environment, the high flat F0, i.e. the fortis cue, becomes perceptually salient. In "widen/whiten", on the other hand, there is an upward F0 glide from the low value of the preceding syllable right into the stressed vowel, and it is only the final 130 - 160 ms that are actually flat. After the closure pause, there is a substantial F0 fall of 20 Hz. In this context, the high flat F0 is integrated into a macroprosodic rise-fall pattern and is, therefore, perceptually far less salient, thus losing its fortis cue strength.

GENERAL DISCUSSION

The results of the 3 experiments point to the following prosodic influences on lenis/fortis stop perception in German and English.

1. A flat F0 across a stressed prestop vowel in a focused utterance-final disyllable is a fortis cue, compared with falling F0 patterns, in both German and English, as long as the flat F0 is clearly detachable from a macroprosodic utterance intonation as a microprosodic manifestation. In German, a flat + falling F0 is also differentiated from a continuously falling F0 as a stronger lenis cue.

2. In English, the category boundary between lenis and fortis is located at lower duration ratios. This leads to a coalescence of the identification functions for flat + falling and continuously falling.

3. A stop release with regular voicing of high amplitude and an F0 fall (below the focus peak) weakens the preconsonantal microprosodic fortis cue.

4. The microprosodic effects of prestop flat and flat + falling F0 are obliterated when they are integrated into macroprosodic utterance pitch patterns.

5. The interaction of pre- and poststop microprosody and of global utterance macroprosody explains why a prestop F0 influence on lenis/fortis perception can only arise under special circumstances and, therefore, not provide a basis for tonogenesis (cf. /1/).

REFERENCES

/1/ J.M. Hombert, J.J. Ohala, and W.G. Ewan, "Phonetic explanations for the development of tones", Language 55, pp. 37-58, 1979.

/2/ K.J. Kohler, "F0 in the perception of lenis and fortis plosives", J. Acoust. Soc. Am. 78, pp. 21-32, 1985.

/3/ K.J. Kohler, "Preplosive F0 in the perception of /d/-/t/ in English", Proc. Montreal Symposium on Speech Recognition, pp. 34-35, 1986.

COMPONENTS OF PROSODIC EFFECTS IN SPEECH RECOGNITION

ANNE CUTLER

MRC Applied Psychology Unit
15 Chaucer Rd.
Cambridge CB2 2EF, U.K.

ABSTRACT

Previous research has shown that listeners use the prosodic structure of utterances in a predictive fashion in sentence comprehension, to direct attention to accented words. Acoustically identical words spliced into sentence contexts are responded to differently if the prosodic structure of the context is varied: when the preceding prosody indicates that the word will be accented, responses are faster than when the preceding prosody is inconsistent with accent occurring on that word. In the present series of experiments speech hybridisation techniques were first used to interchange the timing patterns within pairs of prosodic variants of utterances, independently of the pitch and intensity contours. The time-adjusted utterances could then serve as a basis for the orthogonal manipulation of the three prosodic dimensions of pitch, intensity and rhythm. The overall pattern of results showed that when listeners use prosody to predict accent location, they do not simply rely on a single prosodic dimension, but exploit the interaction between pitch, intensity and rhythm.

Speakers place accent on the most important words in an utterance. Thus by finding accented words, listeners can efficiently locate the most central parts of a speaker's message. Previous studies have shown that listeners do indeed actively use sentence prosody to tell them where accented words are going to occur. Cutler [2] produced pairs of sentences varying in prosodic contour. An example is (1):

(1) (a) The couple had quarrelled over
a BOOK they had read.
(b) The couple had quarrelled over
a book they hadn't even READ.

Upper case represents sentence accent. In (1a) the main sentence accent falls on *book*, in (1b) on *read*. These sentences were used as materials in a phoneme-monitoring experiment, in which listeners are asked to respond as quickly as possible to the presence of a specified word-initial phoneme. In (1), the target phoneme is /b/, so the target-bearing word is *book*. Targets on accented words are responded to faster than targets on unaccented words in this task. In Cutler's experiment, the target-bearing word itself was actually spliced out of both sentence contexts and replaced in each by identical copies of a neutral rendition of the same word. The result of this manipulation was a pair of sentences with acoustically identical target-bearing words, which were preceded by identical sequences of words; the only difference between the members of each pair was the prosody applied to the words preceding the target. In one case the prosodic contour in which the target-bearing word occurred was consistent with accent falling upon that word; in the other, it was consistent with the target-bearing word being unaccented. Under these conditions, the 'accented' targets still elicited faster responses than the 'unaccented' targets, and since the only relevant differences between the two sentences in each pair lay in the prosody, Cutler concluded that listeners must have used cues in the prosody to direct their attention to the location where sentence accent would fall.

Prosody, however, is not a unitary phenomenon. The separate dimensions of rhythm, pitch and intensity all contribute to the prosodic structure of an utterance. Cutler's experiment did not examine *how* listeners were exploiting prosody to predict accent, or whether any one prosodic dimension was more informative than others.

Cutler and Darwin [3] subsequently found that removing pitch information - i.e. monotonising the sentences - did not remove the accent effect; in monotonised spliced sentences like (1) the 'accented' targets are still responded to significantly faster than the 'unaccented' targets.

From this, Cutler and Darwin concluded that pitch information could not be a necessary component of the accent prediction effect. They speculated that no prosodic dimension might prove necessary for listeners to predict upcoming accents, but variation in any prosodic dimension might prove sufficient.

In the present studies, the three prosodic dimensions of pitch, rhythm and intensity are separately manipulated in an attempt to analyse the accent effect in further detail. Unlike the study by Cutler and Darwin, which simply removed the dimension of pitch by setting it to a single value across each utterance, the present studies investigate the effects of the separate prosodic dimensions when they are *interchanged* between the two members of a sentence pair. To begin with, using dynamic time-warping techniques in a system developed by Jeffrey Bloom at the Polytechnic of Central London [1], we exchanged the rhythmic patterns within each pair of sentences (for examples like [1], where naturally different contours were produced by having a slight variation in the text at the end of the sentences, the rhythmic patterns were exchanged up to the point at which the two members of the pair diverged). Thus (1a), for example, was given the rhythm of (1b) but retained its original pitch and intensity contours; (1b) had the rhythm of (1a) but its own pitch and intensity patterns.

In Experiment 1, phoneme-monitoring response times were measured in these rhythmically manipulated sentences, and in the same sentences with intact prosody. The intact sentences were LPC-analysed and resynthesised to control for acoustic effects of resynthesis. The words bearing the target were acoustically identical in all four sentences belonging to a set such as (1).

There were 20 such sentence sets. Forty listeners, in four groups of ten, took part in the experiment. Each group heard only one sentence from each set, and the two variables of 'accented' versus 'unaccented' targets, and intact versus rhythmically manipulated prosody, were counterbalanced across subject groups.

Subjects were tested individually. Response times, measured from a click (inaudible to the subjects) aligned with target onset, were collected by a microcomputer using programs developed by Norris [4]. After the experiment subjects were given a short recognition test, and their response times were analysed only if they scored at least two-thirds correct on this test.

The results of this experiment are shown in Fig. 1. The intact sentences, in which rhythm, pitch and intensity contours are preserved from the original utterance, show the advantage of 'accented' over 'unaccented' targets which was found in the earlier experiments. This indicates that the resynthesis alone is not interfering with listeners' ability to use prosodic contours to predict the location of accent. The difference in this condition is significant (F1(1,36) = 21.36, p <.001). In the rhythmically manipulated sentences, however, the advantage of originally accented over originally unaccented targets is less than half as large as the difference in the prosodically intact sentences, and it is not statistically significant (F1(1,36) = 3.55, p >.05).

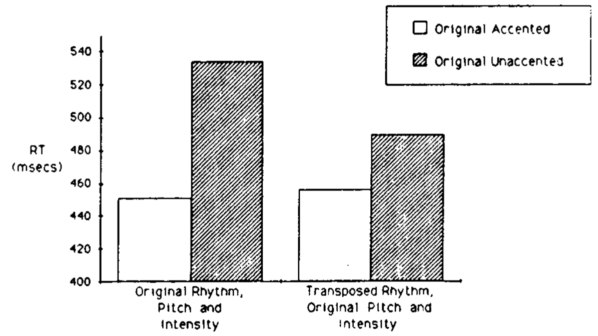


FIG. 1. Phoneme-monitoring response time (msecs.), Experiment 1.

This experiment shows that the rhythmic manipulations have severely affected the accent effect. Each of the utterances which had undergone this rhythmic manipulation had an unnatural, indeed a conflicting, prosodic structure - pitch and intensity contours signalled one prosodic pattern while the rhythm signalled another. It is clear that listeners did not base their prosodic processing on one aspect of the prosodic contour alone.

One possible interpretation of this result is that listeners are simultaneously processing all three prosodic dimensions, and that the separate contributions of each prosodic dimension to the predicted accent effect are simply additive. The attenuated, but still positive, effect in the rhythmically manipulated sentences would, on this simple story, be attributable to the combination of positive effects contributed by the pitch and intensity contours, set against a negative effect contributed by the rhythmic contour.

This interpretation was tested in Experiment 2. This experiment investigated prosodic manipulations which were the reverse of those in Experiment 1. The pitch and intensity contours were transposed between originally accented-target and unaccented-target members of a sentence pair, leaving the rhythmic contour, alone, intact.

This manipulation was possible because the time-warping applied to the sentences in Experiment 1 produced pitch and intensity contours which, although they preserved the contour shape from the utterance they had originally belonged to, were aligned with the rhythmic pattern of that utterance's pair. Therefore these contours could simply be transposed onto that pair. These transpositions were realised using prosodic editing routines devised by Kim Silverman.

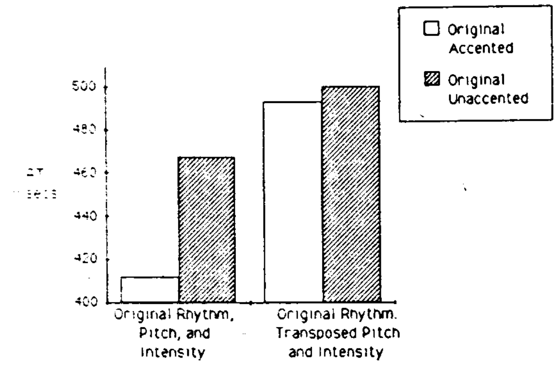


FIG. 2. Phoneme-monitoring response time (msecs.), Experiment 2.

Experiment 2, like Experiment 1, included the resynthesised utterances with intact prosody; these were compared with the utterances in which of the original prosody only the rhythm was preserved intact, the pitch and intensity contours being transposed between members of a pair. Again, the target-bearing words were acoustically identical in all sentences from any set.

Forty listeners, who had not taken part in Experiment 1, were tested; design and procedure were as in Experiment 1. The results are shown in Fig. 2.

It can be seen that once again the utterances with intact prosody showed a strong accent effect, i.e. response time advantage for 'accented' over 'unaccented' targets. This difference was statistically significant (F1(1,36) = 6.85, p <.02). In the utterances with transposed pitch and intensity contours, there was virtually no response time difference between originally accented and originally unaccented targets (F1 <1).

The results of this experiment rule out the very simple explanation of Experiment 1 offered above. Had listeners been simply evaluating all three dimensions of prosody in an additive fashion, we might have expected the reverse of the result found in Experiment 1 - that is, we might have expected an advantage of originally unaccented targets over originally accented targets of about half the magnitude of the difference in the opposite direction produced by the prosodically intact utterances. However, the conflicting prosody in this case wiped out any difference in response times as a function of original accent location.

This result raises the possibility that transposition of prosodic contours might itself interfere with listeners' ability to predict accent location by extracting relevant information from the prosody. In order to rule out this possibility, a further experiment was conducted in which all three prosodic dimensions were transposed.

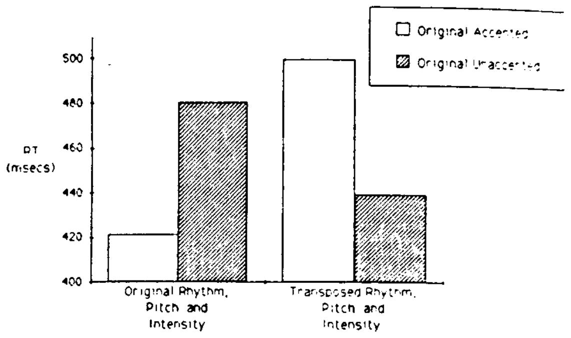


FIG. 3. Phoneme-monitoring response time (msecs.), Experiment 3.

In Experiment 3, the resynthesised utterances with intact prosody were again tested, and compared in this case with utterances in which rhythm, pitch and intensity contours had all been transposed between members of a sentence pair. The manipulated utterances in this experiment therefore exhibited the maximum of transposition, in that every utterance had rhythm, pitch *and* intensity contours which had originally been applied to another utterance. However, they exhibited the minimum of prosodic conflict, since rhythm, pitch and intensity contours were always in accord.

As in the previous experiments, the target-bearing words were acoustically identical in all sentences from any set.

Forty listeners, none of whom had taken part in Experiments 1 and 2, were tested. Design and procedure were as in the preceding experiments. The results are shown in Fig. 3.

Once again there was a significant advantage for 'accented' over 'unaccented' targets in the prosodically intact sentences (F1(1,36) = 10.38, p <.005). Moreover, there was a significant difference in the reverse direction, i.e. a response time advantage of originally unaccented over originally accented targets, in the prosodically manipulated sentences (F1(1,36) = 6.83, p <.02). That is, when all three components of the prosodic contour signalled that accent would occur at the position where the target occurred, the target was responded to faster; and this was true whether the consistent prosody was applied to its original utterance or to its original utterance's pair.

This result allows us to dispose of the suggestion that prosodic transposition might interfere with listeners' prosodic processing. Instead, it is clear that what interfered most strongly with listeners' prosodic processing in the two preceding experiments was prosodic conflict. When one prosodic dimension was in conflict with the other two, listeners were unable to arrive at a consistent interpretation based on prosodic information. One effect of this was that significant accent effects disappeared.

However, the results from the prosodically manipulated conditions in Experiments 1 and 2, though they were both statistically insignificant, seem to differ. This might suggest that more sensitive experimentation could yet uncover differential contributions to the accent effect on the part of rhythm, pitch and intensity respectively. For the present, though, we may conclude with confidence that listeners' processing of prosody is not simply an additive evaluation of separate dimensions; the interaction between prosodic dimensions is of paramount importance. When the three dimensions rhythm, pitch and intensity agree, listeners exploit them efficiently and consistently. When they conflict, this exploitation is significantly impaired.

ACKNOWLEDGEMENTS

This research was supported by a grant from British Telecom to A. Cutler and D.G. Norris. Acknowledgement is made to the Director of Research of British Telecom for permission to publish this paper. The author is grateful to Jeffrey Bloom and Kim Silverman, without whose prosodic manipulation techniques these experiments would not have been possible, as well as to Steve Bartram, Sally Butterfield, John Williams and John Culling for technical assistance.

REFERENCES

[1] Bloom, P.J. Use of dynamic programming for automatic synchronization of two similar speech signals. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* 1984, 2.6.1-2.6.4.

[2] Cutler, A. Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics, 20,* 1976, 55-60.

[3] Cutler, A. & Darwin, C.J. Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. *Perception & Psychophysics, 29,* 1981, 217-224.

[4] Norris, D.G. A computer-based programmable tachistoscope for non-programmers. *Behavior Research Methods, Instrumentation and Computers, 16,* 1984, 25-27.

HIGH FREQUENCY SPEECH PERCEPTION: PHONETIC ASPECTS AND APPLICATION

MÁRIA GÓSY

Department of Phonetics, Linguistics Institute of HAS
Budapest 1250 Pf. 19. Hungary

ABSTRACT

The so-called speech frequencies (100-3000 Hz) seem to be both necessary and sufficient for perception and understanding. The role of speech elements occuring above 3000 Hz is unclear. They might be totally unnecessary, on the one hand or, on the other, they might have a secondary acoustic cue function which is demonstrated experimentally by removing the lower frequencies. Experiments were carried out with Hungarian native listeners both with normal and with impaired hearing. The results are given in detail.

INTRODUCTION

The so-called speech frequencies seem to be both necessary and sufficient for the perception of vowels and consonants. The acoustic information in this frequency range is generally suitable for understanding running speech. However, a lot of comprehension problems arise if only these frequencies can be used. This can be demonstrated with the telephone where general conversation can easily be carried out without any problems in understanding. However, identification of names or comprehension of suddenly changed topic of dialogue can cause difficulty. It is known that people with hearing loss at high frequencies (above 3000 Hz) suffer from perceptual and understanding difficulty. There is no doubt that the first two energy maximums, the formants, contain the main information for the identification of vowels and certain consonants. Moreover, components of some other consonants - like [s] or [t͡s] - occurring below 3000 Hz are sufficient for their identification. The role of high frequencies (above 3000 Hz) in perception, however, has been little investigated [1]. The acoustic information contained in the high frequencies may be purely supplementary; alternatively it may play an independent and special role in perception. To bring this problem a little closer to a solution, experiments were carried out with Hungarian-speaking native listeners.

METHOD AND MATERIAL

The material used consisted of (i) 25 sound-sequences without meaning and (ii) 102 monosyllabic, phonetically balanced Hungarian words. The bisyllabic sound-sequences contain almost all Hungarian speech sounds. The acoustic structure of part of them corresponds to Hungarian phonotactic rules while that of another part of them contradicts them. All the words consist of three sounds: a vowel between two consonants. The words range from well-known ones, in everyday use, to ones very rarely used. They belong to different grammatical categories. Attempts were made to choose booth the sound-sequences and words containing consonants and vowels in different phonetic positions and in different environments. The speech material was recorded by a male announcer who pronounced it as isolated statements in random order. The recording was made with a professional tape recorder and microphone under laboratory conditions. An 8 s pause was left between the sound-sequences/words. The intensity level of sound-sequences and the words varied within ± 6 dB. Two types of filtration method were used for testing: pass-band and high-pass filtering by an Audio Filter. The filter slope was always 36 dB/octave. The cut-off-frequencies were for words 2200, 2700, 3300, 3900 Hz and 2200-2700, 2700-3300, 3300-3900, 3900-4700 Hz; for sound-sequences 2200, 2700 Hz and 2200-2700, 2700-3300 Hz. These values were chosen in view of the fact that the highest acoustic cue for Hungarian vowels appears in general to be about 2200 Hz; it is the second formant for the [i] sound. There are 8 different materials for the words and 4 for sound-sequences. In order to examine the role of the upper frequencies, those below 2200 Hz were removed. The frequency analyses were made of filtered material by the Sound Spectrograph (Type 700 of Voice Identification). Each of the 12 test materials was administered to 10 adult normal-hearing subjects, totally 120 subjects, half of them females and half males. The experiments were conducted in a silent room. The listeners' task was to write down the sound-sequences or words they could perceive/understand. In order to obtain statistically significant results, we used our own Psychotest program.

RESULTS AND DISCUSSION

The experimental data for sound-sequences and for words are summarized in Table 1. These show that (i) the perception/understanding of sound-sequences/words was better under pass-band filtering than under high-pass filtering; (ii) perception/understanding decreased under high-pass filtering according to the change of the cut-off-frequency; (iii) a frequency band seems to occur with the highest perception and understanding ratio: 2200-2700 Hz. The differences between the filtered groups proved to be significant at the .01 level.

Table 1

Cut-off-frequencies of filtering (Hz)	Correct identification (%)	
	words	sound-seq.
2200 h.p.	67	49
2200-2700 p.b.	98	78
2700 h.p.	72.5	35
2700-3300 p.b.	95	75
3300 h.p.	74	
3300-3900 p.b.	95	
3900 h.p.	46	
3900-4700 p.b.	95	

The abbreviations mean high-pass and pass-band filtering.

These results led us to the conclusion that there are frequency bands in which more acoustic information about the same word/sound-sequence seems to disturbing to the decoding processes [2]. The supposed idea is that the upper part of the acoustic structure of certain speech sounds does not remain characteristic for them when the lower part is lost. In other words: these high frequencies do not contain unambigous information about the sounds or cannot be acoustic cues used for identification. The components appearing at these frequencies have been thought to play a supplementary role in recognition. Results obtained from examinations using the low-pass filtration method confirm this [3]. If this were the case, the high elements would have been redundant. Our new results have not confirm this assumption, and, indeed, they seem to contradict it. The data have supported an alternative hypothesis, namely that certain speech sounds and sound combinations have special 'cue-like' components above 2200 Hz. This 'secondary-cue' hypothesis was further investigated by means of spectrographic analyses. These showed that, as expected, the main difference in acoustic structure between pass-band and high-pass filtered groups lies in the presence or absence of the higher frequencies.

By way of illustration let us look at the bilabial nasal consonant [m]. The original acoustic structure of [m] contains cues at about 500 and 1500 Hz. In the absence of these frequencies it is not possible to identify [m] without elements above 2000 Hz. Spectrographic analysis of [m] shows further components at about 2800 and 3700 Hz. The word mos 'washes' was understood accurately when frequencies below 2000 Hz were removed by filtration. When the component at 2800 Hz was reduced in intensity by further filtration, identification of the consonant became impossible (Fig. 1).

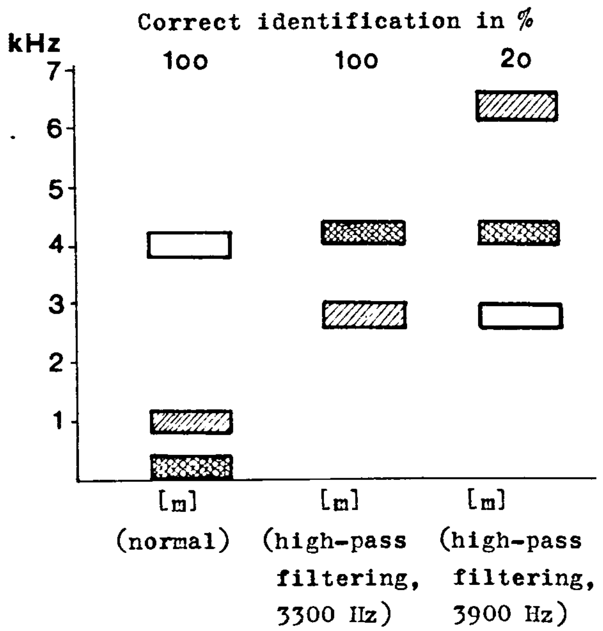


Figure 1. Consonant [m] in word mos 'washes' and its identification in %

One of the questions to be asked in this respect is: why can we not used the whole information of the high frequencies in perception, why do they seem to cause difficulty? Moreover, why do these perceptual problems disappear when there are only frequency bands? This suggests that, in contrast to the main acoustic cues (below 2200 Hz), the secondary cues act alone and independently of the disturbing higher components. As to the explanation of 'disturbing higher components' let us document it with an example. The perception of the Hungarian long [a:] vowel was analyzed. The correct identification of this sound in the sound-sequence tádó [ta:do:] after high-pass filtering is 0% and after pass-band filtering is 100% (with the cut-off-frequencies of 2200 and 2200-2700 Hz). The false responses in the first case were: [tøldu, tødu, tøldo:, tɔdu]. Instead of [a:] dominantly [ø] was perceived. In the acoustical structure of [a:] there are components between 2000 and 4000 Hz with very different intensities. This is assumed to cause the perceptual differencies. The spectrographic analyses show, however, an important difference depending upon the context (sound combination). The [a:] mentioned above was perceived more correctly when it occurred between fricative or fricative and nasal consonants. This can be explained by the transition phases which act as acoustic cues in this respect.

Finally the role of meaning should be taken into consideration. If we compare the frequency of use and the correct identification percentages of the test words, it seems clear that meaning generally does not play as important a role in this case as is supposed in literature. There are frequent words with low understanding ratio, and rarely used words with high percentage values. There are a lot of items which cannot be used in isolation in Hungarian, e.g. pác 'pickle' (70%) and zöm 'bulk' (20%). There are words with similar meaning or frequency and their understanding is quite different; and words with similar acoustic structure and different percentage values. The grammatical category of the words seems to be of lesser importance as well.

By way of final conclusion the following idea will be presented. All the results have supported that perception and understanding are better in certain high frequency bands especially in 2200-2700 Hz. This finding led us to the hypothesis that hearing-impaired people with special hearing losses can perceive/understand speech in the 2200-2700 Hz range better than in a wider band which also contains the 'disturbing' elements.

A supplementary experiment was carried out with the participation of 10 hearing-impaired adults having hearing losses of different types and extents. Table 2 shows the responses of a mixed-type hearing-impaired woman for the words with their normal acoustic structure and after pass-band filtering.

Table 2

Original words	Responses of a hearing-impaired adult	
	normal sounding	after pass-band filtering (2200-2700 Hz)
moʃ	moʃ	moʃ
kør	kol	kør
meɲ:	-	meɟ
la:b	ɔd	la:b
ʒeb	ʒe	ʒeb
hi:d	hi:g	hi:d
si:n	sẹ:p	si:n
fɔl	-	fɔl

The results confirm that the secondary acoustic cues can, indeed, ensure the perception/understanding of speech in case the normal decoding process cannot work because of hearing problems.

What criteria should the high frequency components fulfil in order to act as acoustic cues? (i) Identification should reach a significant level and, (ii) frequency values should be defined for correct perception. On the basis of our data it can be supposed that the components appearing in certain frequency bands correspond to the above-mentioned expectations.

Further research should show how these findings can be applied in audiological examinations, phoniatric work and in speech therapy.

REFERENCES

[1] Fant, G.: Speech Sounds and Features. Cambridge, Massachusetts, and London 1973.

[2] Rosen, S. - Fourcin, A.J.: When less is more - Further work. Speech Hearing and Language No. 1. 1983, 1-27.

[3] Gósy, M.: Magyar beszédhangok felismerése, a kísérleti eredmények gyakorlati alkalmazása./Identification of Hungarian speech sounds, the application of experimental results. Hungarian Papers in Phonetics No. 15. 1986, 3-100.

Inter-aural Speech Spectrum Representation by Spatio-Temporal Masking Pattern

Tatsuya Hirahara

ATR Auditory and Visual Perception Research Laboratories

Twin 21 Bldg. MID Tower, 2-1-61 Shiromi, Higashi-ku, Osaka, 540 Japan

ABSTRACT

In this paper, several speech sounds are examined by a masking method to show typical examples of speech spectrum in the auditory pathway represented by a spatio-temporal masking pattern and to clarify differences between interaural and physical representation of speech spectrum. Three types of Japanese speech, monosyllables, continuous speech and a monosyllable reproduced time reversely, are chosen for masker sounds. Using 1/3 octave band noise bursts with 25msec. duration as maskees, simultaneous and temporal masking are measured for the whole period of each masker. Spatio-temporal masking patterns thus obtained are an inter-aural speech spectrum. Compared with the physical spectral pattern: speech onsets and the formant structure, in particular, the transition of formants are emphasized and represented prominent in the masking patterns. These spectral emphases in the auditory pathway are composed of three functions, AM/FM masking, forward/backward masking, and adaptation. Further, taking into account the considerable differences between inter-aural and physical representation of speech spectrum, the inter-aural spectrum can be implemented as better representation of speech spectrum in speech feature extraction and speech signal processing by computers.

INTRODUCTION

Spectrum analysis in the human auditory system is performed by cochlear function and neural network processing. These characteristics are assumed to be different from those of spectral analysis techniques based upon digital signal processing we usually use.

A number of psychophysical and neuro-physiological studies have been carried out to date to obtain knowledge on this auditory spectral analysis characteristics[1,2,3]. These studies indicate that the auditory system has its own signal processing functions such as, critical band filtering, lateral inhibition, adaptation, saturation, combination tones generation, masking and so on. Therefore, the inter-aural spectrum, i.e. sound spectrum representation in the auditory pathway, is different from the physical spectrum. Also, the remarkable abilities of the human auditory system to detect, separate, and recognize speech sounds are assumed to be performed using these inter-aural spectrum as input data for higher level signal processing. Therefore, inter-aural spectrum is superior to the physical spectrum representation when discussing perceptual cues of speech sounds.

From this standpoint, recent efforts have been made to develop a speech analysis method based on auditory functions. Several researchers have reported studies that simulate some auditory functions and a number of them have tried to apply their results, in part, to the field of automatic machine speech recognition [4,5,6,7,8,9,11].

Very few reports, however, have been given on studies concerned with inter-aural representation of dynamically varying and/or complex structured sound, such as speech [10,12,13,14,15]. It is the purpose of this paper to observe speech sounds from the viewpoint of spatio-temporal masking pattern, and show typical examples of speech sound representation in the auditory pathway. Differences between inter-aural spectrum and physical spectrum representation of speech are also clarified.

METHODS

Basically, two methods have been used to measure inter-aural spectral patterns. One is a neuro- physiological method, by which activities of the auditory nerve fibers measured directly correspond to sound stimuli inputs [17]; however, this method can not be used to study human auditory system. Another is a psychophysical method, by which activities of auditory system are measured indirectly. Three major psychophysical methods used to measure peripheral activity are the masking method [10,16], the pulsation threshold method [19] and the cancelling method [18]. In this paper, two traditional masking methods, temporal and simultaneous masking methods, were chosen since they are most appropriate for measuring inter-aural spectral representation of speech sound of wide range, time-varying spectral dynamics.

A masking value $M(m;t,s)$ is defined as the threshold shift of maskee signal s overlapped with masker sound m at time t. from masker onset. That is,

$$M(m;t,s) = L(m;t,s) - L(s) \quad [\text{dB}] \qquad (1)$$

where $L(m;t,s)$ and $L(s)$ are the hearing threshold level of maskee signal s with and without masker sound m. present at the time t. When the maskee signal s is a function of frequency f, $M(m;t,s)$ is also a function of frequency f. Therefore, a three dimensional masking pattern for the masker sound can be obtained by measuring $L(m;t,s(f))$ at various t and f. This three dimentional masking pattern is considered to be an inter-aural spectrum representation of a masker sound after peripheral auditory processing.

EXPERIMENTS

Three experiments were carried out. Maskers were different types of speech sounds, while maskee signals and experimental procedure remained the same throughout the experiments.

Masker Experiment I : Japanese monosyllables */e/*, */re/*, */be/* and */de/* of 300 to 400 msec. duration were chosen for maskers. Experiment II : A continuous sentence speech */Are dewa eberesutoni noborenai/* (He can not climb Mt.Everest.) was chosen for the masker. This sentence was selected because it included the monosyllables */e/*, */re/*, */be/* and */de/*. The sentence duration is 1.6 seconds. Experiment III : Reversally reproduced monosyllable */re/* was chosen for the masker to investigate how the time axis, inverse of the masker, affected the masking pattern. These speech samples were uttered by a male speaker in a soundproof room. Their average fundamental frequency was about 100Hz.

Maskee Maskee signals were sixteen 1/3 octave band noise bursts of 25 msec. duration with a linear rise and fall time of 5 msec. Their center frequencies *fc* covered 100Hz to 4kHz.

Setup Experimental setups and the time chart of the stimuli are shown in Fig.1. The masker and maskee were D/A converted simultaneously via different channels. Both of them were low-pass-filtered(Fc=5kHz,-96dB/Oct.), individually attenuated to a certain level, mixed together, then presented to a subject monauraly through headphones (STAX SR-5) in a soundproof room. The presented level of masker was fixed at 70dBSPL.

Procedure Every threshold value was determined by the method of limits. At the beginning of the experiment, the maskee level was set below the threshold. Subjects were instructed to judge whether or not the maskee signal could be heard with the masker sound for each presented stimulus by pushing 'Yes' or 'No' button on the switch box. Every time the 'No' button was pushed, the system increased the maskee level by 1 or 2dB automatically. The maskee level gave the threshold value when the 'Yes' button was pushed for the first time. To allow a judgement to be made correctly and easily, subjects were allowed to use two additional buttons: 'Again' to repeat the same stimuli, and 'Check' to repeat the masker sound only.

Two well trained male subjects participated in the experiments. Measurements were repeated at least 3 times for every threshold $L(m;t,s)$ and at least 10 times for every $L(s)$ on different days for each subject.

RESULTS

The sound spectrogram and speech waveform of the monosyllable masker /de/ are shown in Fig.2 (a) and (b), respectively. In Fig.2 (c), the spatio-temporal masking pattern measured every 25 msec. for this masker is depicted. The fine spectral structures of the masker sound, in particular, the first formant transition (e.g. at t = 125 to 175 msec.) and the vowel formant structure (e.g. at t = 175 to 300 msec.), are clearly observed in the masking pattern.

Figure 3 shows masking spectra and 1/3 octave-band spectra for */de/* at t = 250 msec. Solid lines represent masking spectra, i.e. the inter-aural spectra, and broken lines represent 1/3 octave-band power spectra, i.e. the physical spectra. Thick and thin solid lines represent differences between the two subjects. In Fig.3, the first formant (F1) in the masking spectra appears more prominent than that in the power spectra since masking values in the lower frequency region were small.

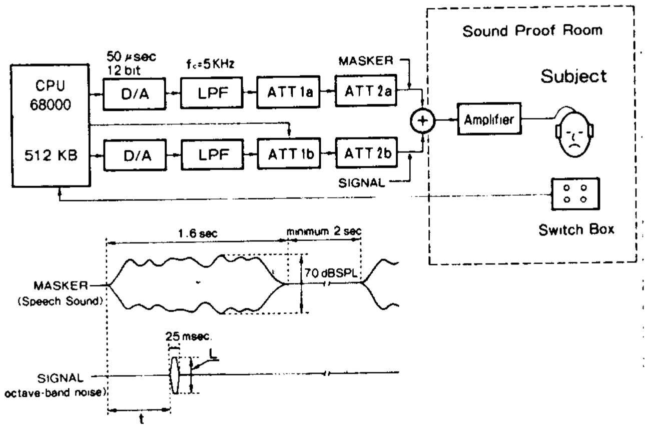


Fig.1 Experimental setups and the time chart of the stimuli. Both masker sound and maskee signal are D/A converted (20kHz, 12bits), low-pass-filtered (Fc=5kHz,-96dB/oct.), individually attenuated, mixted togather, then presented to a subject monauraly.

Figure 4 (a)-(c) show masking patterns and a 1/3 octave band power spectral patterns for the masker sound */de/* as a function of time. When compared with the power spectral pattern, three distinctive characteristics are observed in the masking pattern. (1) Masking does not take the value of 0 dB at the time before the beginning (t = -25 msec.) and after the end (t = 400 msec.) of masker speech. (2) Masking value increases remarkably at speech onset (t = 25 msec.) and at the transitional part of the formant. (3) Masking value decreases gradually in the vowel part. These characteristics were commonly observed in each masking pattern measured with respect to other monosyllable masker sounds.

The spectrogram and the speech waveform of the continuous speech masker are shown in Fig.5 (a) and (b). A spatio-temporal masking pattern measured every 25 msec. for this masker is depicted in Fig.5 (c). Formant structures and formant transitions are clearly represented in Fig.5 (c) as well as Fig.2(c).

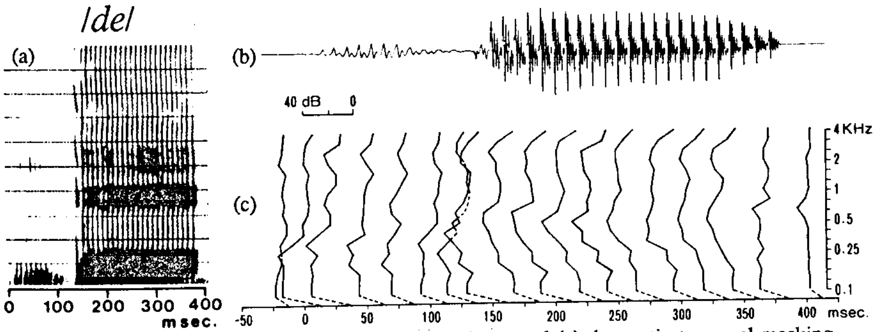


Fig.2 (a) Wide band spectrogram, (b) speech waveform, and (c) the spatio-temporal masking pattern measured every 25 msec. for the monosyllable masker /de/.

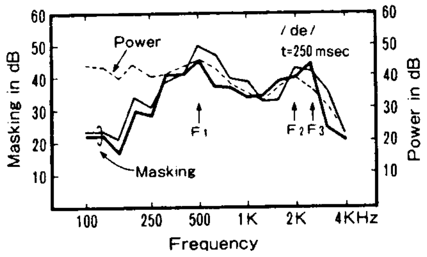


Fig.3 The masking spectra (solid lines) and 1/3 octave band power spectra (broken line) for /de/ at t=250 msec. Thick and thin lines represent differences between the two subjects.

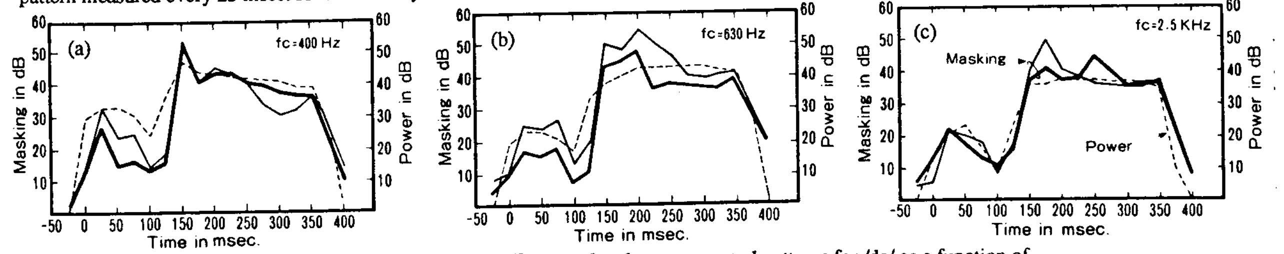


Fig. 4 Masking patterns and 1/3 octave band power spectral patterns for /de/ as a function of time at three frequency bands: (a) fc=400Hz, (b) fc=630Hz and (c) fc=2.5kHz.

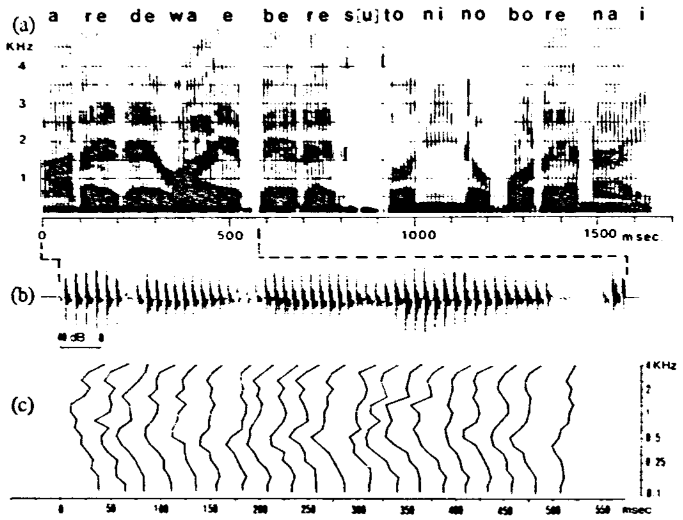


Fig.5 (a) Wide band spectrogram , (b) speech waveform and (c) the spatio-temporal masking pattern measured every 25 msec. for the continuous sentence speech */aredewa eberesutoni noborenai/*.

Figure 6 (a) - (c) represent masking patterns and power spectral patterns for continuous speech as a function of time. Dips seen in the masking patterns at each syllabic boundary, around t = 75, 200, 300, 475 msec., are deeper and more noticeable than those in the power spectral patterns. One reason for the dip depths in the masking patterns being prominant is that the masking values proceeding and succeeding the dips are large.

Figure 7 shows a monosyllable speech spectrogram for */re/* in normal time axis. This monosyllable and a time reversally reproduced one (reversal */re/*) are the maskers in the third experiment. Figure 8 (a) - (d) show masking patterns using */re/* (solid lines) and the reversal */re/* (broken lines) for a whole period of the masker sound. Comparing both masking patterns, masking values increase at the onset of each masker sound, whether it is reproduced reversely or not (i.e. at t = 25 msec. for */re/* and at t = 275 msec. for reversal */re/*). This phenomenon appears most remarkable at frequency band fc= 160Hz. Masking values of */re/* are larger in 5 to 10 dB than those of reversal */re/* at around t =50 to 100 msec. at frequency band fc= 315Hz and 1.6kHz. These frequency bands are those within which, Fl and F2 transition occurs, although their transition direction is different between */re/* (i.e. upward) and the reversal */re/* (i.e.downward).

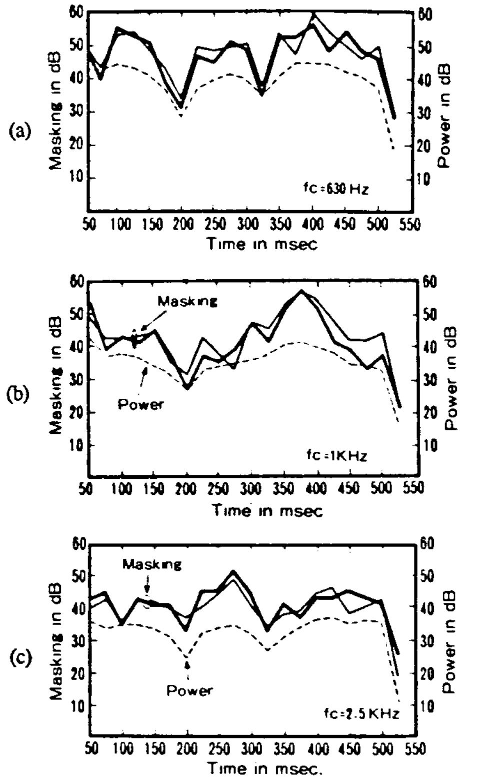


Fig.6 Masking patterns (solid lines) and 1/3 octave band power spectral patterns (bloken lines) for the continuous speech as a function of time at three frequency bands: (a) fc=630Hz, (b) fc=1kHz and (c) fc=2.5kHz. Thick and thin lines represent differences between two subjects.

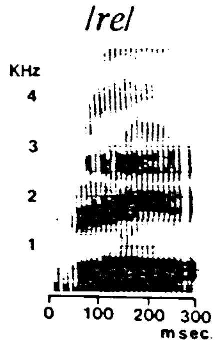
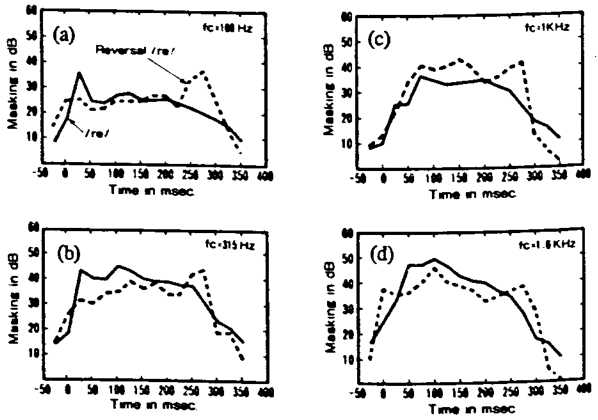


Fig.7 Wide band spectrogram of the monosyllable masker sound /re/ in normal time axis. This monosyllable and a time reversally reproduced one (reversal /re/) are the maskers in the third experiment.

Fig.8 Masking patterns for /re/ (solid lines) and reversal /re/ (bloken lines) at four frequency bands: (a) fc=160Hz, (b) fc=315Hz, (c) fc=1kHz and (d) fc=1.6kHz



DISCUSSION

Results show several important characteristics which seem to play important roles in physical to inter-aural spectral transformation by means of the non-steady part emphatic functions. Three of these characteristics found in comparing masking patterns with physical spectral patterns are discussed in this section.

First, speech onset is emphasized in masking patterns. This onset emphasis is caused by a temporal increase of the amplitude component, that is an upward amplitude modulated (AM) component. There exists a downward AM component due to temporal amplitude decrease at speech offsets. Although, the offset emphasis produced by the downward AM component is smaller than the onset emphasis.

Second, formant transitions, in particular, Fl and F2 transitions, in masking patterns are more prominent than those of the physical spectra. This is an inter-aural emphasis caused by formant movement which is composed of both AM and frequency modulated (FM) components. These AM components are produced by temporal change of each harmonics level. One of the FM components is produced by the resonance frequency movement itself as seen in broad band spectral patterns. In a strict sense, a formant transition is not a real movement of a physical existing frequency component, such as sweep tone, but a movement of spectral envelope peaks estimated from several resonated harmonics of the fundamental frequency. However, this formant movement increases masking values as well as frequency sweep tone [20]. Another FM component included in the formant transition is fluctuation of harmonics frequencies. This fluctuation is a physically existing movement of the FM component due to fundamental frequency change.

Third, formants in middle and higher frequency ranges become prominent in masking patterns resulting from small masking value in the lower frequency range. This is due to a general masking characteristic that lower frequency components mask higher ones more effectively than higher frequency components do lower ones. In this paper, suppression effects along the frequency axis, which is seen in the results given by the pulsation threshold [12,13], are not reflected on the masking pattern since traditional masking procedures were used.

On the other hand, a decrease in masking values at the middle part of vowel is noticeable in the monosyllable masking patterns, but not so noticeable in the continuous speech patterns. This phenomenon is an adaptation effect caused by the steady state vowel part which has a several hundred milisecond duration. In a continuous speech masker, vowel part are not long enough to cause the adaptation effect. Since the adaptation decreases masking values at long steady vowel part, non-steady parts of speech (including onset and formant transitions) preceding and/or succeeding these vowels are relatively emphasized in the auditory pathway.

Furthermore, as shown in the results of the third experiment, reversing the time axis of a masker sound gives us completely different masking patterns. Two spectra with the same exact frequency structure have two different masking values. This suggests that spectral change direction and interaction between temporarily adjacent components play important roles in the physical to inter-aural spectral transformation.

To summarize, it is clear that temporal amplitude varying features, transition of formant frequencies and structures, which are considered to be important cues in speech perception, are emphasized and more prominent in the auditory pathway than those in physical spectrum patterns. It is expected that inter-aural spectral representation will bear better results than physical spectral representation when implemented in speech signal processing by computers. The physical to inter-aural spectrum transformation discussed in this paper can be described quantitatively by simulating AM/FM component emphasis, backward/forward masking, adaptation and lateral inhibition. This transformation can be implemented in a automatic speech recognition preprocessor as a better representation of speech spectrum capable of discriminating two utterances with confusable physical spectra.

CONCLUSION

In this paper, three types of speech sounds are examined by a masking method to show typical examples of inter-aural representation of speech spectrum represented by a spatio-temporal masking pattern and to clarify differences between inter-aural and physical representation of speech spectrum. Our findings are summarized as follows:

1) Compared with the physical spectral pattern: speech onsets and formant structure, in particular, formant transitions are prominent in the masking pattern.

2) Spectral emphasis is presumably composed of three auditory functions: AM/FM components emphasis, forward/backward masking and adaptation. These play important roles in physical to inter-aural spectrum transformation.

3) The direction of AM/FM component movements in speech sounds is of great importance and strongly affects the process of producing the inter-aural spectrum pattern.

4) Taking into account the considerable differences between inter-aural and physical representation of speech spectrum, the inter-aural spectrum can be implemented as a better representation of speech spectrum in speech feature extraction and speech signal processing by computers, particularly in automatic speech recognition by machine.

ACKNOWLEDGEMENTS

This work was mainly done at NTT Integrated Communications Labolatories. The author is grateful to Dr. Ryouhei Nakatsu for his continuing guidance and encouragement, and to Dr. Noboru Sugamura for his fruitful discussions. The author also wishes to thank Dr.Yoh'ichi Tohkura for his helpful advice in writing this paper.

References

[1] E.Zwicker and E.Terhardt (Eds.) (1974) *Facts and Models in Hearing*, Springer-Verlag, New York.

[2] R.Plomp (1976) , *Aspects of Tone Sensation; A psychophysical study*, Academic Press.

[3] R.Carlson and B.Granstrom (Eds.) (1982) *The Representation of Speech in the Peripheral Auditory System*, Elsevier Biomedical,

[4] D.H.Klatt (1980), "SCRIBERand LAFS: Two new approach to speech analysis," in *Trends in Speech Recognition* ,W.A.Lea (Ed.), Prentice-Hall, pp. 529-555

[5] J.B.Allen (1985)," Cochlea Modeling," IEEE ASSP Magazine, Jan., pp. 3-29

[6] R.F.Lyon (1982), "A Computational Models of Filtering, Detection and Compression in the Cochlea," Proceedings of ICASSP, pp. 1282-1285.

[7] S.Seneff (1986)," A Computational Model for the Peripheral Auditory Sysytem Application to Speech Recognition Research," Proceedings of ICASSP, pp. 1983-1986

[8] E.Zwicker (1986), "Peripheral Preprocessing in Hearing and Psychoacoustics a: Guidlines for Speech Recognition," Proceedings of the Symposium on Speech Recognition, Montreal, pp.1-4.

[9] S.A.Shamma (1986)," The auditory processing of speech," Proceedings of the Symposium on Speech Recognition, Montreal, pp.14-17

[10] T. Ifukube (1973), " Masking by Frequency Modulated Tone," J. Acoust. Soc. Japan, vol.29, No.11, pp. 679-687.

[11] T. Ifukube (1975), " Auditory Masking of Amplitude Modulated tone and its Analysis by Analog Simulation," J. Acoust. Soc. Japan, vol.31, No.4, pp.237-245

[12] T.Houtgast (1974) ," Auditory Analysis of Vowel-Like Sounds," Acoustica vol.31, pp.320-324

[13] R.S.Tyler and B.Lindbolm (1982)," Preliminary study of simultaneous-masking and pulsation-threshold pattern of vowels," J. Acoust. Soc. Am., vol.71, No.1, pp. 220-224.

[14] B.C J.Moore and B.R.Glasberg (1983), " Masking patterns for synthetic vowels in simultaneous and forward masking," J. Acoust. Soc. Am., vol.73, No.3 pp.906-917

[15] A.Sidwell and Q.Summerfield (1986), " The Auditory Representation of Symmetrical CVC Syllables," Speech Communication, vol.5, No.3,4, pp.283-297

[16] E. Miyasaka (1983), " Spatio-temporal characteristics of masking of brief test-tone pulses by a tone-burst with abrupt switching transients," J. Acoust. Soc. Japan, vol.39, No.9, pp.614-623

[17] B.Delgutte and N.Y.S.Kiang (1984), "Speech coding in the auditory nerve I-III," J. Acoust. Soc. Am., vol.75, No.3, pp. 866-896.

[18] J.L.Goldstein (1966), " Auditory Nonlinearity," J. Acoust. Soc. Am., vol.41 No.3, pp 676-689

[19] T.Houtgast (1973), "Psychophysical Experiments on Tuning Curves and Two-Tone Inhibition," Acoustica, vol.29, pp.168-179.

[20] T. Hirahara (1985)," Auditory Response by Formant Transitional Stimuli from the Viewpoint of Simultaneous Masking," Proceedings of the Autamn meeting of Acoust. Soc. of Japan, pp. 249-250

AUTOMATIC WORD STRESS DETECTOR

L.Zlatoustova N.Kozlenko M.Khitina L.Zakharov

Department of Philology, Moscow State University
Moscow, USSR, 119899

ABSTRACT

This paper presents the results of measurements of Russian word-stress parameters (using acoustic and statistic methods). There are also demonstrated some specific peculiarities of Russian word stress which are employed in the computer model of an automatic word stress detector.

INTRODUCTION

The analysis of the literature on the word stress reveals that the Russian stress is not distinguished in current speech by certain specific parameters. It rather aims at structuring or shaping of a phonetic word on the whole. The peculiar character of the Russian stress presents certain difficulties for automatic stress detection.

Among the acoustic correlates usually considered for stressed vowels characteristics are fundamental frequency, duration, intensity and spectrum. The absolute and relative values are of interest.

It is necessary to specify the rhythmic organization of phonetic words and frequency of occurence of phonetic words and their rhythmic structures (RS). A rhythmic structure characterizes a single word or a few words, autonomous or syntactic, forming a stressed group. RS type is designated by a fraction where the number of syllables in a phonetic word is a nominator and the ordinal number of the stressed syllable is a denominator. RS variety is designated by a succession of consonants and vowels in a RS which is shown in terms of C (consonants) and V (vowels).

The rhythmic pattern of a Russian text does not permit two or more succesive stressed syllables or a long succession of unstressed syllables. The average length of an interval between two successive stresses varies from I to 3 syllables, the most frequent being 2 syllables /I/.

In the initial and final RS there are usually no more than 2 prestressed or poststressed syllables. The RSs containing from I to 4 syllables are most characteristic for Russian syntagmas. The most frequent ones consist of 3 RS, average syntagma length being 2,8 RS.

The phonetic word of 2 or 3 syllables are predominant. The data obtained by prof. L.V.Zlatoustova show that the 6 most frequent RS types: I/I, 2/I, 2/2, 3/2, 3/3, 5/3 cover approximately 70% of any Russian text. The mentioned RS types and 3 more: 3/I, 4/2, 4/3 can cover about 90% of any Russian text (e.g. the text of a dialogue) /2/. The distribution of general RS types in different languages is demonstrated for fiction and newspaper texts /3/ (see Table I).

Table I. Frequency of occurence of RS types (%)

RS TYPE	LANGUAGE			
	RUSSIAN	BULGARIEN	ENGLISH	GERMAN
I/I	I3	I3,8	27	I7,6
2/I	I6,8	I6,5	I8,8	I6,9
2/2	2I,3	I0,6	2I,5	I6,7
3/I	6	7,8	I,4	7
3/2	I9,6	I6,6	5,5	23,3
3/3	7	2,3	3,8	3,2
4/I	I,8	I,8	0,7	-
4/2	4	I0,9	I,9	2
4/3	9,2	II,7	I,9	2,8
4/4	I,4	I,3	0,5	0,5
5/3	5	2	I,4	I

Stress in Russian is normally placed on one of the initial three syllables of a phonetic word. Preferably it is one of the central syllables of a word.

RS REALIZATION IN SPEECH

The specific character of RS realization depends on frequency of occurence of a word, its position in an utterance or syntagma. It also depends on a variety of conditions: prepared reading vs. spontaneous speech, an artistic reading by an actor vs. neutral reading by a layman, RS constituting a one-word utterance or being part of a context, the character of a text (variety and genre), normative vs. dialectical speech,etc. Extralinguistic factors are also to be taken into account.

To determine an RS type is necessary to learn the number of syllables in a RS, the stressed syllable and its position relative to unstressed syllables, consonant and vowel component markers typical of beginning and end of a RS.

To detect the stressed syllable we took a number of RSs with stressed syllables (and vowels) clearly dissimilar to the unstressed syllables of the same utterance. It should be mentioned that stressed and unstressed syllables may have similar values of such parameters as fundamental frequency, duration, intensity and spectrum. So it is preferable to choose the most typical and frequent samples as the test material for developing an automatic teaching system (ATS).

Thus our minimal text consisted of I to 4 utterances, an utterance consisted of 2 syntagmas and so on. A separate utterance also can form a text. In the utterance of 2 RSs the phonetic word of any of the 9 rhythmic types may occur as the first component. The second component may be one of the following types: I/I, 2/I, 2/2, 3/I, 3/2. 4/2. Such RS types as 3/3, 4/3, 5/3 can succeed every RS type exept 3/I, 4/2, 5/3. The third phonetic word in a three word utterance is chosen in a likely manner.

The most frequent words are preferable to be chosen to form utterances. The comparison of the structural types of word entries in word counts shows that independently on the material used (written vs. oral text) and the size of analysed selection (400 thousand vs. I million occurences). One can note a certain similarity of rhythmic types and varieties in both selections.

The important finding was that the most frequent structural models found in word counts are the most frequent in the texts. Thus, there is a limited number of basic structural types of words and phonetic words in Russian /4/.

One of the structurally important variables of a phonetic word is the relative duration of its vowels in strong and weak positions. It has been repeatedly mentioned that vowel length change accounts for phonetic word duration variance due to conditions such as separate vs. contextual occurence, initial vs. final position in an utterance, emotional vs. neutral content, whether or not a RS bears the phrase accent,etc.

The temporal structure of a phonetic word is essentially conditioned by the relationship of broad and narrow vowels in strong and weak positions. Finally, it is important whether the initial syllable of a word is covered and the final syllable is open.

RESULTS AND DISCUSSION

The comparison of durations of the stressed and the first prestressed broad vowels in the RS of VCV́CV(C) and CVCV́CV(C) varieties revealed that in the final position the first RS vowel is always shorter than the stressed vowel irrespective of its being open or covered.

In the RS of the CVCV́CV(C) variety with the covered initial syllable in the beginning of an utterance the first prestressed vowel is shorter than the stressed vowel. The first prestressed vowel that starts an utterance is always longer than the stressed vowel.

Initial position (one-syntagma utterance)

The unstressed vowel in the absolute beginning of a RS of the 3/2 type and VCV́CV(C) variety is in 89% instances longer than the stressed one. In case of a covered initial unstressed vowel (RS of the CVCV́CV(C) variety) is in 90% instances shorter than the stressed one.

Final position (one-syntagma utterance)

In the final position of a RS in an utterance both the initial vowel of an open prestressed syllable and the vowel of the first prestressed covered syllable (i.e. RSs of the VCV́CV(C) and CVCV́CV(C) varieties) are in 96% instances shorter than the stressed vowel.

Table 2. The relationship of durations of prestressed and stressed vowels connected to narrow vs. broad types of vowel sounds /2/

VOWELS	RS	COEFFICIENT
prestressed narrow stressed broad		0,535
	2-syllable	0,486
	3-syllable	0,59I
	4-syllable	0,454
prestressed narrow stressed narrow		0,867
	2-syllable	0,82
	3-syllable	0,926
	4-syllable	0,827
	5-syllable	I,043
prestressed broad stressed broad		0,796
	2-syllable	0,826
	3-syllable	0,722
	4-syllable	0,827
	5-syllable	0,88

The mean intensity of RS components have been analysed as function of RS position in an utterance, qualitative nature of the components in a RS, and of syllable types. One could justifiably expect that in final position of a RS in an utterance the intensity of all the components have been lower than in utterance initial RSs. This regularity is connected with the phrase intensity contour and has been repeatedly mentioned recently.

The mean intensities allow to estimate intensity changes in strong and weak elements of a RS.

The question of the absolute prominence of stressed vowels (stressed syllables) especially in the RS with a narrow stressed vowel (in a closed syllable) and a broad prestressed vowel is of special interest.

Our experiment shows that a tendency exists for the stressed vowel (independently on its position) to be more prominent in a RS in the absolute beginning of an utterance with similar vowels and consonants. If the first syllable is stressed it is most certain to be marked by increased intensity.

In the absolute beginning in a RS of the VCV́CV variety the usual distribution of mean intensity is like this: the poststressed vowel is the least intensive, the most intensive is the stressed one, and the final vowel, however short, is more intensive than the preceding one. A different distribution is revealed in the absolute ending of a RS: the most intensive is the stressed vowel, the least intensive - second poststressed. In some cases the intensities of the stressed and the poststressed vowels are equal (in the initial position of a ES). However, the stressed vowel may be as intensive as the following unstressed if the former is front, high and the latter – low.

The utterance final stressed vowel in ES of [illegible] or [illegible] varieties is not, as a rule, marked by increased intensity, the stressed vowel, however, is more intensive if a ES is under phrase intensity contour termination /5/.

In the beginning of an utterance there is a marked tendency for the initial syllables to be stronger. This is probably due to their greater importance as information carriers.

The syllable intensity may alter the observed relationships if compared syllables have different sound structures. Thus the analysis of mean intensity suggests the tendency for the stressed vowels in the utterance initial and the utterance final ESs (the latter with the first syllable stressed) to be marked by an increase of energy. Though the case may be that the intensities of stressed and unstressed vowels would not differ due to some sound dissimilarities.

Table 3. Mean intensity /dB/

POSITION IN AN UTTERANCE

beg.	end.	beg.	end.	beg.	end.
V́		V – I poststressed		V – 2 poststressed	
43,6	34,0	38,8	24,8	35,0	I9,0

beg.	end.	beg.	end.	beg.	end.
V – I prestressed		V́		V – I poststressed	
44,0	36,0	45,0	34,8	41,7	21,3

beg.	end.	beg.	end.	beg.	end.
V – 2 prestressed		V – I prestressed		V́	
4I,8	35,8	42,7	35,4	43,25	30,7

Thus the mean intensity estimates of the vowels allow to determine the position of a ES in an utterance, characterize its sound structure but cannot be used as a reliable stress detector in Russian.

THE STRESS DETECTION ALGORHYTHM

In accordance with the above-mentioned we propose to use in stress detection algorhythms the most important acoustic parameters directly related to the phonetic word structure.

To determine the ES type in any position only one variable (duration) or two (duration, intensity) are sufficient. This is true for a one-word utterance with the first stressed syllable (ES types: I/I, 2/I, 3/I – with the covered initial and closed final syllables).

We suggest the stress detection algorhythm based on the method of loudness measurement. The objective methods of subjective loudness /Ψ/ measurements are available now as the international standard N 532)

In addition to such variables as intensity /I/, duration /t/, energy /E/, fundamental frequency /Fo/, etc. loudness allows to take into account the spectrum shape /S/ and the sound field shape /P/. Thus the loudness is a functional of a series of variables $\Psi = \Psi(I, t, Fo, S, P)$ and as a tool of measurement is more adequate than most of physical (acoustic) parameters. As far as energy /E/ is a function of intensity and duration, the loudness is a complex function of energy. In the present study both I and E were used and a comparison has been made.

The test material has been tape-recorded and re-recorded by means of an analog-to-digital converter on a digital tape which has been computer analysed by the Automatic System of Scientific Research (ASSR). The system is able to measure loudness and loudness level (in accordance to international standard N 532) and possess the necessary service software /6/, /7/.

The result of such an analysis are the automatic graphs: oscillograms with the I/20 msec time marks, multiparametric graphs reflecting intensity, loudness level and fundamental frequency as function of time.

The algorhythm has been tested on a limited material where the compared vowels were prosodically different. It demonstrated that loudness integral maxima corresponded to stress vowels in all cases. After parameter optimization the algorhythm has been tested on a more varied material with a new variable of the word position in an utterance. A new phrase test was compiled to compare vowels:

I. Stressed and strong unstressed vowels (non covered in the first prestressed syllable and open in the first poststressed syllable).

2. Vowels of different proper duration and intensity.

3. Vowels in different phonetic environments.

After the linguistic correction the reliability of the algorhythm increased to 92,3%.

CONCLUSION

The attained reliability justifies the use of loudness along with other parameters in working out algorhythms of automatic word stress detection in current speech. It's use as a part of the ATS is recommended.

REFERENCES

/I/ Kagarov E.G. O ritme russkoy prozaicheskoy rechy. Nauka na Ukraine, I922, N 4, s.324-332.

/2/ Zlatoustova L.V. Foneticheskie edinitsy russkoy rechy. M.G.U., I98I. – Io6 s.

/3/ Zlatoustova L.V. Universalii v prosodicheskoy organizatsii teksta (na materiale slavianskikh, germanskikh i romanskikh jazykov). Vestnik MGU, ser. 9, Philologia, I983, N 4, s. 69-78.

/4/ Khitina M.V. Edinitsy ritma russkoy rechy i ikh ispolzovanije v razlichnykh tekstakh: Avtoreph. diss. ... kand. philol. nauk. M., I986. – I9 s.

/5/ Zlatoustova L.V. Phoneticheskaja priroda russkogo slovesnogo udarenija: Avtoreph. diss. ... kand. philol. nauk. L., I953. – 23 s.

/6/ Kozlenko N.P., Rezviakina Z.N. Izmeritel' urovn'a sluchainykh protsessov. B'ulleten' isobretenii i otkrytii i tovarnykh znakov. BIOTZ,N I9, I968.

/7/ Kozlenko N.P. Avtomaticheskoje raspoznavanije slukhovykh obrazov na osnove modeli integralnogo vosprijatija kachestva zvuka. Materialy Vsesojuznogo seminara "Avtomaticheskoje raspoznavanije slukhovykh obrazov" (ARSO-I3). Novosibirsk, I984, s. 75-76.

ACOUSTIC – PHONETIC BASIS OF SPEECH RECOGNITION ALGORITHMS

G. G. RODIONOVA

Computer Centre of the USSR Academy of Sciences,
Vavilova str., Moscow, USSR, 117333

ABSTRACT

The algorithms based on the wider use of acoustic-phonetic (APh) information are described. These algorithms include APh-clustering of training set and APh-classification on unknown message. An APh-structure description of speech signal is presented.

1. INTRODUCTION

Automatic speech recognition (ASR) is a key problem of the modern speech technology. In last years a variety of approaches to ASR have been explored and certain progress has been made in this field. This progress is largely due to the use of widely adopted formalistic techniques such as the most popular dinamic-programming (DP) method. DP-technique is based on whole-word template matching making it's performance quite high due to the absence of segmentation error and other advantages. However such problems as large time and storage requirements, dicrimination of similar words, account of coarticulation effects arise. Now it is quite clear that this approach is not promising. An inevitable return to accounting for "human" aspect of speech signal requires the design of acoustic-phonetic information based algorithms. In this paper we briefly describe the algorithms containing a set of procedures used for APh-clustering in training and recognition.

2. PARAMETRICAL DESCRIPTION OF SPEECH SIGNAL

The accuracy of recognition is evidently dependent on reliability of every level of a recognition system, and errors in coding and parametrical description are the most essential.

The speech signal processing hardware has been developed in the Computer center of the USSR Academy of Sciences. This hardware is based on the principle of maximal account of acoustic and phonetic features of a speech wave. Special devices are designed to extract and input into computer a variety of parameters, characterising the 10-20 ms intervals (time-segments) such as:

F_1, F_2 - average first and second formant frequency,

F_0 - pitch frequency,

N_0 - number of zero crossings,

A_0 - total energy etc.

Among these parameters a set of the most informative ones $\{P_i\}$, $i = 1,\dots,4,5,6$ has been extracted. These parameters ensure that the requirements of minimal time and maximal accuracy of recognition are fulfilled.

When an algorithm operates with templates and uses the APh-information, it is necessary that these APh-features provide the distinction between all of the templates in the conditions of real-time processing.

We have developed computer programs to obtain some lexical parameters which reflect the phonetic structure of a message. So the parameters show the presence (or absence) of certain phoneme-like subword units (ph-segments) and their order in a word. As a result of the procedures the secondary characteristics from a set of primary ones have been obtained; e.g.:

$R_1 = 1$, if a word S_i contains the noise consonant (NC) segment of a duration τ^i, which does not exceed a threshold value, τ_{th}.: $\tau^i < \tau_{th}$, $N_0^i \leq N_{0th}$, $A_0^i < A_{th}$;

$R_4 = 1$, if the stressed vowel is of the "a" type, i.e. $\tau^i > \tau_{th}$, F_1^i and F_2^i are lying in their standart domain of mean values, $A_0^i > A_{th}$,

etc.

Vector $W = \{R_i\}$, $i = 1,\dots,8,16$ reflects a certain information about message phonetic structure, for example, the word "сани" is characterised by vector:

$$W = \{1,1,0,1,0,0,1,0\}.$$

This means:

$R_1 = 1$: there is a noise consonant (NC) during the given word realization (WR);

$R_2 = 1$: this NC is in the beginning of the word;

$R_3 = 0$: there are no second NC in the word;

$R_4 = 1$: the stressed vowel is of the "a" type;

$R_5 = R_6 = 0$: the stressed vowel is not of "y" or "и" type;

$R_7 = 1$: the NC has the energy maximum in the high frequency region;

$R_8 = 0$: R_8 is not computed when $R_3 = 0$; if $R_3 = 1$, then the value of R_8 depends on N_0 of the second NC.

Such a description is rather rough, but it is of a reliable nature. When the number of secondary features is equal to 16 or 24, the vector W is more informative, but in this case the description has some evident desadvantages. Thus, the original speech signal is presented by means of full description (FD), being two kinds of parameter, having different levels of extracting and different powers of adequacy.

Namely FD consists of:

(a) primary description - a temporal matrix $\|P_T\|$, where T is the message duration in 10 ms time-segments and

(b) secondary description - a vector W of binary lexical features (ph-segments).

This representation is more accurate then the one obtained with whole-word templates, where phonetical variations can be expressed only by adding other templates. It also shows better the distinctions between similar words. Such a description provides a more natural way of dealing with acoustic-phonetic information and, on the other hand reduces considerably the required amount of training material.

3. APh -CLUSTERING OF A TRAINING SET

The recognition system software consists of two parts: a teaching one, which provide a training set and a recognizing part, destined to carry out the search operations. The training set is formed by means of pronouncing every position of given vocabulary $\{S\}$(a word or a word combination with their attributes indicated: word code, speaker name, etc.). Input and full representation (primary and secondary) for every utterance is made. These descriptions are stored in two computer memory domains. Then the training set of the length M: $\{S_i\}$, i = 1,..,M,(which is at the beginning structureless) is clustered on phonetical features base, i,e, is divided into J structural clusters C_j, j = 1,...,J. The clustering process is performed with the aid of vector W components, lying in the nodes of a binary logical tree (BLT). These components are previously arranged and BLT is constructed after the user manner. It should be noted that the total number of terminal clusters J is not equal to 2^k, where k is the length of vector W. It is so since every branch of BLT does not contain all of the theoretically possible nodes due to the special nature of W. Thus we can estimate the mean number of templates, N_j, in the cluster C_j:

$$N_j \approx M/J.$$

In case of phonetical nonbalanced vocabulary this estimation may turn out to be rather approximate, but this fact is not of a great importance. The point of the method is that a cluster, C_j, contains a template which is relevant to unknown utterance with the same phonetical label j. APh-clustering is carried out automatically with the help of especially developed procedures of speech recognition system.

4. RECOGNITION ALGORITHM BASED ON APh-CLUSTERING

Spoken message pertaining to a given vocabulary $\{S\}$ is recognized by looking for a relevant pattern through a subset of templates that maximizes a measure of similarity with the input signal. The main feature of algorithms under consideration is that the search for a maximal similar candidate is made within the templates that form one cluster without any resortion to the remaining templates. The sample to check the recognition algorithm was defined by a given vocabulary $\{S\}$. First, the imput utterence S^i was transformed into primary parameter description, the matrix $\|P_T\|^i$. For the same message a secondary parameter sequence - vector W^i was calculated (in real-time) and an APh-classification was made by the values of vector W^i's components. So the S^i got a structure label, i.e. it was marked by the number of "its" cluster, j. APh-classification procedure was performed by using the same learning binary logical tree as in the learning stage. The fact that both the template and the searched for descriptions belong to the same terminal APh-cluster C_j makes it possible to restrict the search for relevant candidate $\widetilde{E}$ to objects of C_j only: $\widetilde{E} \in C_j$.

The choise of search strategy is of great importance to the outcome (error rates and recognition time), but the described algorithms are independent of this strategy. In our case the relevant candidate $\widetilde{E}_j$ was found by comparing the parametrical matrix $\|P_T\|^i$ with the matrices of templates composing cluster C_j.

If APh-clustering technique is well developed, the algorithms under consideration not only shorten the recognition time on the average by a factor of J times, but also improve the accuracy. The latter takes place because the smaller the number of processed templates the lower the error in classification.

CONCLUSIONS

Adequate description of a speech object is always of great importance. But in the problems on speech recognition dealing with an object that is highly variable in time and in the parameter space, the question of optimal formalization of this object is decisive. The algorithms described may be of interest for those who develop speech recognition systems and who realize that the role of acoustic-phonetic information should be strengthened.

SOUND IMAGE RECOGNITION BY HOLOGRAPHIC MEANS

V.R. IMAKOV, V.N. SOBOLEV

All-Union By Correspondence Electrotechnical Institute of Communications, Moscow, USSR 123855

ABSTRACT

Optical methods of sound image recognition are discussed as an alternate to recognition systems with von Neumann's architecture. The main design principles and algorithms are described.

INTRODUCTION

Most sound image recognition systems are based on computer systems (CS) with a von Neumann architecture (with a sequential instruction stream). As is known such CS are not equipped with functions (including input-output) required to process non-numeric data, such as speech, graphic images, etc. Due to the difficulties of real-time parallel processing of acoustic signals, it appears expedient to shift to customized optical computers (COC) for solving sound image recognition problems. Such COC may use both coherent and non-coherent light emissions, or their combination. Correlation type COC are the most widely used due to the simplicity and efficiency of complex signal transformations, such as convolution, correlation, Fourier transforms, Hankel transforms, multiplication of matrices, etc. In contrast to traditional digital CS in which the elementary operation is a comparison by mod 2, in COC of the correlation type an elementary operation is a complex functional or integral transformation with an execution time determined only by the time of light travel through the optical media and assemblies, which can be some 10 ns to 10 ps. Another feature of COC is their ability to perform multiparallel signal processing. Optical computers are equivalent to CS with 10^6 to 10^{12} inputs. The number of output channels can range from 1 to kn, where n is the number of inputs and k = 1, 2,3... Another feature of COC is the ease and simplicity of processing multidimensional objects. COC which should simultaneously, in fractions of a microsecond add and divide hundreds of millions of numbers or multidimensional matrices can be designed without undue engineering problems, while such speeds in traditional digital CS are unattainable, especially if the simultaneous processing of great data bulks is taken into account. Many researches tend to treat acoustic signals as a unidimensional $f_t(x)$ one, rather than three-dimensional $f_t(x,y,z)$ signals, this being hardly always justified. Holography is an ideal means of mathematical simulation of three-dimensional objects, and holographic methods provide an exhaustive description of acoustic signals at all stages of its processing.

IMPLEMENTATION

Optical computers can be designed as analog, digital, or hybrid devices and may include various electronic and mechanical assemblies and units. COC functioning is based on the principle of generalised image delineation. The acoustic signal is fed to optical channel mostly via the so-called spatial-time light modulators (STLM) in the form of a two- or three-dimensional matrix consisting of several hundred or thousand cells controlled by the acoustic signal or its electric equivalent. The acoustic or electric signal causes a charge image to be formed on the modulator surface and this in turn modulates the light beam. STLMs may be operated both in the light transmission or light reflection modes. One of the STLM modifications is the controlled liquid crystal matrix (with acoustic, electric, or light control) with modulation frequencies up to about 60 kHz which is usually adequate for acoustic signal processing. The most advanced acoustic light modulators (of the Phototitus type) are based on CRTs [1,2], with a special crystal serving as the target inside the CRT and two electron guns for information recording and erasure, respectively. The charge image pattern on the crystal surface is formed by a controllable electron beam. During information readout the passing coherent light is phase and amplitude modulated. Real-time operation is provided by a second electron gun with a wide beam to remove the surface charge. As demonstrated [2], noncoherent optical processing is essentially reduced to linear operations with the image. In the classical non-coherent optical processor [3] the correlated output signal appears on a background of a constant bias. In the past, applications of such systems have been hampered by the low output signal-to-noise ratio and the difficulties of handling complex data. In the non-coherent optical speech processing system under study a higher signal-to-noise ratio is obtained and the constant bias is eliminated by modulating and demodulating the carrier. This makes it feasible to preprocess complex data to a form suitable to be input to the main coherent processor; this is acomplished with the aid of an obscure aperture of special shape. Consider two-dimensional functions: the recognized acoustic pattern f(x,y) and the reference pattern g(x,y) which are to be compared by closeness. In the general case, they can be complex quantities. Using their optical image, coded transparencies with transmission intensities f_c and g_c are generated:

$$f_c = 0.5|f(x_1,y_1)|\left\{1+\cos[2\pi\nu_c x_1 + \arg f(x_1,y_1)]\right\} \quad (1)$$

$$g_c = 0.5|g(x_1,y_1)|\left\{1+\cos[2\pi\nu_c x_1 + \arg g(x_1,y_1)]\right\} \quad (2)$$

where ν_c is the carrier frequncy used in the coding operation. Functions f_c and g_c are realized as intensities caused by biasing the cosine carrier. At $|f| \leq 1$ and $|g| \leq 1$, we have $0 \leq |f_c| \leq 1$ and $0 \leq |g_c| \leq = 1$. This means that processing the coded transparencies is equivalent to processing the initial functions. Correlation between f_c and g_c is provided by the base non-coherent processor (Fig. 1). Fresnel holograms for the plane of obscure P' were generated, with transmittance functions $g_c(x_1,y_1)$ in the input plane P_1 corresponding to various phonems and their combinations (dyads). The transparency modulated by f_c was positioned in the P_1 plane and thus the light intensity in the output plane P_1 was $f_c \circledast g_c$:

$$I_2 = f_c \circledast g_c = 0.25\ |f| \circledast |g| + 0.25\ |f \circledast g|\ \cos[2\pi\nu_c x_2 + \arg(f \circledast g)] + 0.25\ |f| \circledast |g|\, |\cos[2\pi\nu_c x_2 + \arg(g)]| + 0.25\ |g| \circledast |f| \quad (3)$$

If ν_c is sufficiently large, the signal spectra of main frequency band with a modulated carrier in Eqs. (1) and (2) will not overlap in the frequency domain. Since correlation is equivalent to multiplication in the frequency domain, the last two terms in Eq. (3) will be zero and the pattern in plane P_2 will be reduced to:

$$I_2 = f_c \circledast g_c = 0.25\ |f| \circledast |g| + 0.25\ |f \circledast g| \cos[2\pi \nu_c x_2 + \arg(f \circledast g)] \qquad (4)$$

To obtain the desired complex function $f \circledast g$ from the distribution in the P_2 plane the pattern in this plane was scanned by a raster in the x_2 direction, with the spatial carrier ν_c being transformed into a time carrier $S\nu_c$ (S is the scanning speed). Passing the video signal through a band-pass filter removes the first term of Eq.(3) and the second term then depicts the absolute value and phase of the $f \circledast g$ signal. In the transformation device used masks for the DC component and first, third, fifth and seventh derivatives of the spatial-temporal acoustic signal were provided, with the even derivatives zeroed out by an appropriately selected obscure function. Differentiation and averaging were holographic.The masks were programmed to provide a pseudo-formant representation of the speech signal, this ensuring an adequate invariance relative to different dictors. Pseudo-formants are more descriptive than formants, least of all prone to change, and are relatively easy to separate [4]. Non-coherent optical speech processing is limited to linear transforms only. Nonlinear transforms of acoustic signals are readily produced by coherent optics techniques, using the "Kristal" facility with a "Phototitus" modulator. Recognition was performed using the multirange delineation and modified image disfocus methods [2,5].

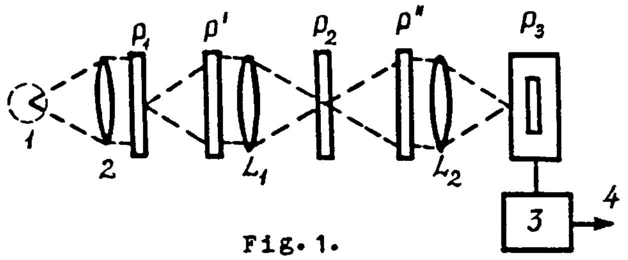


Fig. 1.

Basic non-coherent optical correlator
1 - source; 2 - condenser lens; 3 - decoder; 4 - output signal; P', P" - obscures; P_1, P_2 - mask-transparencies; L_1, L_2 - lenses; P_3 - integral matrix

PRINCIPLES OF ALGORITHM CONSTRUCTION

Delineation of "visible speech" patterns by means of a controlled photoelectrooptical liquid-crystal matrix is based on the photosensitive surface being exposed both to a focused image and defocused image, the former providing a pulse response in the form of a delta function and the latter - in the form

$$1/(R_o^2 \mathrm{ciRc}(\sqrt{x^2 + y^2}/R_o)),$$

where R_o is the defocusing factor. The contour is determined by the difference between these images which is generated during readout. Such processing is analoguous to photography with an "unsharp mask" [3].Generalizing Casasent's transform [2,6] by introducing normalization to time and combining geometric transformations with integrated optical processing provides addressing a considerably wider class of phonem speech decoding problems, in particular by including "visible speech" image recognition when the pattern differs from the reference one in scale, positioning, orientation and time dependence. A multigraph is generated in the COC memory as result of holographic speech signal processing, this multigraph containing various interpretations of the recognized words, syllables and phonems. Studies show the optimal recognition algorithm to correspond to the minimal evaluation by Kolmogorov's intricacy criterion. Some relations, describing associative signs are outlined from the versatile relations class. The effects of actually implemented algorithms on the image being recognized is limited to the screening operator which is in the form of a special mask and which is equivalent in effect to convolution of an associatived sign matrix with a versatile relations matrix. In the intelligent system thus created particular calculus of natural deductions is widely employed. Digital holography was used to design the optimal filter, the initial data being produced by passing the visible speech images through special masks, such as chess field, concentric alternating dark and light bands, moire grid, etc. Computer processing of these prefiltered images produced a program of grid plotting for a precision plotter, with a photo image of this grid reduced by 70X used as an optimal matched filter. The same program was used to control the electron beam path during readout of the recognized visual speech image. Beam deflection was corrected by means of a special associative mask which served as a multiversion prompter. The most probable beam paths were run first with less probable paths following. The artificial intelligence system made wide use of contiguity and hint relations. As compared to frame artificial intelligence systems, this system features the advantages of associative links and a considerably higher version search rate for speech pattern recognition.

FURTHER DEVELOPMENTS

The artificial intelligence system described was run mostly under stringent program control. To make the system more flexible it is expedient to complement its intelligent and customized processors by a so-called instrumental processor. The function of this latter is to generate the CS of variable architecture and structure, depending on the stage of the task being performed. The instrumental processor determines the number of atomic evaluators and their networking into a semantic net to optimize the search of a reference pattern for the image to be recognized and select the most efficient algorithm for the present stage. Thus, the intelligent processor sets the strategy, while the instrumental processor determines the tactics of recognition. Mathematical simulation of both processors utilized Petri nets.

REFERENCES

1. Y.Saito, S.Komatsu, H.Ohzu. Scale and rotation invariant real-time optical correlator using computer-generated hologram. - Optics Communication, 1983, v.47, No.1 pp.8-11.

2. D.Casasent, P.Psaltis. Positional, rotational, and scale invariant optical correlation. - Applied Optics. 1976, v.15 No.7, pp.1795-1800.

3. А.З.Дун, С.Ю. Маркин, Е.С.Невеженко и др. Исследование фотоэлектрического модулятора света в режиме обработки изображений. - Автометрия, 1982, с. 24 - 30, № 2.

4. Trends in speech recognition. Ed.W.Lea. Prentice-Hall,Inc.,Englewood Cliffs, N.J, 1980.

5. О.А.Бутаков, В.И.Островский, И.Л.Фадеев. Обработка изображений на ЭВМ. М., Радио и связь, 1987.

6. С.А.Майоров, Е.Ф.Очин, Ю.Ф.Романов. Оптические аналоговые вычислительные машины. Л., Энергоатомиздат, 1983.

SOME PRINCIPLES OF CONSTRUCTION OF SPEECH RECOGNITION SYSTEM

G.V.KRYUKOV

Dept. of Mechanics
and Mathematics
Moscow State University
Lenin Hill, Moscow
USSR, 119899

A.V.MIKHALEV

Dept. of Mechanics
and Mathematics
Moscow State University
Lenin Hill, Moscow
USSR, 119899

V.N.TRUNIN-DONSKOY

Computer Center of
Academy of Sciences of USSR
Vavilov Street, 40
Moscow, USSR, 117333

ABSTRACT

This paper is based on the theory of fuzzy sets as a mathematical means of description of speech and language. It is suggested to consider the problem of parametric representation and following analysis of speech signals with the purpose of recognizing as a problem of successive transformations from fuzzy subsets to usual ones and vice versa and an analysis of obtaining results at every step.

INTRODUCTION

The idea to use the theory of fuzzy sets for speech recognition was suggested in /1-5/. Partially it was connected with difficulties arising with attempts to use the traditional mathematical methods of description of speech and landuage, particularly when constructing nonadaptive systems of speech recognition, i.e. speaker-independent speech recognition systems for an arbitrary speaker which is a carrier of pronunciation norms for given languge. Experiments of Klatt and Stevens/6/ shows that uncertainty of speech signals in acoustic field is the main propety of speech. At the same time under phonetic decoding linguists are managing without any complicated mathematical means and even without sufficient current information. The studies on blind-spectrogram-reading experiments with various speakers which have been carried out over the years on the faculty of Philology of Moscow State University /7/ and constructions of speech understanding system based on the analysis of phonetician-expert-experience with decoding speech spectrograms confirm our thesis. Using only general lingustic knowledge on nature of formant parameters, on intensity of signals and harmonics expert-liguists make successfully phonetic decoding actually basing on fuzzy linguistic variables (for example, high 1 formant, very low 2 formant, big total energy etc.)

FUZZY TRANSFORMATIONS IN SPEECH RECOGNITION

It has been assumed that elements of thinking of man would enable to imagine as fuzzy subsets for which the function of belonging takes not only 1 or 0 but real numbers beween 0 and 1. In practical systems of recognition it is not necesary on the parametric level to try to find some global decisins and on phonetic level to obtain unique decision on belonging to a definite class, but it is possible to accept several variants of sounds as it is doing in systems for understanding continuous speech when forming of phonetic lattices. Remark that uncertainty on the phonetic level may be setted using higher linguistic levels (lexical, syntactic and semantic). Thus extension of uncertainty of solutions for ill-definable classes of sounds on lower recognition levels may give more possibilities for right and reliable recognition of speech sounds than raising rigidness of decision-making on the same levels.

It is possible to describe speech which is by its nature rather complicated information phenomenon with the help of fuzzy linguistic variables connected with features of prime parameters which in their turn are results of some objective measurings and do not contain fuzzyness. Thus on the stage of prime data reduction of speech signal we have direct transformation from a fuzzy set to an usual one. At the same time labels indicating on belonging of learning set to suitable classes may be fuzzy. It means, that classification object must be going on with taking into account all of the probable classes. Thus on the stage of fuzzy classification we have an inverse transformation from usual subsets to fuzzy ones.

In the case of data reduction of speech signals using the notion of usual α- level subsets $A_\alpha=\{\bar{x}/\mu_{\underset{\sim}{A}}(\bar{x})\geq\alpha\}$,we may formulate the general principle of construction of acoustic-phonetic processor, which consists in search of α-level which controls passage of speech signals throught some key schemes for purpose of subsequent processing. Experimentally α-levels are selected such i)basic noice of apparatus, stationary noice and room noice do not stand out against a background in pauses; ii)there is information on place of formation of weak fricative sounds (f,h) when they appear. As a result we have a direct transformation from the fuzzy subset of analogic speech signals to the usual subset of discrete figure codes corresponding to these analogic signals, i.e. $\underset{\sim}{U}(t)\rightarrow\{A_i^j\}$;where $\underset{\sim}{U}(t)$ is a fuzzy set of analogic speech signals; A_i^j is a set of discrete readings of parametres representing of a speech signal; $i=1,2,\dots M$ is a type of parameter;

$j=1,2,\dots N$ is a number of reading.The fuzzy decision on belonging of these discrete readings to the nearest samples is described in /8/.In the process of work of fuzzy decision algorithm we have the inverse transformation from the fuzzy set of discrete readings of parameters representing of speech signal to an usual subset of hypothesis on pronounced word.

Such transformation may be realized using decomposition theorem /9/,which from the chain of usual subsets $A_0 \subset A_1 \subset A_2 \subset \dots \subset A_{n-1} \subset A_n$ of acoustically similar families of segments of words with phonetic labels defined by parametric matrices give us a fuzzy subset of hypothesis on pronounced word (here $A_0 \rightarrow a_1^{(0)} \dots a_{N_0}^{(0)}$, $A_n \rightarrow a_{1n}^{(n)} \dots \dots a_{N_n}^{(n)}$ are phonetic transcription of acoustically similar words; $a_i^{(j)}$ is a phonetic label of i-th segment and j-th word; n is the word number; N_n is length of phonetic transcription of n-th word.The inclusion relation of acoustically similar words is defined by relative embedding coefficient.

DEFINITION 1.By a relative inclusion coefficient K_e from A_0 in A_e is meant the maximal number of coincidence of phonemes in transcription arranged on the order of appearence: $K_e = K(A_0, A_e) = max(m) | \exists a_{im}^{(e)} = a_{jk}^{(0)},\ a_{i2}^{(e)} = a_{j2}^{(0)}, \dots, a_{im}^{(e)} = a_{jm}^{(0)}$.
$i_1 \leq i_2 \leq \dots \leq i_m;\ j_1 \leq j_2 \leq \dots \leq j_m$

DEFINITION 2.Let $A_{e_1} \subset A_{e_2}$ be meant that $K_{e_1} \geq K_{e_2}$.

The fuzzy subset $\underset{\sim}{A}$ of hypotheses on pronounced word defined as follows: $\underset{\sim}{A} = \underset{\alpha_i}{MAX}[\alpha_1 \cdot A_{\alpha_1}, \alpha_2 \cdot A_{\alpha_2}, \dots, \alpha_n \cdot A_{\alpha_n}]$ (here the meanings of α_i for A_i are such that $\alpha_1 > \alpha_2 > \dots > \alpha_n$).Meanings $\alpha_1 \dots \alpha_n$ have significance of related between speech waveform and nearests samples.

The following transformation (from a fuzzy subsets of hypotheses on pronounced word to a precise belonging of introducted realization to a definite standard) may be realized,for example, with regard for syntax and semantics of languge.

On these grounds we may consider the problem of parametric reprezentation and subsequent analysis of speech recognition as a problem of subsequent transformations from fuzzy subsets to usual ones and vice versa and analysis of results on each step.Such transformations may be described using Galois relations for fuzzy sets /10/. These theses were used when working out of acoustic-phonetic and phonologic processor and hardware-software speech recognition systems.

REFERENSE

/1/R. DE MORI,«Computer Models of Speech Using Fuzzy Algoritms» ,New York,Plenum Press,1983.

/2/P.DEMICHELIS,R.DE MORI,P.LAFACE , M. O KANE,« Computer Recognition of Plosive Sounds Using Contextual Information »,IEEE transactions on acoustics,speech and signal processing,vol.ASSP-31,p.p.359-377, April 1983.

/3/R.DE MORI and G.GIORDANO,« A parser for segmenting continuous speech into pseudo syllabic nucleir»,Proc.IEEE-ICASSP,Denver, Colo.,p.p.876-879,1980.

/4/Э.Б.Донбаев,В.Н.Трунин-Донской, Моделирование систем понимания киргизской речи на ЭВМ ,Изд."ИЛИМ",Фрунзе,1977.

/5/А.Ж.Чымбаев, Предварительная сегментация и маркировка слитной речи ,М.ВЦ АН СССР,1979.

/6/D.KLATT,K.STEVENS,«Sentens Recognition from Visial Examination of Spectrograms and Macineaided Lexical Searching»,IEEE Proc. of 1972 Conference on Speech Communication and Processing.

/7/Н.В.Зиновьева, Особенности интерпретации "слепых" сонограмм при работе с временным окном ,В кн."Тезисы докладов и сообщений 13-й Всесоюзной школы-семинара "Автоматическое распознавание слуховых образов (АРСО-13)",ч.2,Новосибирск,1984,ИМ СО АН СССР,с.30-32.

/8/Г.В.Крюков,А.В.Михалев,В.Н.Трунин-Донской, Относительная глобальная проекция в автоматическом распознавании образов , В кн. УII Всесоюзная конференция "Проблемы теоретической кибернетики",Тезисы докладов, I часть,Иркутск,1985.

/9/А.Кофман, Введение в теорию нечетких множеств ,М.,"Радио и связь",1982.

/10/A.ACHACHE,«Galois connection of fuzzy Subset», Fuzzy Sets and System,8(1982),p.p. 215-218.

SPECTROGRAM READING AND EXPERT METHODS FOR ACOUSTIC-PHONETIC SPEECH SIGNAL DECODING

Nina Zinovyeva

Department of Philology, Moscow State University
Moscow, USSR, 119899

ABSTRACT

This paper presents the results of a large series of experiments in reading spectrograms of Russian utterances. Our experiments have enabled us to reveal the most general principles of human speech behaviour in spectrogram processing and expert acoustic-phonetic decoding strategies. We discuss here these aspects of human expertise and also address the problem of expert knowledge implementation in designing speech recognition algorithm, e.g. the algorithm for segmentation of speech wave into segments corresponding to phonemes.

INTRODUCTION

As it was stated in a recent series of papers /I/ - /5/, the experiments in spectrogram reading demonstrated the richness of phonetic information that can be derived from the most widely used three-dimensional /frequency-time-intensity/ visual display of the signal produced by visible speech spectrographs. It was also pointed that "rules for extracting and interpreting this information can be explicitly formulated" /2/ and thus used to improve the segmentation and labeling performance of present speech recognition systems /3/, /5/. It is also worth mentioning that despite the other-than-auditory modality of speech signal processing, the spectrogram reading is of particular interest as it can provide some useful insights into the human speech perception as such /I/.

Here we report on the results of a long-term investigation in reading spectrograms of Russian utterances. The experiments have been conducted at the Philological Department of the Moscow University since I979. At the very beginning of the research the participants were not skilled spectrogram readers, so one of the goals of our study was to acquire acoustic-phonetic decoding competence in dealing with visual speech signal representation. That was the reason for rather simple experimental tasks set on an early stage of the work and their gradually increasing complexity in the following experiments. It was achieved by using more complex speech units /from isolated words of limited vocabulary and their syntactically and semantically anomalous combinations to nonsense words and nonsense phrases, syllables, extracted from words and phrases and so on/, by increasing the number of speakers /on the whole the utterances of I4 speakers were examined during our research/ and complicating the conditions of the experiments /using different "time windows" with duration ranging from 300 to 50 msec, noising speech signals etc./.

In our investigation we used to read wide-band spectrograms produced on a "Kay Sona-Graph" /Model 7029A/ with different frequency ranges. About 800 spectrograms were analyzed in total. In each experiment were taking part from 3 to 4 human readers, who in course of the research /during first 2-3 years/ mastered the skill of spectrogram acoustic-phonetic interpretation to the highest degree. We shall further refer to them as experts.

The results of acoustic-phonetic decoding achieved in our experiments were as follows: the skilled experts were able to correctly transcribe about 87% of segments with an average I,2I labels produced to each segment in the case of isolated word interpretation and about 83% with an average of I,2I labels for the connected speech.

It would be interesting to compare our results with those obtained on the basis of other languages. It was reported /I/,/3/ that for American English the mean accuracy of labeling ranges from 80% to 90% with an average of I,53 labels to each segment. For French the first measurement is approximated to 85% and the second - to I,5 /4/, /5/.

The comparison of these results makes it clear, that they are very similar in the first measurement, which reflects the accuracy of phonetic interpretation of spectrograms, and differ in the second, reflecting the ambiguity of phonetic decisions. We suppose that this difference is due to the different phonetical and phonological structures of languages under discussion, specifically to the different numbers of vowel phonemes. Vowel segments, highly influenced by the surrounding context, are more ambiguous than consonants, but relatively small set of alternative phonetic labels for vowel identification in Russian decreases the ambiguity of phonetic decisions.

The close examination of our results revealed some other factors, which influence experts' performance in spectrogram reading. This performance depends on the skill level of the expert due to the training period of spectrogram reading, on the type of analyzed speech material /connected speech versus isolated words/, on using short "time window" and on the speaker's specific pronunciation features /foreign accent or speech deficiency/. At the same time the accuracy of acoustic-phonetic decoding practically does not depend on speaker's voice quality.

It is worth mentioning, that dealing with spectrograms the experts did not make precise measurements of spectral parameters, because measurement process increased difficulties in reasoning about spectrograms and tempered the results.

All the facts mentioned above, as well as the close examination of the protocoles provided by the experts and of the tape-recorded discussions which they conducted during some spectrogram reading sessions enabled us to formulate the most general principles of speech spectrogram acoustic-phonetic interpretation.

THE GENERAL PRINCIPLES OF ACOUSTIC-PHONETIC DECODING

We have formulated four most important principles of spectrogram acoustic-phonetic decoding /APhD/. It should be pointed out that we have revealed them from our own experience, but the later analysis of the literature on the problem has shown that practically all of them are somehow mentioned in the papers of other researchers /5/ - /9/. This leads us to conclude that these principles characterize experts' speech behaviour as such, unrelated to the language structure and perhaps even to the speech perception modality.

I. The phonetic identification can't be deduced immediately from the continuous acoustic-parametric representation of the signal. There exists an intermediate level of speech signal processing which serves as a kind of bridge across the representational gap between acoustic substance and it's underlying phonetic form /8/. The acoustic information on this level is described in the most compact and abstract manner, without absolute numerical measurements of spectral parameters. Such qualitative descriptions suppose the detection from spectrogram the most important and closest to phonetic categories acoustic properties, free from the signal variability due to extralinguistical sources. We believe that the training period for spectrogram reading is mainly connected with evolving in expert's mind this specific interface device for an other-than-auditory modality.

Nowadays almost all researchers in the field assent to the idea of existence of this specific intermediate representational level in speech processing. The units of this level are called acoustic cues or descriptors /I/ - /6/. Our attempt to sketch the system of these units is represented in the section below.

2. The APhD is a highly active process combining two processing directions: bottom-up and top-down. It means that the lower speech representation level is analyzed from the point of view of higher level units. The acoustic-parametric representation is judged by the set of acoustic cues existing in expert's mind. The interpretation of these acoustic cues is the result of producing and gradual decreasing of a number of phonetic hypothesis. That is why the results of spectrogram reading depend in particular on the number of units in different phonetic classes.

3. The general procedure of APhD is divided in two separate stages: a) segmentation by partial sound specification corresponding to the manner-of-articulation categories and b) identification of the place-of-articulation features /in larger sense including the front-back and high-low qualities of vowels/. It should be pointed out that segmentation does not precede labeling but is conducted by partial recognition of sound stretch. At the same time segmentation preceeds full phonetic identification because contextual evidences are used to determine the place-of-articulation features. At this stage the segment boundary placement may be refined but in any case segmentation is resulted from recognition and includes various procedures for detecting and extracting from the sound wave groups of sounds different in manner of articulation or the so called "broad phonetic classes" /I/, /3/.

4. The APhD is highly context-dependent process. The use of contextual evidences in transformation of acoustic cues into phonetic categories is obvious and generally acknowledged. But we believe that contextual information is important as well in measurement-to-descriptor mapping. The experts are not aware of this because they reach an intermediate speech processing level by unconscious mechanisms of human visual system. But our current experiments with digital representation of the signal have demonstrated the significance of contextual information at the very first stages of APhD.

THE SYSTEM OF ACOUSTIC CUES FOR PHONETIC UNITS RECOGNITION

In the present section we attempt to sketch the inner structure of the acoustic cue level /AC-level/ and it's relations with the preceding and subsequent levels of speech signal processing. We suppose that this structure reflects the expert APhD strategy.

In examination of the spectrograms the experts proved to be highly efficient in detecting some "primitive visual objects" or PVO /7/ that serve as the basis for AC-descriptions. The acoustic cues carry information about presence/absence of PVO, their sequential relations, and about duration, intensity and frequency modifications of PVO. The term "modification" in our case means some qualitative /not quantitative/ characteristics of PVO, described as being "high/mid/low", "long/short", "strong/weak" /3/, while relative characteristics are described in terms of "higher/lower", "stronger/weaker" and "longer/shorter". The information about PVO's changes through the time, reflecting in their spectral trajectories, is also used to achieve AC-descriptions.

The AC-level in turn can be roughly divided into three sublevels: AC-I, AC-2, AC-3. These sublevels are differentiated according to their place in the whole analysis and the relative degree of approximation to the preceding and subsequent levels /fig. I/. Thus units of AC-I are closer to the acoustic-parametric representation, while AC-3 units are closer to the phonetic level. The AC-I analysis is practically independent of the phonetic structure of a certain language, but is involved in speech/non-speech detection of the acoustic signal. On the contrary, the AC-3 analysis highly depends on the phonetic system of a language. In this respect the AC-2 level is in an intermediate position.

The aforementioned sublevels are interconnected because higher level descriptions as a rule incorporate units of the preceding level. In addition, within each sublevel AC-units differ in their physical nature /duration, intensity or frequency/.

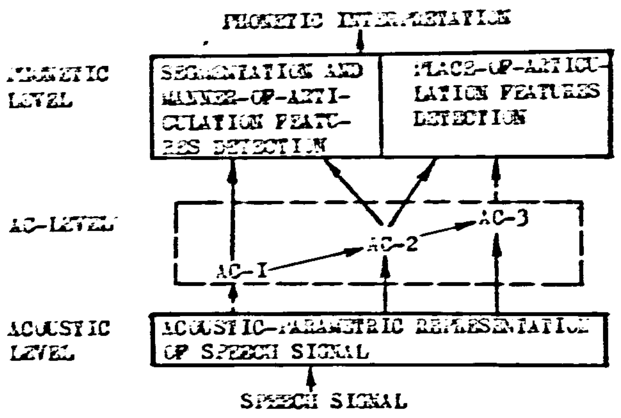

PHONETIC INTERPRETATION
PHONETIC LEVEL
SEGMENTATION AND MANNER-OF-ARTICULATION FEATURES DETECTION
PLACE-OF-ARTICULATION FEATURES DETECTION
AC-LEVEL
AC-I
AC-2
AC-3
ACOUSTIC LEVEL
ACOUSTIC-PARAMETRIC REPRESENTATION OF SPEECH SIGNAL
SPEECH SIGNAL

Fig. I
THE GENERAL SCHEME OF THE APhD
We have marked with broken line those components which we believe to be of paramount importance and complexity in view of speech recognition.

The presence/absence information about PVO and their combinations constitutes the very first layer of AC-I. Using these cues the speech wave is splitted into primary subphonetic segments /PSS/ such as "voiced closure", "unvoiced closure", "voiced noise", "unvoiced noise" and "vocal segment" which incorporates both vowels and sonorant consonants. AC-I also contains the sequential cues which make it possible to combine some non-vocal PSS /such as closures and following them noises/ into aggregate segments. Further on the durational AC-2 are used on these segments. In this case durational AC-2 presuppose relative estimation of lengths of PSS in terms of "longer/shorter". This procedure enables to identify stops, affricates and stop-fricative clusters with different place of articulation of their components /such as [ks] /.

Vocal segments, including vowels and sonorants, are estimated by intensity and frequency AC-2. These acoustic cues are rather multiple as a result of the necessity to consider different positions of sonorants within the vocal segments and different frequency locations of the formants due to the nazal/non-nazal sonorant discrimination. At the same stage the duration AC-2 are used to detect vibrants. In this case the AC-2 are formed according to the expert knowledge of minimal possible duration of the PSS.

The procedure described above results in extracting segments corresponding to the manner-of-articulation categories. All the obtained pieces of speech wave are estimated by durational AC-2, detecting segments that are likely to contain more than one phone of the same manner of articulation /e.g. [lm] /. For these segments /and for the segments corresponding to stop-fricative clusters as well/ boundary placement decisions are improved on higher levels.

The first stage of expertise concerning segmentation by identification manner-of-articulation categories doesn't comprise any significant difficulty for the experts /the accuraey of segmentation is about 98% correctly placed boundary markers/. But in order to have the computer mimic the segmentation performance of the experts, we need to evolve some special device for automatic extraction of the PVO. This problem is very difficult to solve. It lies beyond our competence /and beyond the field of phonetics/ as it is related to the mechanisms of human visual object recognition.

A more complex intellectual task for the spectrogram readers is to derive place-of-articulation values from the AC descriptions. To achieve this aim they at first use intensity and frequency AC-2, describing contextually independent modifications of consonants /mainly of their noise components/. But very often it is rather difficult and sometimes even impossible to specify accurately the place of articulation for consonants only by their intrinsic characteristics. In such cases the experts use AC-3 descriptions, carrying information about PVO's /mainly vowel formants'/ changes through the time. The AC-3 analysis is highly contextual and presupposes parallel inference of both consonant and vowel phonetic specification, based on the same acoustic evidences.

Formant trajectories are described qualitatively in terms of movement directions, relative spectral locations and frequency ranges, shape and slope. The phonetic decisions are deduced according to the expert knowledge of acoustic manifestations of the coarticulation processes in Russian which make it possible to interpret formant trajectories in terms of implied acoustic targets, closely connected with place characteristics.

To implement the rules of AC-3 inference in designing speech recognition systems it is necessary not only to reveal and formulate all the knowledge items of this level, which is not minor problem in itself, but also to describe their possible and impossible combinations, resulting in different confidence values of phonetic decisions. Besides, it is necessary to evolve quantitative methods to create qualitative descriptions, used by the experts. So, the conversation of AC-3 units into phonetic features is the second most difficult step of APhD to mimic it in the computer programs.

IMPLEMENTATION OF THE EXPERT APhD PRINCIPLES IN ALGORITHM FOR SPEECH SIGNAL SEGMENTATION

At present we devote our efforts to adapt the spectrogram reading methods to digital representation of the signal which can serve as an input into a speech recognizer. To bypass the problem of automatic extraction of the PVO, we were to select suitable digital representation for indirect interpretation in terms of AC-I. We came to the conclusion that frequency band analysis creates appropriate representation for this purpose because the information about energy balance in frequency bands can be qualified as a result of the PVO being present or absent. It proved impossible to select uniformal for all speakers frequency range division into bands, but we believe that this problem can be solved by evolving a rather simple adaptation procedure to define frequency band boundaries according to the speaker's voice quality.

In general the selection of frequency bands depends on the ranges of functioning of acoustic objects corresponding to the PVO /"voice", "FI", "FII", "FIII", "low velar noise", "mid alveolar noise" and "high dental noise"/. It is evident that in this case frequency bands would overlap because of the overlapping functional ranges of the above-mentioned acoustic objects /e.g. frequency band distribution for one of the speakers: up to 300 Hz, 200-900 Hz, 500-3500 Hz, 3500-7000 Hz/.

Using the achieved digital representation /i.e. a series of parameter vectors, reflecting frequency band energy concentrations, one every centisecond/, we have designed a segmentation algorithm, implementing all the expert APhD principles described above. The algorithm consists of different procedures /=sets of rules/ for detecting different in manner of articulation groups of phones /segmentation by recognition/. Each procedure performs goal-directed search for those pieces of speech wave that are consistent with the preconditions of its rules /active search/. The procedures include rules of interpretation of frequency band information in terms of AC-I /intermediate representation/. The algorithm doesn't perform centisecond phonetic labeling but interprets each parameter vector with respect to adjacent vectors /contextual analysis/.

We'll dwell on the last principle in more detail. The expert analysis of initial digital representation revealed that vast majority of centisecond parameter vectors were characterized by high phonetic ambiguity which can be decreased significantly by taking into consideration values of the adjacent vectors. In this case very similar /or even identical/ vectors can acquire different labels while quite different ones can get identical labeling depending on context information.

To simulate context-dependent interpretation of parameter vectors performed by the experts we have introduced the notions of centers of PSS and their periphery considering centers to be the most prominent and distinct vectors to interpret them in terms of AC-I. At first the sound string is analysed for centers which later guide the search for boundaries of PSS. The detected centers of the PSS and their readings in terms of AC-I serve as a context for the interpretation of their environments. It means that the reading given to the center is extended to all adjacent vectors /both preceding and subsequent/ untill they are consistent with the preconditions of the periphery-detecting rules which are significantly weaker than those of the center-detecting rules.

The whole procedure results in extracting PSS, which are later combined into segments corresponding to phonemes, as it was described above. The search for sonorants within vocal PSS is also organized as center-guided process. This enabled us to design various procedures for sonorant identification in pre-center, inter-center and post-center positions.

The algorithm is hierarchically organized according to confidence of inferences of the procedures included. The more confident procedures are activated earlier, narrowing the search area for succeding procedures.

Our work on the algorithm proved the spectrogram-reading approach to be very promissing and productive in view of speech recognition.

REFERENCES

/1/ 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.A. Cole, ed., Lawrence Erl. Ass.,1980, pp.3-50.
/2/ M.A.Bush, G.E.Kopec, V.W.Zue, "Selecting acoustic features for stop consonant identification", Proc. ICASSP-83, pp. 742-745.
/3/ V.W.Zue, L.F.Lamel, "An expert spectrogram reader: a knowledge-based approach to speech recognition", Proc. ICASSP-86, pp. 23.2.I-4.
/4/ N.Carbonell, D.Fohr, et al., "An expert system for the automatic reading of French spectrograms", Proc. ICASSP-84, pp. 42.8.I-4.
/5/ N.Carbonell, J.-P.Damestoy, et al.,"APHODEX, Design and implementation of an acoustic-phonetic decoding expert system", Proc. ICASSP-86, pp. 23.3.I-4.
/6/ M.Liberman, "On the role of phonetic structure in automatic speech recognition", Proc. X-th ICPhS: Plenary Sessions. Symposia. Utrecht, 1983, pp. 315-318.
/7/ J.Johannsen, T.MacAllister, et al.,"A speech spectrogram expert", Proc. ICASSP-83, pp. 746-749.
/8/ P.D.Green, A.R.Wood, "A representational approach to knowledge-based acoustic-phonetic processing in speech recognition", Proc. ICASSP-86, pp. 23.4.I-4.
/9/ J.Caelen, N.Vigouroux, "Producing and organizing phonetic knowledge from acoustic facts in multi-level data", Proc. ICASSP-86, pp. 23.5.I-4.

МОДЕЛИРОВАНИЕ ДИНАМИЧЕСКОГО ВЗАИМОДЕЙСТВИЯ РЕЧЕВОГО СИГНАЛА С ПАМЯТЬЮ В СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

Г.Н.Вандакурова, Р.Я.Гумецкий, Л.А.Мелень

Кафедра биофизики и математических методов в биологии
Львовский государственный университет, Львов, СССР, 290000

АННОТАЦИЯ

Распознавание речевого сообщения рассматривается как целенаправленный динамический процесс взаимодействия ассоциативно организованной памяти с речевым сигналом, позволяющий генерировать гипотезы о каждом из элементов сообщения, которые проверяются на соответствие воспринимаемым акустическим образам. При этом максимально используется информация многоуровневой памяти распознающей системы, а ее рецептивные возможности задействуются на глубину, достаточную для однозначной смысловой интерпретации речевого сообщения.

ВВЕДЕНИЕ

В естественной речи многие из номинальных элементов сообщения не могут быть непосредственно восприняты, так как они зачастую недостаточно четко реализованы или даже редуцированы в произнесенной фразе.Решения о них могут появляться при восприятии речи только ретроспективно, после того, как сообщение становится осмысленно понятым. Поэтому в плане выполнения функции восприятия речи принципиально важная особенность распознающей системы заключается в способности текущего предсказания возможных в данной ситуации элементов сообщения, которое уточняется до непротиворечивости (соответствия) воспринимаемому речевому сигналу [I].

Методологическая основа рассматриваемого подхода к моделированию восприятия речи основывается на предположении, что многоуровневый процесс распознавания речевых сообщений заключается не столько в процедурах таксономии и классификации конкретных акустических образов, сколько в целенаправленной динамической актуализации ими языковой памяти распознающей системы, что в конечном итоге позволяет однозначно интерпретировать сообщение на основе переданных в речевом сигнале смыслоразличительных признаков, достаточных для осуществления этой цели. Наряду с выявлением таких информативных признаков, в этом процессе основополагающую роль играют механизмы дедукции, включая активный поиск аргументов для оценки соответствия воспринимаемых акустических образов предполагаемым на каждом шаге анализа элементам сообщения [2].

С этой точки зрения процедуры активного взаимодействия между памятью, организованной с учетом лингвистических и иных ассоциаций, и специфическими механизмами обработки речевого сигнала не могут игнорироваться и должны быть введены в общую структуру модели распознавания речи. Возможные алгоритмы взаимодействия речевого сигнала с памятью распознающей системы исследовались в настоящей работе.

ОРГАНИЗАЦИЯ МОДЕЛИ

Описанная ниже модель основывается на дедуктивном подходе к распознаванию речи [I,4]. В соответствии с его основными положениями, параллельно с процессом обработки воспринимаемого речевого сигнала текущим образом генерируется обоснованные в данной ситуации гипотезы о распознаваемом речевом образе, которые верифицируются путем сопоставления предполагаемой фонемной структуры с ее акустической реализацией в произнесенном сообщении. При этом моделируются три принципиально различных вида перцептивной активности, определенные как:

I/ мыслительная активность (выдвижение словесно сформулированных гипотез в соответствии с лексиконом и закономерностями используемого для речевой коммуникации языка), управляемая результатами распознавания сообщения;

2/ слуховая активность (обработка и представление речевого сигнала в виде динамики обнаруживаемых в нем фонетических качеств, подготавливающие его к сопоставлению с фонемными матрицами слов, выдаваемых на данном отрезке сообщения алгоритмом предсказания);

3/ активность принятия решения (верификация предсказываемых фонемных последовательностей по результатам фонетической интерпретации речевого сигнала),в результате которой определяется максимально правдоподобная фонемограмма высказывания [I].

В распознающей системе на используемый словарь всевозможных слов, находящихся в ее памяти (лексикон), априорно действует или текущим образом накладывается целый ряд ограничивающих факторов, которые постоянно задают и перестраивают подмножества слов-гипотез, актуальных на каждом шаге анализа конкретного сообщения. Ограничения, которые позволяют упорядочить (при обучении) и целесообразно задействовать память при распознавании речи, могут задаваться как изнутри самой системы, так и из внешних источников. В первом случае имеет место воздействие на процесс восприятия речи внутренней структуры используемого проблемно-ориентированного языка (лексики, синтаксиса, семантики). Внешние источники информации имеют место, когда известна ситуация речевого общения (тематическая, контекстуальная, в т.ч. и текущие результаты распознавания сообщения).

Таким образом, на основе взаимодействия априорной информации языковой памяти с текущей информацией воспринимаемого сигнала формируется результат распознавания пословно предсказываемого сообщения как последовательность слов, наилучшим образом соответствующая фонетической структуре высказывания. Сама мера подобия речевого сигнала эталонам верифицируемых элементов речи (слов) может быть оценена с помощью различных решающих правил и мер близости.

АЛГОРИТМИЧЕСКИЕ РЕШЕНИЯ

В моделируемом алгоритме распознавания речевого сообщения максимально используется априорная информация, заданная в памяти воспринимающей системы, а рецептивные возможности системы задействуются при анализе определенного участка речевого сигнала в той мере, в какой это необходимо для однозначного декодирования соответствующего лексического элемента сообщения.

На каждом шаге распознавания текущее подмножество актуальных лексических гипотез из заданного словаря определяется на основе пересечения подмножеств слов, которые:

I/ задаются априорно вводимыми прагматическими и лингвистическими ограничениями на высших уровнях системы;

2/ определяются текущим образом, исходя из лингвистических ограничений, на основе решений о предшествующих отрезках речи;

3/ определяются по результатам оценки интегральных надежно регистрируемых характеристик очередного участка речи на низших ступенях его анализа.

Для реализации последнего в модель введена динамическая организация памяти по свойствам речевых сигналов, которые не являются традиционными свойствами лингвистической категоризации. В этом плане испытывалась ассоциативная группировка слов (осуществляемая параллельно с эталонизацией словаря) по близости их интегральных акустических характеристик, типичный пример которой рассмотрен в работе [3].

На этапе обучения системы для разбиения заданного словаря на подмножества близких по акустическому образу слов могут использоваться процедуры кластеризации обучающих реализаций слов по различным - структурным, просодическим, маркирующим и другим надежно регистрируемым характеристикам их первичного описания. Обнаружение соответствующих характеристик в акустической реализации распознаваемого слова существенно уменьшает область поиска решения. Алгоритм группового распознавания слов, который испытывался в составе описываемой модели, рассмотрен в работе [3].

В результате наложения всех указанных выше ограничивающих факторов на выходе памяти формируется осмысленное, акустически и лингвистически обоснованное предсказание для каждого лексического (следовательно и фонемного) элемента конкретного сообщения, неоднозначность которого (если она имеет место) разрешается оценкой соответствия эталонных описаний слов-конкурентов фонетической структуре воспринимаемого речевого сигнала. Таким образом, составом актуальной части лексической памяти на каждом шаге восприятия сообщения постоянно управляют как первые ступени, так и конечный результат текущего анализа очередного отрезка речи.

Алгоритм принятия решения об элементах речевого сообщения состоит в том, что поступающие из памяти эталоны актуальных слов-гипотез в нужный момент сопоставляются с очередным отрезком речевого сигнала. Существенно, что при этом происходит предварительная унификация описаний речевого сигнала и эталонов с использованием для этой цели многомерного пространства надежно различимых фонетических качеств, достаточных для сегментного представления речи. Последняя в одном из вариантов модели проводилась на основе двоичного структурного описания речи, предложенного в работе [5] и позволяющего проводить нелинейное во времени сопоставление без использования процедур ДП-согласования.

ЭКСПЕРИМЕНТАЛЬНАЯ РЕАЛИЗАЦИЯ МОДЕЛИ

В диалоговом режиме с использованием ограниченных сменных словарей моделировались типичные задачи распознавания фраз-команд в автоматизированных системах управления (речевое управление вычислительной машиной, речевой запрос в ИПС и другие ситуации). Аппаратурно-программные средства реализованных экспериментальных систем включали:

а/ устройство выделения и ввода речевых признаков в ЭВМ (УВРП);

б/ базовый комплект устройств мини-ЭВМ (УВК СМ-4);

в/ программное обеспечение речевого управления (ПОРУ).

Исходным дискретным описанием речевого сигнала, вводимым в ЭВМ, являлась последовательность выдаваемых УВРП каждые 20 мс векторов качественных и количественных речевых признаков, отсбражающих акустические характеристики источника, способа и места образования звуков в речевом тракте человека. Вторичное структурное описание речи определялось наличием-отсутствием в речевом сигнале конкретной последовательности сегментов с различным фонетическим качеством из потенциально возможной последовательности сегментов всех различимых типов. Переход от временного описания к структурному осуществлялся с одновременным определением наличия пауз - границ слов.

Алгоритмы, разработанные для речевого обращения к ЭВМ, были реализованы в ПОРУ в виде комплекса: I/ универсальных управляющих программ, которые обеспечивают работу системы в режимах обучения (эталонизации, группировки произносимых слов используемого словаря) и распознавания речи (с выполнением воспринятых речевых команд), и 2/ ряда общих подпрограмм (ввода речи, формирования структурного описания, вывода текущих результатов и др.). Связи между группами слов, которые могут встретиться в определенных местах фраз или ассоциируются по акустическому подобию, представлялись в памяти в виде графов, задаваемых списочными структурами.

В режиме распознавания программно реализованы и опробованы:

а/ структурный метод распознавания речи на основе бинарного признакового описания речевых сигналов и хеминговой меры их близости к эталонам слов из памяти системы;

б/ метод распознавания сообщений по полному признаковому описанию речевых сигналов с ДП-согласованием в качестве меры их сходства с эталонами.

. Исследовалась работоспособность распознающих моделей при объеме экспериментальных словарей до I00 слов. В условиях шумов до 75 дБ оценена надежность распознавания речевых команд различных операторов, которая при заданных ограничениях составляет 90±3% для лиц, имеющих опыт работы с системой.

ВЫВОДЫ

Результаты исследования рассмотренных алгоритмов взаимодействия параллельных потоков актуальной информации из памяти с потоком текущей информации речевого сигнала, полученные в частных моделях, подтверждают эффективность описанного подхода к распознаванию речевых сообщений формализованных проблемно-ориентированных языков. В теоретическом аспекте - показана возможность реализации разработанных алгоритмов в виде целостной иерархической модели восприятия речи.

Следует особо подчеркнуть значение целенаправленной активности в работе всех уровней и систем, принимающих участие в процессе восприятия речи. Это означает, что и все задействованные в распознающей модели блоки неизбежно должны находиться под воздействием управляющих факторов, наиболее мощным из которых является фактор языка, направляющий восприятие речи. Учет всего многообразия факторов, как внутренних, так и внешних, в том числе и самой речи, которые управляют корректностью и адекватностью упреждающей активности распознающей системы, представляется нам как феномен ее интеллектуализации. Именно в таком плане следует понимать значение предварительной организации памяти распознающей системы и ее динамического взаимодействия с распознаваемым сигналом, вне которого осмысленное восприятие, т.е. понимание речи автоматом представляется невозможным.

ЛИТЕРАТУРА

[I] Derkach M. Deductive approach to automatic recognition of russian spoken sentences // Proc. ICASSP, 1980. - pp. I04I-I044.

[2] Динамические спектры речевых сигналов / М.Ф.Деркач, Р.Я.Гумецкий, Б.М.Гура, М.Е.Чабан; под ред. М.Ф.Деркача. - Львов: Вища школа. Изд-во при Львов. ун-те, I983. - I68 с.

[3] Гумецкий Р.Я., Вандакурова Г.Н., и др. Двухступенчатый алгоритм распознавания слов, использующий информацию о совокупности и последовательности акуст.-фонет. признаков в речевом сигнале // Вест.Львов.ун-та, сер.биол.,I983.С.I05

[4] Вандакурова Г.Н., Гумецкий Р.Я, Мелень Л.А. К разработке систем распознавания речевых команд из ограниченных сменных словарей // Тез.докл. Всесоюз. шк.-семинара АРСО-I3. Новосибирск, I984. - С. I26-I27.

[5] Гумецкий Р.Я., Гура Б.М., Мишин Л.Н. Использование двоичного структурного описания речевого сигнала в задачах распознавания речи // Там же. - С.I36-I37.

THE PHONEMOPHONE TEXT-TO-SPEECH SYSTEM

Boris LOBANOV

Institute of Telecommunication
Leninsky pr., 113-20, MINSK, 220023, USSR

ABSTRACT

The paper report the results of the theoretical and experimental studies aimed at designing a universal text-to-speech synthesis model covering both the full range of intralanguage phenomena and applicable for multilanguage synthesis. The overall algorithm used in the text-to-speech synthesizer "Phonemophone" is described.

INTRODUCTION

The problem of converting a text into speech signal has been approached by several authors through the application of algorithmic synthesis or synthesis by rules [1,2] . A long list of rules and exceptions would be generally used, their number sometimes going up to thousand.

The main concern of the present study is to reduce the number of rules to a limited set of generalized categories capable of covering all the significant intralanguage phenomena and applicable at the same time to various languages. It is hoped that the lingual-acoustical model of the present work meets the above requirements.

The problem of speech synthesis can be split into two comparatively independent subproblems related to adequate synthesis of phonemic and prosodic structures of the text. At the acoustical level these subproblems correspond to the tasks of synthesizing the current formant parameters, on the one hand, and the current values of fundamental frequency, duration and intensity, on the other.

The present model of speech synthesis from text is based on two principles:

1. A limited set of acoustic invariant structures called portraits of phonemes and prosodems is used. They help to describe all the linguistically significant units of the phonemic and prosodic structures of the text.
2. A limited set of algorithmic rules is used to transform the portraits of phonemes and prosodemes into current acoustic features of running speech.

The exact number and the types of phoneme and prosodeme portraits are determined by available linguistic information about a given language. For instance, English phonemic units must be described by 20 vowel and 24 consonant portraits, in Russian 6 portraits of vowels and 36 portraits of consonants are required.

The exact number as well as the types of transformation rules for the portraits of phonemes and prosodemes rely on the up-to-date data in the fields of experimental phonetics, speech production and speech perception. Thus, for example,the number of transformation rules for the phoneme portraits must take account of the well-known effects of soundscoarticulation, reduction and assimilation.

1. FORMANT PORTRAITS OF PHONEMES

Phoneme and formant are fundamental notions of speech synthesis from text. A phoneme is an elementary and meaningful unit for any texts recording. The problem of transferring a written text into a phonemic one has already been algorithmically resolved for a number of languages and now doesn't present any difficulty. A formant, rather a formant parameter,is a universally unit for acoustic synthesis of any language sounds. A modern formant synthesizer can ensure the quality of sounds very near to natural.

Cur model of speech synthesis employs the following set of operating formant parameters:

F1,F2,F3,Ff - frequencies of three voice formants and the generalized frequency of fricative formants (F-parameters); Av,An, Aa,Af - amplitudes of voice, nasal,aspirative and fricative formants (A-parameters). This set corresponds with the parallel-consequentive design of the formant synthesizer of speech signals.

The formant portrait of a phoneme is built on the 5 consequentive time segments: 0 - introductory, 1 - basic,2-3 - additional and 4 - the final segments In the phoneme portrait T0=T4=0, T1 always exceed zero, whereas T2, T3 can be equal or different from zero. Certain formant

values of F- and A-parameters are given for each time segment. F-parameters are given for the 0-3 segments by three values of F, α, τ where F is inherent formant frequency, α - coarticulation coefficient, τ - formant transition duration. At 4th segment parameters are set at a value of τ^f only; A-parameters are set at segments 1-3 of values of A, τ^A and at segment 4 - by a value of τ^A only.

Thus, each phoneme portrait is described by 83 formant properties.

The portrait is graphically presented in Fig. 1.

The values of formant properties are established experimentally by analysing the behaviour of phonemes in natural speech. The minimal requirements to the experimental material are the following:

- each consonant is to precede and follow each vowel, as well as a pause;
- all the material is to be recorded by the same speaker.

Formant parameters and their properties are obtained by examining the sonograms of speech signal. The process of experimental analysis includes the phoneme fragments segmentation procedure, the normalization of the measured formant parameters, the determination of the inherent values of frequencies and coarticulation coefficients, the measurement of the formant transition duration.

As a result of the investigation phoneme portraits for Russian, Byelorussian, Ukrainian, Bulgarian, English, German and French were obtained and those later on proved valid in building the polylanguage system of speech synthesis.

2. TRANSFORMATION RULES FOR PHONEME PORTRAITS

The rules transforming the phoneme portraits into current values of formant frequencies are based on modelling allophonic variation in natural speech.

The major reasons of phonemes acoustic variation in connected speech are those of articulatory effects caused by coarticulation, reduction and assimilation.

Let's consider one of the most essential components of phonemes modification - that of coarticulation. [3] describes the model of coarticulation at the acoustic level. It has been shown in CV-syllable the formant frequency of the consonant F^{cv} can be expressed by inherent frequencies of the consonant F^c and of the vowel F^v by the equation:

$$F^{cv} = F^v + (1 - \alpha^c) F^c \qquad (1)$$

where $0 \leq \alpha^c \leq 1$ is the consonant coarticulation coefficient. It is easy to show that the inherent consonant frequency and the coarticulation coefficient have the geometrical essense of consonant "focus" coordinates.

Fig. 2 illustrates the trajectories of formant frequency variation F2 (continuous lines) for the consonant /p/ and /t/ within syllables (PV, PA, PI).

The dotted lines indicate their continuation along the pause of the consonant with the point of intersection at the "focus" (point "a"). From the similarity of the triangles abc and cde it follows that

$$F^{cv} = \frac{\Delta 1}{\Delta 1 + \Delta 2} F^v + \left(1 - \frac{\Delta 1}{\Delta 1 + \Delta 2}\right)\varphi \qquad (2)$$

From equation (2) with (1) follows that the $\varphi = F^c$ and $\Delta 1 / (\Delta 1 + \Delta 2) = \alpha^c$.

Let's consider a more general case of a syllable containing more than one consonant, i.e. of the type C2 C1 V0, C3 C2 C1 V0 and the like. Spectrographic examination reveals in this case the dependence of the consonant formant frequency C2 not only on the vowel frequency V0 but on the consonant frequency C1. By analogy consonant formant frequency C3 is dependent on the frequencies C2, C1, V0 and so on.

To take this phenomenon into account the following recurrent equation was applied:

$$\begin{cases} F^{(1)cv} = \alpha^{(1)c} \cdot F^{(0)v} + (1 - \alpha^{(1)c}) \cdot F^{(1)c} \\ F^{(2)cv} = \alpha^{(2)c} \cdot F^{(1)cv} + (1 - \alpha^{(2)c}) \cdot F^{(2)c} \\ \vdots \\ F^{(n)cv} = \alpha^{(n)c} \cdot F^{(n-1)cv} + (1 - \alpha^{(n)c}) \cdot F^{(n)c} \end{cases} \qquad (3)$$

In the formula (3) the top indexes (n) denote the number of a consonant that comes in succession in a syllable beginning with a vowel marked (0).

Some coarticulation effects are also observed in vowel formant frequencies. For instance, in a particular environment F2 of vowels is considerably increased in the position before dental consonants and is reduced in bilabial environment. Modifications of vowel formant frequencies are calculated from the formula:

$$F^{vc} = \alpha^v (\gamma_1 F^{c1} + \gamma_2 F^{c2}) + (1 - \alpha^v) F^v , \qquad (4)$$

where Fv, α^v are the inherent frequency and the coarticulation coefficient of a vowel phoneme; F^{c1}, F^{c2} - intrinsic frequencies of the adjacent consonants (both left and right); γ_1, γ_2 - weight factor. Algorithmic modification rules for the portraits of consonants and vowels affected by coarticulation are based on formulas (3), (4). The properties of Fc, α^c, Fv, α^v are taken from the tables of phoneme portraits. The phoneme portraits also carry the information required to simulate the effects of sound reduction and assimilation.

3. PROSODIC PORTRAITS

Speech prosodic features are intended as a means of realising suprasegmental linguistic phenomena, those of stress and intonation in particular. In the proposed model speech prosodic units are hierarchically arranged in the following succession: syllable, word, accentual group, syntagm, phrase and, finally, phrasal unity (phonetic paragraph). Syntagm is the smallest independent prosodic unity. It is assumed that a limited set of syntagm portraits that can allow a relatively accurate description of speech prosodic features for reading any text can be established. The number of prosodic portraits is selected with a view to the need of realising such linguistic phenomena as communicative intention (statement or question, and the like), syntactical peculiarities (enumeration, contrast, etc), modality or emotion (command, surprise, etc). The number of prosodic portraits in a given language may go to several scores.

A syntagm prosodic portrait is established by a consequentive combination of prosodic portraits of the accentual groups, which fall into three types: initial, intermediate and final. These categories of accentual groups are unequal in their importance for the construction of the syntagm contour portrait. Major variety of contours is associated with the final accentual group carrying syntagmatic stress considerably smaller variety is observed within initial or intermediate accentual groups.

An accentual group corresponds to one or several words in a syntagm that are united into a single prosodic (pitch, rhythmic or dynamic) contour bound to one stress. When the number of accentual groups is the same as the number of words in a syntagm and an accentual group is identical to a phonetic word. But an accentual group may consist of more than one word, and in this case one of the words is made more prominent. The prominence of a word within the accentual group involves prominence of its word stress. Other words in an accentual group are less prominent and receive weak stress. It is more common that the main stress falls on the first word of the accentual group. This fact allows to define a very simple rule of determining the boundaries of accentual groups: the left boundary coincides with the beginning of the word having the main stress, the [illegible] one comes to the beginning of the word of the next accentual group.

Pitch, rhythm and energy of the accentual group are established by the normalized values of fundamental frequency, duration and intensity over the three normalized time segments. They are nucleus, prenucleus and postnucleus. The nucleus of the accentual group is always the stressed syllable marked as the main stress. So the prenucleus and postnucleus are other phonemes of the accentual group that precede or follow the nucleus. The prosodic portraits of accentual groups are established with the help of tables of numbers having from two to four marked intervals over every time segment. Fig. 3 is an illustration of the pitch component of the Russian prosodic portraits of the final accentual group for the six intonational types (statement, question, exclamation, enumeration, contrast, parenthesis). Pitch curves are outlined with the help of normalized coordinates "time - frequency". The normalized time interval (0-1/3) is a correlate of the prenucleus, (1/3-2/3) - of the nucleus, and the interval (2/3-1) is that of the postnucleus. The interval of the normalized fundamental frequency (0-1/3) indicates its low pitch level, (1/3-2/3) - the middle, and (2/3-1) - its high pitch level.

4. RULES OF PROSODIC PORTRAITS TRANSFORMATION

Pitch, rhythmic and dynamic features of accentual groups appear to be the minimal units that make up the intonation pattern of a syntagm, phrase and a text. Mention has already been made that the prosodic portrait of an accentual group is determined for the three normalized time segments - prenucleus, nucleus and postnucleus.

Within a natural text these segments may comprise a different number of phonemes, the latter in their turn may be of different duration in natural speech. Rules of prosodic portraits transformation are built with a view to these factors.

An example of pitch portrait transformation for the phrase 'Do it on Saturday' is presented in Fig. 4.

When an accentual group has no prenuclear or postnuclear part their function is attributed respectively to the left or right fragment of the nuclear vowel (1/4 of its duration). In case a syntagm includes more than 3 accentual groups the required number intermediate groups is added, and if the number of accentual groups is smaller than 3 the intermediate and secondly the initial accentual group are cut off. With regard to the position of a syntagm within a phrase or phrasal unity the portraits of accentual groups are subjected to further transformation that mainly consists in proportional modification of their scale.

5. TRANSFORMATION ALGORITHM TEXT-TO-SPEECH

The above described rules of establishing phonemic and prosodic portraits as well as the rules of transforming the current values of formant parameters from the ta-

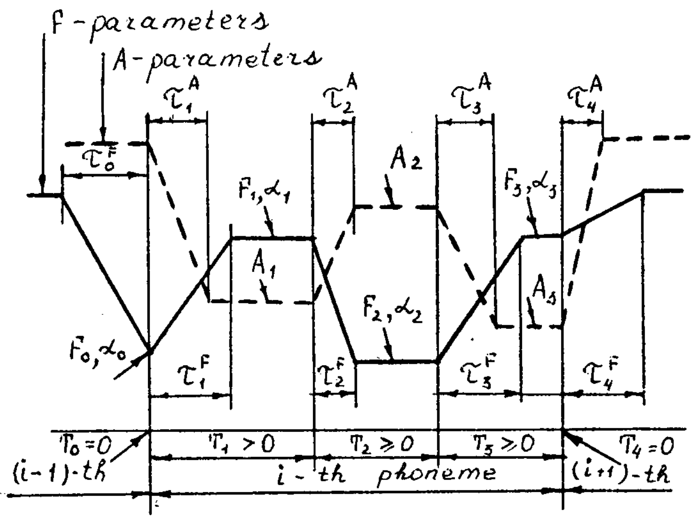


Fig. 1. Graphic presentation of a phoneme portrait

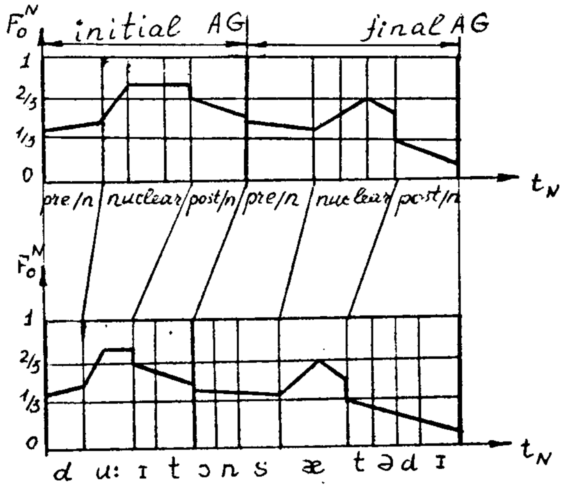


Fig. 4. Example of pitch portrait transformation

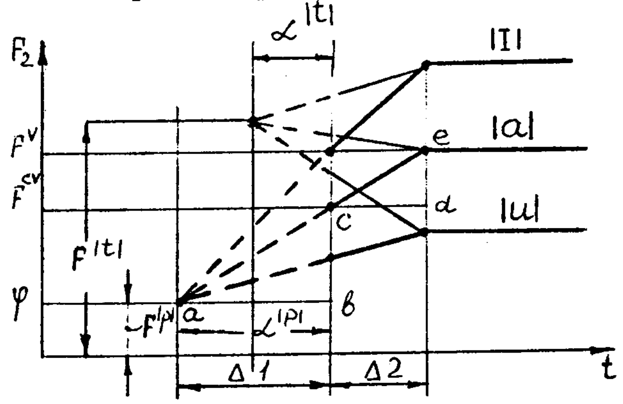


Fig. 2. Geometrical sence of F^c and α^c

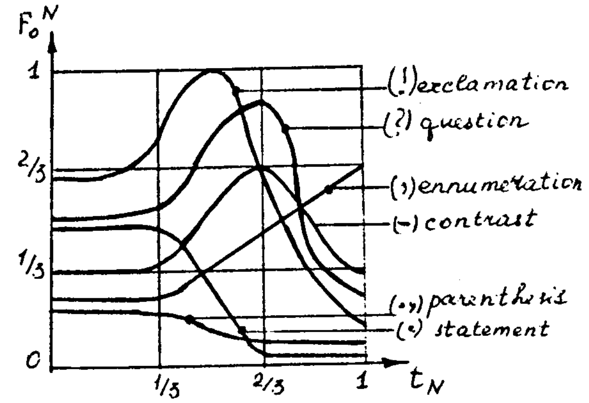


Fig. 3. Pitch component of the final accentual group prosodic portraits six intonational types (Russian)

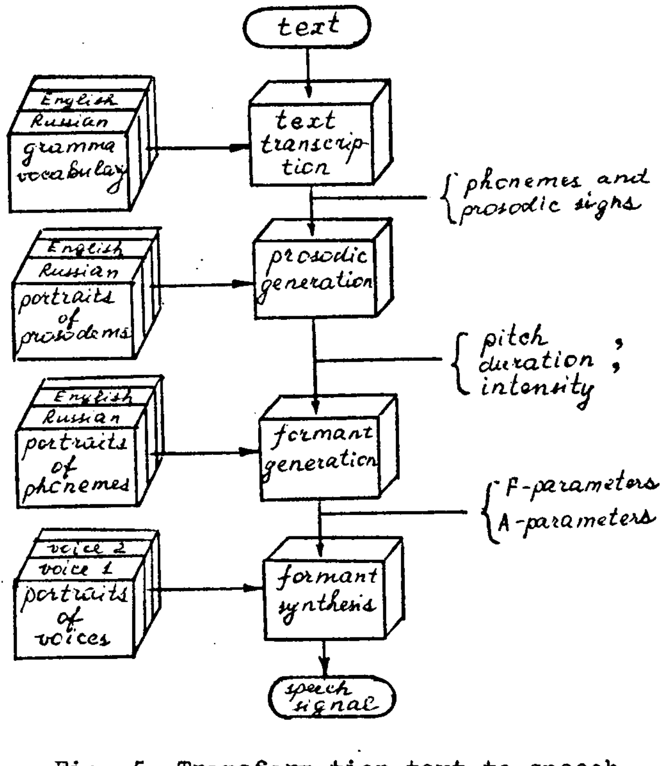


Fig. 5. Transformation text-to-speech algorhythm

sis of transformation algorythm text-to-speech in the Phonemophon system. The block-scheme of the algorythm is presented in Fig. 5.
At the first stage a written text is changed by certain rules, including a morphological vocabulary into a phonemic one which is provided with prosodic notation.
At the second stage prosodic parameters as based on the prosodic portraits and the rules of transforming frequency, duration and intensity are being generated. At the third stage formant parameters formed on the basis of phoneme portraits and the rules of transforming them into F-and A-parameters are being generated. From thus obtained sets of parameters many voices formant synthesis of speech signal is being performed.
The peculiarities of building phoneme and prosodeme portraits for multi-language speech synthesis and the rules of their transformation are described in [4] .

SUMMARY

The above presented strategy of speech synthesis from text formed a basis for compyling a series of Phonemophon devices. It has covered the distance from Phonemophon 1 to Phonemophon 5 since 1972 to 1987. On their basis since 1982 a mass production of speech synthesizers from text has been launched. The latest version of Phonemophon 5 is a single-card device built by digital microprocessors. It ensures a bylingual speech synthesis from text (Russian and English for instance),it is supplemented with controlled voice characteristics (3 male and 2 female) and with the controlled speech tempo. It is also well provided with the interface with a computer and telephone.

ACKNOWLEDGEMENTS

I gratefully record my thanks to Elena Karnevskaya for the beneficial cooperation in creating essential linguistic principles. I should also like to thank Michael Marchenkov and Irina Aksyutina for their participation in the constraction of the programmed model. I must thank Valeriya Afanasyeva and Yuri Zimitsky who have built the device itself.
I am greatly obliged to my Institute's authorities and the colleagues of the laboratory who have readily supported this work.
And last not least I should like to thank Ludmila Leladze for the magnificent preparation of English version of this paper.

REFERENCES

1. D.H.Klat. The KLATTalk text-to-speech conversation system. In. Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. Paris, 1982, pp. 1589-1592.
2. Hertz S. "From Text to Speech with SRS". J. Acoust. Soc. Am. 72(4), 1982, pp. 1155-1170.
3.B.M.Lobanov. On the Acoustic Theory of Coarticulation and Reduction. IEEE Int. Conf. Acoust., Speech, Signal Processing. Paris, 1982, pp. 915-918.
4. E.B.Karnevskaya. The Linguistic aspect of multi-language speech synthesis. In this volume.

A RYTHM-BASED PROSODIC PARSER FOR TEXT-to-SPEECH SYSTEMS IN FRENCH

Christel SORIN, Danièle LARREUR and Régine LLORCA

Centre National d'Etudes des Télécommunications 22301 Lannion FRANCE

Abstract

The prosody is one of the main factors deciding the quality of text-to-speech synthesis systems. We present here a system allowing for a prosodic parsing and an automatic prediction of a French prosody which makes no use of syntactic analysis. The system was derived from studies on the prosody used in commercial announcements. In the first step, a sentence is divided into Prosodic Groups (PG's) which consist of lexical words located between two grammatical words. In the second step, the length and relative location of PG's determine the insertion of pauses and the specific prosodic categories attributed to each PG. Finally, simple right-to-left derivation rules furnish the prosodic category of each word inside the PG. Predefined Fo and duration rules are then applied depending on the prosodic category attributed to each item.

Introduction

The automatic generation of prosody in text-to-speech system consists into two phases :

Phase 1 : definition of prosodic rules allowing to automatically derive Fo and duration contours from prosodic markers (manually) introduced in the text.

Phase 2 : definition of parsing rules allowing to predict the location of the prosodic markers automatically.

Existing text-to-speech systems for French include different sets of prosodic rules (see for example, Emerard, 1977, for the CNET synthesis sytem, O'Shaughnessy, 1984 and Bailly, 1986, for the INRS system, Lienard et al, 1977, for the LIMSI system and Carlson and al, 1982, for the KTH system). These rules were mainly defined by studying Fo contours of read sentences. Another prosodic speaking-style is that used by radio or TV speakers for news or commercial announcements. This "speaking-style" largely uses lexical emphasis and aims to be maximally intelligible and convincing. It could therefore be well adapted to speech synthesis system towards counterbalancing the negative effects of the segmental defaults of synthetised speech.

In the first part of this paper, we present a new set of prosodic rules trying to mimick French "commercial" prosody. In the second part, the prosodic paser will be described that allows to generate, in the CNET's synthesis system, both types of prosody the "reading" prosody and the "commercial" prosody.

I- Rules for "commercial" prosody generation in French

The rules system consists into 3 modules :

- a "duration" module
- a "macroprosody" module
- a "microprosody" module.

1/ Duration rules

Two different sets of duration rules were defined. The first one is intimately related to a diphones-based synthesis system. The duration rules aims to complete the duration effects already captured inside the stored diphones by durational modifications which appear inside a sentence. Established 11 rules include the lengthening of the last word-syllabe before a main prosodic boundary, the shortening of consonant clusters inside a word, the shortening of middle syllables inside long plurisyllabic words, a special treatment for monosyllabic lexical words etc... These rules use the informations provided by the intonation markers which will be described in the following paragraph.

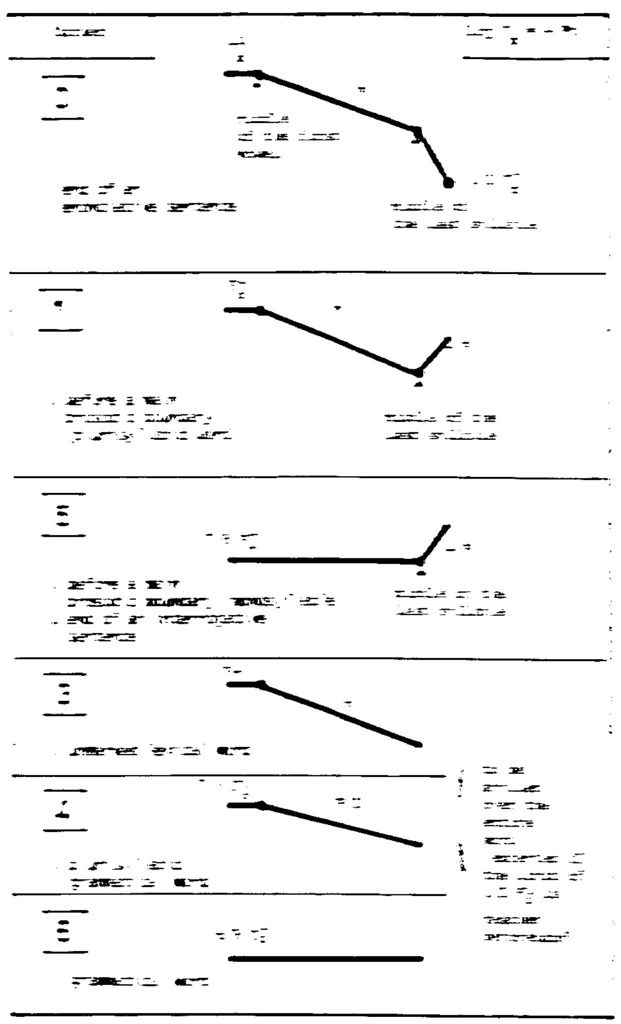
However these rules only modify the intrinsic segmental duration of the stored diphones. Therefore, the criteria used for choosing the diphones (both the environnment from which they were extracted and the segmentation criteria) still strongly influence the segmental durations of resulting synthesised sentences.

A second set of rules was developped so that the duration module would be independent from the type of synthesis system (formant or diphone-based). This predictive model of segmental duration (Bartkova et Sorin, 1985) was tested on three corpora : the mean differences between measured and predicted segmental durations were less than the Just Noticeable Difference (JND) for duration in connected speech (Huggins, 1971).

2 Macroprosodic rules :

Macroprosodic rules were based on the study of the prosody used in commercial announcements. The basic corpus contained 100 sentences recorded by a professional female speaker for a commercial purpose (informations on a new product). The observed Fo contours were well described by 6 typical Fo contours for which the Fo evolution was formalised. Table I presents these 6 schematized Fo contours that are associated to 6 prosodic markers. They can apply either on a word or on a sequence of words located at the left of the prosodic marker inserted into the sentence. These contours are mainly defined by their Fo initial value (Fo^i), the slope P, the final Fo value (for the sentence final Fo contour). Two supplementary markers + and - can be associated with any marker. They allow to increase or decrease the Fo initial value by 1/10 Fo^i steps. For long words or words-sequences involved in an unique Fo contour, some rules modify slightly the above presented contours (for example, the P-slope is divided by a factor 2, the Fo^i value is maintained over the 2 first syllables etc...).

TABLE I



At this level, the pauses are indicated by 2 supplementary markers (" for a long 300 ms pause, ' for a short 100 ms pause) that can be associated with every other Fo marker.

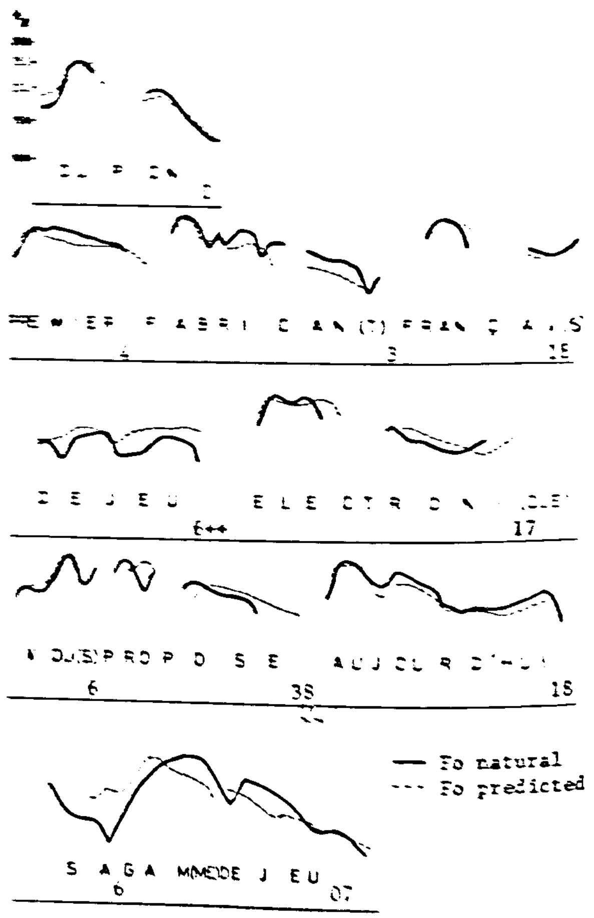
3 Microprosodic rules :

This module contains three set of rules which are automatically applied :

- microprosodic rules for vowels in an unvoiced context (smoothening of the vowel Fo contour when preceeded or followed by an unvoiced segment),
- microprosodic rules for voiced consonants : a dip is introduced in the macroprosodic Fo contour at the place of the consonant,
- "microfluctuation" rules : to avoid the presence of long sequences having the same Fo value (for example long vowels) some fluctuations are introduced on the flat Fo contours (their magnitude is less than 10 Hz).

This set of rules allows for a good prediction of the observed Fo contours. As an example, Figure 1 displays the Fo contours that were obtained after manual assignment of prosodic markers, in comparison with the original Fo contours of the sentence.

FIGURE 1



II. Prosodic parsing of a sentence in French

In many text-to-speech synthesis systems, the prosody is derived from more or less complex syntactic analysis of the sentence. However, for French, Choppy and al (1975) proposed an automatic generation of prosody that avoids the need of a syntactic analysis of the text. Some recent studies (Wenk and Wiolland (1982), Dell (1984) and Martin (1986)) suggest that rythmical constraints could strongly influence the prosodic structure of the sentence. In the corpus we studied, we observed a strong tendency for segments between pauses or prosodic juncture to have the same number of syllables (generally inferior to 7 syllables).

In these context and for practical reasons (i.e. to avoid the use of an heavy syntactic parser), we developed a prosodic parser that maximally uses (beside the ponctuation) the presence of short grammatical words inside the sentence. These words have, in fact, 2 main characteristics :

- they are indicators of some syntactic structure
- they present frequently a relatively stable low Fo contour, that acts as a trempling before the higher initial pitch of the following lexical word.

A lexicon of 120 grammatical words was built. The words belonging to this lexicon are marked ϕ. Among them, a special group contains the grammatical words that, most of the time, introduce a subordinate phrase (they are marked ϕ**) and another group that allows to detect the presence of a verb (they are marked ϕ*).
The prosodic parsing of the sentence is done in the following way :

1/ detection of the word marked ϕ, ϕ* or ϕ**
2/ introduction of brackets (][) before every word ϕ, ϕ* or ϕ** which is preceeded by a non ϕ-word and before ponctuation signs (like ", ( ) :" etc...).

The sentence is then parsed into segments between brackets. These segments will be designated as "Prosodic Groups" (PG) in the following.

A second module attributes to each PG a specific category which will define the location of the pauses and the main prosodic boundaries. Here, the basic idea was to introduce pauses after long PG in order to simulate breathing pauses. We hypothesised that it was preferable to introduce (in the synthesised sentence rather larger number of pauses than a realistic number of pauses (as in natural spontaneous speech : such pauses could reduce the mental load of the listener due to the heavier processing of alterated speech (Nusbaum and Pisoni, 1982). However, the location of those pauses should be, of course, prosodically plausible.

4 main categories are attributed to each PG as a function of :

- the number of lexical words inside each PG
- the position of the PG inside the sentence
- in some cases, the number of syllables in the PG and the previously attributed categories of the surrounding PG's.

TABLE II

Examples of prosodic parsing rules

- the sentence-final PG	. receives the category I
- PG followed by a comma	. receives the category IV . is followed by a long pause "P"
- PG containing 3 (or more) lexical words	. receives the category IV . is followed by a long pause "P" . attributes the category IV to the preceeding PG . is preceeded by a short pause "p" (facultative)
- PG followed by a PG containing a ϕ* or ϕ**-word	. receives the category IV . is followed by a short pause "p"
Stylistic rule (specifically observed in commercial announcements) - PG preceeding the sentence final PG	- if the total syllables number of the 2 PG's exceeds 7 syllables : . receives the category IV . is followed by a short pause "p" - if not : . receives the category II . attributes the category IV to the preceeding PG . is preceeded by a short pause "p"
- PG containing an unique lexical word	. receives the category V (if no category was previously attributed)
- PG containing 2 lexical words	. a set of contextual rules attribute or the category V or the category IV and a short pause
- sequences of PG having received the category V	. if the total number of syllables exceeds 7, an eurythmic index is calculated : a short pause is introduced between the PG's which delimit the eurythmic structure. Category IV is attributed to the PG preceeding this pause.
- etc...(essentially Pauses-harmonisation Rules).	

Right-to-left derivation rules inside a PG

V ← VI ← V ← II ← I
V ← VI ← V ← IV
V ← VI ← V

The final step of the processing consists of deriving the prosodic markers from the categories attributed to each group. This task is achieved in 2 different ways for the "reading" prosody in one hand and for the commercial prosody in the other hand. In the first case, a simple correspondance-table associates each category to one of the previously defined prosodic markers (Emerard, 1977). In the second case, some right-to-left derivation rules are applied inside each PG : a category is attributed to almost every word in the sentence (some intermediate rules group some monosyllabic word sequences into an unique "prosodic word"). At this level, (which now use 6 categories) a correspondance table associates to each word-category one of the markers which were presented in the first part of this paper (Table III).

TABLE III

Category	Prosodic Marker
I	0-
II	4*
IV	1- or 5- (monosyll.)
V	4*
VI	3-
ɸ, ɸ* or ɸ** word . unique . two . sequence	 6 6 and 6- 4-
short pause "p"	8
long pause "P"	7

Table IV gives some examples of the results both for the PG categorization and for the allocation on prosodic markers for the "commercial" prosody.

Conclusion

The entire prosodic module was tested on a large body of TELEX messages. Special items like surnames, accronyms, numbers, abbreviations, were treated beforehand by a text-preprocessing module. The results were judged to be satisfactory enough to implement this module into a text-to-speech system for reading electronic mail.

Some defaults of this module indicate the limits of a "syntax-independent" prosodic parser : in some cases, rythmical constraints must be subordinated to syntactic structure, which cannot be detected without a profound syntactic analysis. This is the case, in particular, for verbs or verbal forms, as illustrated in Table IV ("mis en place" must be considered as an PG because it is derived from the verbal form "mettre en place"). Corresponding prosodic improvements could then be reached only in using, at least, a large lexicon of verbal forms or a fine syntactic (and may be) semantic analysis which remains to be done.

Acknowledgements

The authors wish to thank S. Maeda for his help in writing the english version of this paper.

Références

Bailly, G. (1986) : "Multiparametric generation of French prosody from unrestricted text", IEEE-ICASSP, 2419-2422.

Bartkova, K. and Sorin, C. (1985) : "Predictive model of segmental durations in French", J. Acous. Soc. Am., 77, suppl. 1, S54 (to appear in speech Comm.).

Carlson, R., Granström, B. and Hunnicutt, S. (1982) : "A multi-language text-to-speech module", Proc. IEEE-ICASSP 82, 1604-1607.

Choppy, C., Lienard, J.S., and Teil, D. (1975) : "Un algorithme de prosodie automatique sans analyse syntaxique", Proc. 6th JEP/GALF, 387-395.

Dell, F. (1984) : "L'accentuation dans les phrases en français", in "Formes sonores du langage", ed. by Dell, Hirst and Vergnaud, Hermann, Paris, 65-122.

Emerard, F. (1977) : "Synthèse par diphones et traitement de la prosodie", Thèse 3° cycle, Univ. of Grenoble.

Huggins, A.W.F. (1971) : "Just noticeable differences for segment duration in natural speech", J. Acous. Soc. Am., 51, 1270-1278.

Lienard, J.S., Teil, D., Choppy, C. and Renard, G. (1977) : "Diphone synthesis of French : Vocal response unit and automatic prosody from text", Proc. IEEE-ICASSP 77, 560-563.

Martin, P. (1986) : "Structure prosodique et structure rythmique pour la synthèse", Proc. 15 JEP/ GALF, 89-91.

Nusbaum, H.C. and Pisoni, D.B. (1982) : "Perceptual and cognitive constraints in the use of voice response systems", Research on Speech Perception, Progress Report 8, Indiana Univ., 203-216.

O'Shaughnessy, D. (1984) : "Design of a real-time French text-to-speech system", Speech Comm., 3, 233-243.

Wenk, B. and Wiolland, F. : "Is French really syllable timed", J. of Phonetics, 10, 193-216.

TABLE IV : **Examples of Prosodic Parsing and Allocation of Prosodic Markers**
(sentences presenting no ponctuation sign)

P = long pause
p = short pause

ɸ words		ɸ* ɸ**	ɸ*	ɸ	ɸ	ɸ
PG categories	IV	IV	V	V	IV	I
Intra-PG categories			VI			V II
Prosodic markers	[Demain] 13-	[vous qui voyagez] 6 6- 13-	[vous pourrez gagner] 5 3 4*	[du temps] 6 3-	[en utilisant] 5 15-	[les voyages vite Air] 6 4* 3- 07-
Prosodic Parsing and PG categories	[Les trois malfaiteurs] [et le complice] [qui les attendait] [au volant] [d'une voiture] V IV_p V V IV_p [ont réussi] [à échapper] [aux policiers] [en dépit] [de l'important dispositif V IV_p V IV_p mis][en place] IVb V policier][en place] IV_p V [dans toute la région] [en emportant] [un butin] [dont le montant] [n'a pas été révélé]. IV_p V IV_p IV_p I_p					

TEXT-TO-SPEECH RUSSIAN SYNTHESIS BY RULE

ALEŠ BUČEK

Tesla Electronics Research Inst.
Prague, Czechoslovakia

JEVGENIJ TIMOFEJEV

Faculty of Pedagogics
Hradec Králové, Czechoslovakia

ABSTRACT

The paper concerns the present-day state of research of automatic conversion of Russian written text into a corresponding acoustic signal.

INTRODUCTION

Our information is limited to the results of research work carried out in the framework of integrated efforts of Tesla Electronics Research Institute in Prague and Faculty of Pedagogics in Hradec Králové. As a result of the result of the research work the first version of the program, which assigns a sequence of short sounds of appropriate amplitude and spectral composition to any Russian text written in a usual form, has been developed. The sounds transmitted by microcomputer rapidly one after another are percepted by users as spoken Russian.

SPEECH SYNTHESIS

Our solution is based on approximation of speech signal on the basis of the basis of two-formant sounds, which are tabulated for one-period length in the computer memory and transmitted into loud-speaker respecting the sound combination of input text. We have used an arsenal of 12 vocal-like sounds and of 1 noise-like sound. By changing the rate of transmision of various digital patterns, by various number of patterns in one period, by various number of periods and various loudness we have obtained much more extensive set of various sounds. By means of these 12+1 sounds and by way of their transmission individual elements of spoken speech are described in the computer memory. We have defined these elements as follows:

↗h - sound initiation
h - sound body
h↘ - sound ending
cv - consonant-vowel combination /each with other/
vc - vowel-consonant combination /each with other/.

Russian word CKOPO /sko:ra - phonetic transcription/ is decomposed in comformity with this definition into the following speech elements:

↗s,s,s↘,↗k,k,ko,o:,or,r,ra,a,a↘

The above decomposition is not entered by the user - the computer carries out the operation without any intervention from the user's part.

The tables of parametric description of individual sounds, which approximate the sounds of speech, have been composed on the basis of prof. M. Romportl's and prof. L. Bondarko's works. We have also used spectrogams of natural speech. Perception tests were the decisive argument of spectrograms interpretation.

In the first version of our synthesis the high degree identification of Russian word accent /when basic prosodic parameters are absent/ is provided by means of:

- greater quantity of stressed vowels vs. unstressed vowels, final stressed a-vowels are double elongated
- stressed vowels are 6 dB louder than unstressed vowels
- stressed vowels are 1/2-tone higher than unstressed vowels
- quality alternation of unstressed o/a-vowels /a-norm pronunciation/, i-norm pronunciation can be also introduced, but the perception of word accent is not improved.

TEXT-TO-SPEECH ALGORITHM

The described method of synthesis of segmental features of speech and approximation of stressed/unstressed vowels phonetic contrast have enabled to produce a synthetic signal of spoken Russian, which has no prosodic feature and sounds somewhat monotonously, but is characterized by high degree adaptability of the users of Russian to this signal with good understanding.

Besides the input text need not be entered in a form of phonetic transcription. The automatic conversion of a usually written Russian text into input phonetic transcription is also provided.

The first version of algorithm of written text-to-phonetic transcription includes four basic stages:

1. Receipt of input text

In the input buffer the system selects only alphabetic letters of written Russian /capital letters of Russian alphabet/, character ":" for word accent, character "," for pause, character ␣ /spacing/ and CR , which ends the receipt of a text.

2. Text transmission into working memory

The text is transmitted character by character from left to right till CR . During transmission some characters or some combination of characters are processed:

- realization of some consonant combination is modified / ПРА:ЗДНИК - prazn'ik, ЧЕ:СТНЫЙ - česnyj, КАТА:ЕТСЯ - katajetca, ЛЁГКИЙ - l'oxk'ij etc./
- realization of adjective inflextion in genetive case is altered / ДОРОГО:ГО - dorogovo/
- a-norm pronunciation is introduced / ХОРОШО: - xarašo/
- realization of pronoun ЧТО and conjunction ЧТОБЫ is altered /što, štoby/
- i-norm pronunciation is introduced in a limited size /in the first unstressed syllable before stressed syllable/
- consonant combination СЧ is substituted by realization of šč /in a limited size/.

3. Text processing in working memory from left to right:

- the letter Ъ before vowels is substituted by its spoken equivalent
- doubled consonant is substituted by single one
- the orthographical Ь is conversed into its phonetic realization
- pronunciation of a preposition with unstressed vowels is realized / ОБО, ПЕРЕДО etc./
- conversion of multiciphered letters /Е, Е, Ю, Я / is ended
- realization of final stressed a-vowels is modified
- the so-called coarticulation in vowel combination / НАУ:КА, СООБЩЕ:НИЕ etc./ is respected.

4. Text processing from right to left:

In this stage the text is processed according to deaf-sonorous assimilation laws of Russian. The text processing is finished as soon as the beginning of the text is reached.

In the present stage of development, our algorithm of automatic transcription contains more than 30 rules and occupies 1,5 KB of ROM-memory. It is universal and every Russian text can be processed. Algorithm development has been based on two methodical principles: approximation and ignoration. For example, the algorithm approximates the pronunciation of all unstressed a/o-vowels as a single realization of a weak /a/ in opposition to a strong stressed /a/. The pronunciation of some strange-origin Russian words with weak unstressed o-vowels is ignored. Nevertheless the user has an opportunity to produce realization with unstressed o-vowels: in this case accent need not be input /the qualitative alternations of unstressed vowels are conditioned by accent input/. The basic criterion for algorithmic rules extension is communicative effect of an acoustic signal and its aesthetic realization. For example, from communicative point of view it is not necessary to modify the consonant combination ЧТ, ЧН into št, šn. The user of Russian will understand text with čt,čn-realization in the same way as with št,šn-realization. But from aesthetic point of view the above modification of text should be desirable. That is why the čt,čn - št,šn conversion has a limited effect and is valid for words ЧТО and ЧТОБЫ only. The rest of words are ignored / Решение конечно. - Конечно, он прав./.

The first version of our text-to-speech algorithm contains the greatest part of Russian written text/phonetic realization differences and is being constantly improved. The practical ideal version of the algorithm is connected with further progress in miniaturization of hardware as well.

TEXT-TO-SPEECH CONVERSION FOR GERMAN USING A CASCADE/PARALLEL FORMANT SYNTHESIZER

GERHARD RIGOLL**

Fraunhofer-Institute (IAO)
Dept. of Advanced Information and Communication Technologies
Holzgartenstr. 17, 7000 Stuttgart 1, West Germany

ABSTRACT

The paper describes some aspects of the use of the cascade/parallel formant synthesizer for German text-to-speech synthesis. Since the algorithms used for speech synthesis are relatively similar for the most languages, the paper emphasizes some novel approaches and special phonetic problems, such as the use of a mathematical model for the cascade/parallel formant synthesizer for the determination of the synthesizer control parameters or the synthesis of phonemes with special articulation, rather than to describe more generally the development of the entire system.

INTRODUCTION

In 1983, the work on the development of German text-to-speech converters, based on the cascade/parallel formant synthesizer developed by D.H. Klatt, has started. In general, the development of a text-to-speech system can be described under different aspects, e.g. emphasizing more the common technical problems, or the letter-to-sound conversion, or the fact that the system might have been modified for a different language, which usually requires a complicated modification procedure that was also performed for the system described here. In this paper, the phonetic aspects of the use of the cascade/parallel formant synthesizer for the German language are especially considered. The accurate determination of the main control parameters for the cascade/parallel formant synthesizer is probably the most important step in order to achieve high voice quality of the final system, although many different steps, e.g. letter-to-sound conversion or prosodics, which are not considered in this paper, are also responsible for the overall quality of the system.

THE CASCADE/PARALLEL FORMANT SYNTHESIZER

The formant synthesizer which was used for German synthesis is a modified version of the synthesizer described in /2/ which was improved by D.H. Klatt during the last years. The current version is now very flexible and capable of synthesizing different voices, including women and children voices. The most important control parameters are still the first three formants and bandwidths and the fundamental frequency. Additionally, it is possible to control special parameters for the voicing source and for prosodics, which is mainly used for the generation of different speaker characteristics. The German vowels and sonorants are synthesized using the cascade branch, while voiceless fricatives and plosives are generated by the parallel branch. Only for voiced obstruents, both branches of the synthesizer are excited.

MATHEMATICAL MODELLING OF THE CASCADE/ PARALLEL FORMANT SYNTHESIZER

There are basically three possibilities to obtain the values for the control parameters of the synthesizer. The fastest and simplest method is a perceptually based method, where the parameters for every phoneme are tuned as long until the synthesis of the phoneme sounds very similar to the original utterance of the phoneme. Although it is possible to find relatively fast a parameter constellation which is leading to an acceptable result for every single phoneme, the overall quality of such a system is mostly poor and many phonemes which used to sound natural when they where tested in isolation, sound unnatural in connected speech. The second method is based on the calculation of the parameters with the use of speech analysis tools, e.g. formant calculation based on an LPC analysis and the generation of the phonemes with the parameters obtained from this analysis for each phoneme. But also this method leads to problems because usually the generation of a sound, using the formant values which were obtained from an analysis of an utterance of this sound, does not lead to a synthetic sound with exactly the same acoustic and spectral properties of the natural utterance. The third method is a spectral tuning of the synthesizer parameters to the spectral properties of one single speaker. This is a very time consuming iterative process, where at the first step the initial values of the synthesizer parameters are derived from an analysis of a natural utterance, as in method 2. In the following steps, the parameters are tuned as long until the spectral analysis of the synthetic utterance is similar enough to the spectral analysis of the natural utterance. In this way, the system is forced to have almost the same spectral features as the single test speaker, which can lead to a high amount of naturalness of the synthetic voice. If such a procedure is used, it is obvious that the success is also dependent on the tools and algorithms which are used for the analysis and comparison of natural and synthetic speech. Beside the more traditional analysis methods like spectrograms and spectra, a novel approach was tested during the development of the German text-to-speech system. This approach is based on a mathematical model of the cascade/parallel formant synthesizer. Based on the fact, that both branches of the synthesizer are composed of digital resonators with the transfer function

$$G(z) = \frac{1 + c_i + d_i}{1 + c_i z^{-1} + d_i z^{-2}} \qquad (1)$$

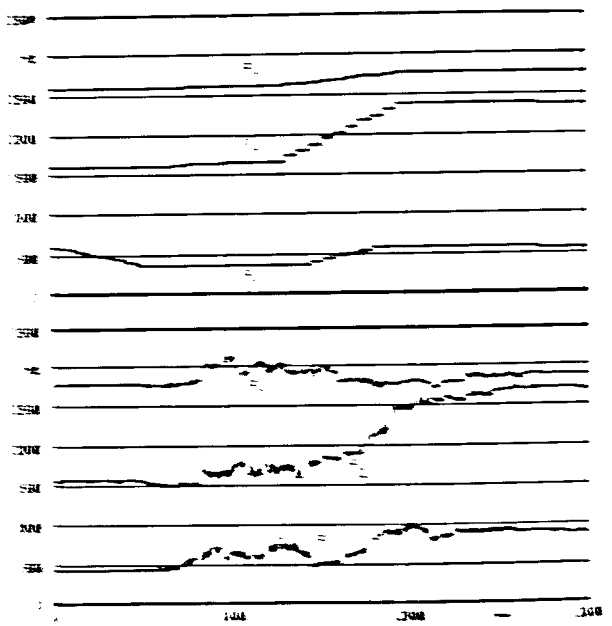
where the resonator coefficients c_i and d_i contain the according formant F_i and bandwidth B_i as nonlinear functions:.

$$c_i = -2e^{-\pi B_i T} \cdot \cos(2\pi F_i T) \qquad (2)$$

$$d_i = e^{-2\pi B_i T} \qquad (3)$$

** The author is currently with the Continuous Speech Recognition Group, Dept. of Computer Sciences, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

[illegible]



[illegible]

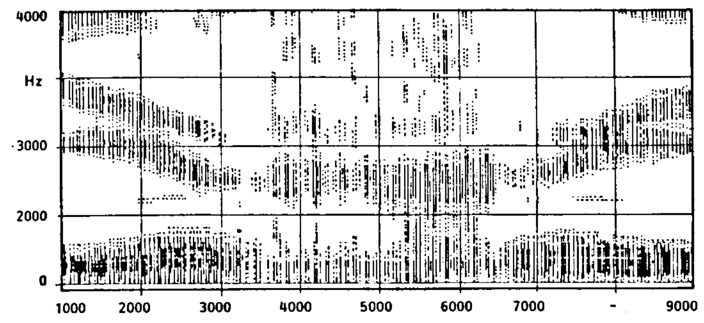


Fig. 2: Spectrogram of the phoneme sequence /iRi/

This periodical change can be demonstrated even better by looking at the formant tracks of this sequence in Fig. 3, obtained from the earlier described nonlinear parameter estimation procedure.

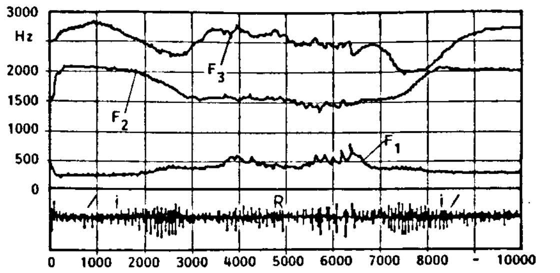


Fig. 3: Formant tracks of the phoneme sequence /iRi/ calculated with the use of a nonlinear parameter estimation algorithm

The synthesis of the uvular /R/ can be performed either by a modulation of the gain AV of the voicing source or by a modulation of the formants. It was decided to use the latter method in the current version, where only the second formant was modulated by a certain percentage of his stationary value, which is shown in Fig. 4. Simultaneously to this modulation, frication noise is added via the parallel branch by setting the gain A_2 of the second parallel resonator to a value different from zero. In this way, the uvular /R/ is handled similar to a voiced fricative.

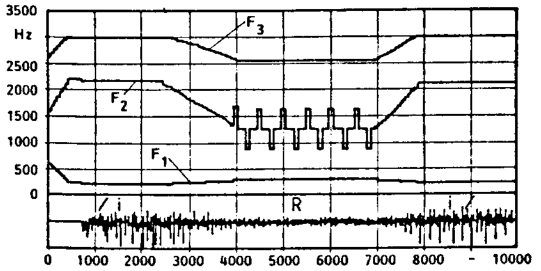


Fig. 4: Formant tracks given to the cascade synthesizer branch for the production of the phoneme sequence /iRi/

COARTICULATION

A very important module of a text-to-speech system are the rules for coarticulation. In the current version, coarticulation is performed in several ways

o the coarticulation of the spectrum of the stationary part of vowels and sonorants with the preceeding and the following phonemes is calculated by a general coarticulation formula which is applied to the formant values and reflects the percentage of the influence of the formant values of the neighbour phonemes to the formant value of the current phoneme

o the variations of boundary values in the transitions of consonants to different vowels are automatically taken into account by the application of the locus theory

o the variations of the spectral properties of consonants in the environment of different phoneme classes is taken into account by modification of the gains of the parallel resonators according to the current environment

o special extended rules are applied to some sonorants which show strong coarticulation. In the German language these are especially the phonemes /l/ and the uvular /R/. Again the /R/ is the most difficult phoneme to handle scince it requires complex rules for the formant values as well as for the control of the frication by the parameter A_2 if it appears in different environments. The coarticulation of these sounds can sometimes be effected only by the left or sometimes only by the right neighbour phoneme and often also by both neighbour phonemes, depending on these phonemes

Since the recording and tuning of phonemes is never carried out with phonemes spoken in isolation, special attention has to be paid to incorporate the coarticulation rules into the process of the spectral tuning of the various sounds because these rules will modify the originally entered parameter values according to the current phonetic environment, in which a specific phoneme is recorded, analyzed and tuned. This is a serious problem scince at the begin of the tuning task, the coarticulation rules are usually not yet known, but they are theoretically required in order to obtain optimal tuning results.

CONCLUSION

Some important steps of the synthesis of German with the cascade/parallel formant synthesizer have been described as well as some new approaches for the analysis of speech to obtain the values for the synthesizer control parameters. The description of the development of the entire text-to-speech system is beyond the scope of this paper. The experiences which were gained during this development have shown that the careful and time consuming tuning of every single phoneme and the consideration of many special cases and exceptions is the key to obtain a synthesizer with a high voice quality.

REFERENCES

/1/ G. Rigoll: The DECtalk System for German: A Study of the Modification of a Text-to-Speech Converter for a Foreign Language. Proc. IEEE-ICASSP, Dallas, 1987.

/2/ D.H. Klatt: Software for a cascade/parallel formant synthesizer. J.A.S.A., Vol. 67, No. 3, 1980.

/3/ G. Rigoll: A New Algorithm for Estimation of Formant Trajectories Directly from the Speech Signal Based on an Extended Kalman-Filter. Proc. IEEE-ICASSP, Tokyo, 1986.

/4/ A. Gelb: Applied Optimal Estimation. M.I.T. Press, Cambridge 1974.

ENLIVENING THE INTONATION IN TEXT-TO-SPEECH SYNTHESIS: AN 'ACCENT-UNIT' MODEL

JILL HOUSE MICHAEL JOHNSON

Dept. of Phonetics and Linguistics
University College London
Gower Street, London WC1E 6BT

ABSTRACT

A new model of intonation for text-to-speech synthesis exploits natural variability within phonological constraints. Patterns are determined with reference to those preferred by an individual speaker.

INTRODUCTION

The output of a text-to-speech synthesis system needs to be intelligible, reasonably natural, and acceptable to the listener. A successful model of intonation will contribute to intelligibility, by clarifying the information structure of the text, and to naturalness, by using F0 contours characteristic of the target speech, aligned to the segmental structure of the text in a phonetically principled manner. To be acceptable to the listener, the output must combine intelligibility with whatever degree of naturalness is necessary to make the act of listening a comfortable, undemanding experience.

For the synthesis of isolated sentences, patterns may be readily specified which are plausible and 'easy on the ear'; but the use of these same patterns over longer texts, of a paragraph or more, leads to repetitiveness which the listener may find tedious: so, acceptability declines. We propose that enhanced acceptability during sustained listening may be achieved by exploiting a further aspect of naturalness: the variability to be found in the intonation patterns of natural speech.

THEORETICAL BACKGROUND

In natural speech, the choice of intonation contour for a text involves a number of separate phonological choices, some of which carry a higher functional load than others. These choices significantly constrain the degree of allowable variability, but within these constraints there is no one single 'correct' intonation pattern applicable to a given text spoken in a given context.

In developing an intonation model for synthesis-by-rule, an early priority must be to identify the sub-systems within which choices are made. For example:

(1) the division of the text into intonational phrases, or 'tone-groups';
(2) the allocation of accents (rhythmically stressed syllables which are also pitch-prominent, in the sense that they interrupt an established pitch contour);
(3) the relative prominence of accented syllables;
(4) the selection of the pitch contour whose starting-point coincides with the final accented syllable (the 'nucleus') of the tone-group -- the 'nuclear tone';
(5) the selection of pitch contour over any remaining (pre-nuclear) syllables.

These sub-systems imply a contour-based analysis which owes much to the 'British school' of intonation, notably /1/ and /2/. We believe that this approach is well motivated at the phonological level.

A theoretically sound synthesis model must allow for those formal differences for which a functional account can be given; ideally it should also model observed formal variations where no functional motivation may be apparent.

While lexical, syntactic and semantic factors play their part, the unifying principle determining intonation assignment is surely a pragmatic one -- the tailoring of an utterance to its context. In a synthesis system using unrestricted text input, any semantic or pragmatic knowledge is bound to be very limited. The rules must exploit any lexical or grammatical knowledge available, but occasional inappropriate choices will inevitably risk lowering the acceptability of the output (cf. /3/). The adverse effect of such errors may be minimised by an output which is otherwise natural-sounding and easy to listen to.

This paper does not directly address the problem of improving the syntactic, semantic or pragmatic knowledge-base. The model described assumes that the input text has been converted to a transcription on which tone-group boundaries and accented syllables are explicitly marked.

THE MODEL

Foundations: auditory analysis of a corpus

The model's phonological units and probabilistic rules were based on close auditory analysis and prosodic transcription of a short corpus of recorded texts. Four texts of 150-250 words each were derived from information bulletins -- reports on road conditions and weather forecasts -- issued over the telephone, using a declarative English style. Recordings of the original speakers (3 male, 1 female) were transcribed orthographically, using suitable punctuation, and presented as written texts to five experienced readers (3f, 1m), who in turn recorded the texts on to PCM tape in an anechoic chamber. A laryngograph signal (Lx), from which subsequent excitation frequency (Fx) analyses were made, was recorded simultaneously. All speakers used a (near) RP variety of English.

The recorded speech was transcribed prosodically, on an auditory basis, using a syllable-by-syllable interlinear notation. Comparison with the derived Fx traces indicated a reasonable match in terms of contour shape and relative pitch levels. No attempt was made to transcribe durational variation.

There was no one preferred reading for any of the texts, with respect to *any* of the sub-systems outlined above. A contour-based interpretation in terms of tone-groups and nuclear tones seemed well motivated, with falling, falling-rising, rising and level patterns all perceptually salient at the ends of intonational phrases. A consistent finding was a high degree of variability in contour-shape in pre-nuclear position. The contours were not readily interpretable in terms of fixed-pattern 'heads' (cf /1/); nor were sequences of accented syllables linked by any kind of automatic contour interpolation (cf /4/). This variability reflected a succession of choices between possible formal patterns. Their functional motivation was unclear, unless it was simply a strategy to avoid monotony. There was some evidence that individual speakers had preferred options among these patterns.

The inventory: units, contours and features

Units. The basic phonological unit chosen for the model is the *accent-unit* (AU) (cf /5/). This consists of an initial accented syllable together with any unaccented syllables following it. The unit is bounded on the right by the next accented syllable or by a tone-group boundary. Minimally, it will contain just the accented syllable; there is no theoretical upper limit, but units may contain more than one rhythmic foot, since some stressed syllables are not pitch-prominent, and are therefore deemed unaccented.

Within a paragraph of text, the largest unit recognised by the model is the *breath-group* (BG). This is normally equivalent to a grammatical sentence, since it corresponds in practice to a stretch of text bounded by /./, /!/ or /?/. A breath-group may be subdivided into *punctuation-groups* (PG) (bounded by /,/, /;/ or /:/), which in turn may contain more than one *tone-group* (TG). Tone-groups are composed of one or more accent-units, together with an optional *prehead* (PH), corresponding to any unaccented syllables preceding the first accent in the group. The final accent of a tone-group is the *nucleus*; preceding (optional) accent-units make up the *head*. All BG and PG boundaries are also TG boundaries. The hierarchical structure linking groups and units is demonstrated below:

(1) [[[["No.$_{AU}$]$_{TG}$]$_{PG}$]$_{BG}$] (= minimal breath-group)

(2) [[[[There are$_{PH}$]["more lane 'closures$_{AU}$]$_{TG}$]

[[between 'junctions$_{PH}$]["thirty$_{AU}$]["two and

'thirty$_{AU}$]["three,$_{AU}$]$_{TG}$]$_{PG}$]][[["north of$_{AU}$]

["Preston.$_{AU}$]$_{TG}$]$_{PG}$]$_{BG}$]

Key: " accented syllable; ' stressed syllable.

Contours and features. Accent-units are characterised by contour. Nuclear unit contours in the corpus represented four basic *nuclear tones*: *falls*, *fall-rises*, *rises* and *levels*. (This formal classification does not distinguish between 'fall-rise' and 'fall + rise'.) Head unit contours fell into three natural classes: *levels*, *falls* and *rises*. An earlier version of this model /6/ treated nuclear and head units separately; the revised version considers both types to belong to the same underlying formal classes. Nuclear units typically involve more salient F0 movement than head units, and are more predictable in their alignment to syllables. Head unit contours showed much variation in this respect, depending not only on the number of syllables in the unit, but on features affecting, for example, the timing of the start and end points of the characterising contour. Stylisations of the basic contour shapes perceived, subsequently adapted for implementation in the model, are shown in Fig. 1.

The use of features owes its inspiration, but not its detail, to Ladd /7/. The unmarked forms in Fig. 1 were those characteristically found in nuclear position, where the marked forms were less common; both marked and unmarked varieties occurred freely in head units, though fall-rises as a class were rare in this position. In practice, there is normally a brief sustention of F0 at the start of the contour, which is accounted for in our implementation; contours are only considered to display the feature [+delayed start] when such a sustention continues into the second syllable. Perceptually level contours may, in fact, follow a shallow declination line; this too is accounted for in the implementation.

Distribution of AU contours in the recorded corpus

Nuclear. Between them, falls and fall-rises accounted for around 75% of all the nuclear tones. All BG boundaries ended in /./, and all were associated with a falling tone (with one exception: the sentence 'Thank you for calling.'). Any of the tones could be found at other PG and TG boundaries. Statistically, fall-rises were most probable here, particularly in the case of BG-initial tone-groups.

Head. Over 90% of the units had either level or falling patterns; rising units were rare. There were few, if any, positional constraints on these contours. However, when they were analysed in relation to the nuclear tone which followed them in the tone-group, certain tendencies came to light. Though levels constituted around 57% of head units overall, this proportion dropped to around 40% when immediately preceding a fall-rise nuclear tone, where they were overtaken by falls. Most of the few rising head-units occurred in tone-groups with a falling nuclear tone. There were no obvious constraints on the juxtaposition of different accent-units, but estimates of the probabilities of certain collocations could be derived from the corpus distributions. There was no clearly identifiable pattern governing the application of the features to these head contours.

The probability values quoted above are derived by adding together the scores for several RP-style speakers, a procedure which allows us to make some generalisations about intonation units used in this variety of English for this discourse style, but which obscures the preferences of the individual speakers. Probabilities based on averaged values, or, preferably, on those appropriate to a particular speaker, may be adopted to implement the model in a text-to-speech system (see below).

Relative prominence of accents

The derived Fx traces relating to the corpus recordings allowed us to make an accurate assessment of the actual and relative peak frequencies of accented syllables. Within a tone-group, there was a marked tendency for the Fx of successive head accents to show some sort of decline. There were no fixed target values associated with accents in any position, but it was possible to define a frequency range within which accents were likely to occur. The position of the tone-group within the breath-group, and of the breath-group within the paragraph, were relevant in determining the height of TG-initial accents. The starting-point of nuclear accents was more varied. a step up from a preceding accent typically reflected a linguistic need to highlight the item in question.

Implementation

This section looks further at the principles guiding the rules in our implementation, rather than at the detailed algorithms, which are still subject to revision. An earlier version of the implementation is described in more detail in /6/; a report on the revised implementation is in preparation.

The model is implemented on the JSRU text-to-speech synthesiser, a system modelled on male RP speech, replacing or adapting the prosodic model of Edward /8/. Rules relate to F0 values only; all other aspects of the system are unchanged.

Reference frequencies. The overall range is based on that of a particular (male) speaker. Values are derived from frequency distribution (Dx) histograms made from the Lx signal recorded with the reading of the texts. The extremes of the range ('HiFx' and 'LoFx') are taken from the 1st order distribution, as is the 'mode' (preferred frequency) value. An additional reference value used in computing the synthetic F0 is 'LoFx2', the lower limit of the range as measured on a 2nd order distribution.

Selection of AU contours. Nuclear contours (nuclear tones) are assigned on the basis of punctuation. For instance, boundaries associated with full stops invariably generate falls; commas and unpunctuated boundaries are associated by convention with the fall-rise, but an algorithm converts this to a rising or level tone in certain circumstances. Head-unit contours are assigned according to tables of probabilities derived from analysis of the speech to be modelled. The tables take account of the transitional probabilities associated with the collocation of different contour types (see /6/ for specific examples). Feature application makes use of tables of stationary probabilitites similarly derived from the corpus.

Contour to F0 conversion. Accent-unit contours are broken down into their constituent levels: H (high) and L (low) for falls and rises, with additional constituents H' and L' to deal with the more complex and marked forms. Level contours consist of H and H' only. The features [delay] and [raised] apply to H or L as appropriate.

F0 values for H and L are calculated on separate criteria. The first H in a breath-group is plotted in relation to the mean value used for such accents by the reference speaker. The H value in subsequent accents will be at a point which is a fixed proportion of the distance between the mode and the previous H. An adjustment upwards is made in a new punctuation-group for the first H, which is related to the previous PG-initial accent. There is a degree of allowable deviation from the computed mean values. Declination between accented syllables is a *derived* effect.

In nuclear units, the value of L coincides with LoFx2, unless it is BG-final, in which case it coincides with LoFx. In head units, L is calculated as a proportion of the distance between its associated H and LoFx2.

Prehead syllables are by default clustered around the modal value.

DISCUSSION

Many of this model's algorithms are still being modified, but even at this early stage, we believe that the output has much to recommend it. It is closely based on observations of natural speech; it allows the distribution of patterns to be modelled on a particular speaker; it exploits natural intonational variability. The inventory could be readily expanded, and the tables modified, to suit different discourse-styles and lects.

In due course, the model will be integrated with improved higher-level rules for phrasing and accent-placement; and at the lower level, a set of microprosodic rules will adjust the essentially straight-line contours now generated (Fig 2) to enhance the phonetic naturalness of the output.

Meanwhile, the model avoids the intonational repetitiveness often associated with synthetic speech. In its present implementation, there is in fact no way of predicting precisely which set of contours will be applied to a given text. A further planned development will be a facility whereby particular patterns may be specified explicitly if required.

ACKNOWLEDGEMENTS

Thanks are due to our sponsors at the Speech Research Unit, RSRE, and to numerous colleagues at UCL for their comments and support.

REFERENCES

/1/ J.D. O'Connor & G.F. Arnold, *Intonation of Colloquial English*, Longman, 1961, 2nd ed 1973.

/2/ D. Crystal, *Prosodic Systems and Intonation in English*, Cambridge University Press, 1969.

/3/ G. Akers & M. Lennig, 'Intonation in text-to-speech synthesis: evaluation of algorithms', *JASA* 77, 2157-2165, 1985.

/4/ J. Pierrehumbert, 'Synthesising intonation', *JASA* 70, 985-995, 1981.

/5/ F. Nolan, 'Auditory and instrumental analysis of intonation', *Cambridge Papers in Phonetics and Experimental Linguistics* 3, Dept. of Linguistics, University of Cambridge, 1985.

/6/ M. Johnson & J. House, 'An accent-unit model of intonation for text-to-speech synthesis', *Proc. IOA: Speech and Hearing* 8, part 7, 409-416, 1986.

/7/ D.R. Ladd, 'Phonological features of intonational peaks', *Language* 59, 721-759, 1983.

/8/ J.A. Edward, "Rules for synthesising the prosodic features of speech', *JSRU Research Report* 1015, 1982.

Fig. 1: Schematised accent-unit contours

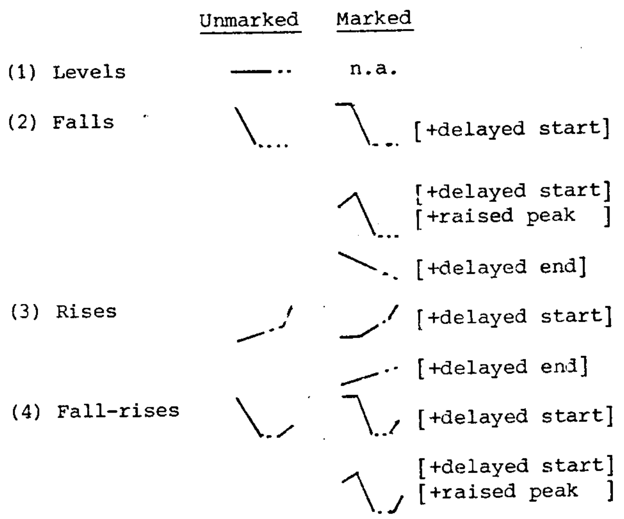
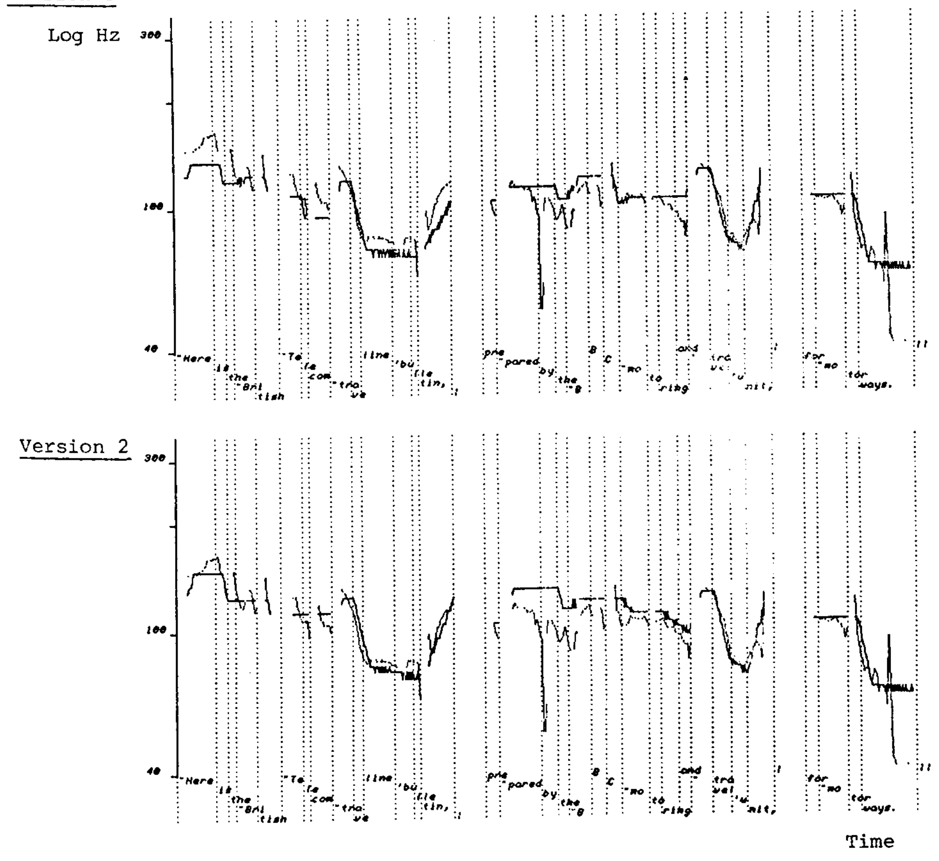


Fig. 2: Comparison between accent-unit contours and an F0 contour derived from natural speech

The solid line in each version is the accent-unit contour; the broken line is the contour derived from natural speech, aligned with the JSRU synthetic segmental durations:
"Here is the "British "Telecom "Traveline 'bulletin, pre"pared by the "BBC "Motoring and "Travel 'Unit, for "motorways.


Version 1

Version 2

In the contour generated by rule, H and L values are calculated in JSRU pitch levels. Interpolation between them is according to a 'moving-target' algorithm to prevent 'steppiness' in synthesis with a 100Hz frame rate.

INTRINSIC PITCH OF VOWELS : AN EXPERIMENTAL STUDY ON ITALIAN

MASSIMO PETTORINO

Ist. Universitario Orientale
Fonetica Sperimentale
Napoli, Italy

ABSTRACT

Many researches have studied the Intrinsic Pitch of the vowels in different languages and from different points of view. There is a general agreement on the existence of this phenomenon and various hypotheses have been formulated in order to explain the mechanism controlling the I. P. The aim of this experimental study is, on one hand, to verify whether the Intrinsic Pitch of vowels does exist in Italian; on the other hand, on the basis of the spectrographic data and Fo tracings obtained from normal and oesophageal speech and from singing, to try to give an account for the phenomenon. The relationship between Fo, opening degree and place of articulation is discussed.

INTRODUCTION

For more than fifty years there has been a general agreement among phoneticians about the existence of an Intrinsic Pitch of vowels. Many experimental studies have dealt with the phenomenon from different points of view and all of them have demonstrated that in many languages, as for instance English /1/, Danish /2/ and German /3/, high vowels tend to have a higher pitch than low vowels, other things being equal.

Even if there isn't disagreement on the existence of the I.P.. problems start when we try to explain why this phenomenon happens.

In fact different hypotheses have been formulated to give an account for the I. P. beginning from the so called "dynamo-genetic" theory proposed by Taylor /1/.

According to Taylor the higher muscular tension of the tongue required to realize a high vowel, radiates to muscles of the larynx causing a higher tension of the vocal folds that, therefore, vibrate at a higher fundamental frequency.

However, Taylor's theory is no longer accepted since "electrical insulation in muscles and nerves is good enough to prevent serial contraction of adjacent muscles by an osmotic spread of excitability" /4/.

Subsequent theories can be grouped into three main categories: "acoustic coupling", "aerodynamic" and "tongue pull" theories.

The first one, based on Flanagan's model /5/ and elaborated by Atkinson /4/, takes into consideration the formant pattern of the vowel: a low F1 attracts Fo giving rise to a higher pitch. This explains why /i/ and /u/, having a very low F1, have a fundamental frequency higher than that of /a/.

The second theory, formulated by Mohr /6/, relates the width of the pharynx with the glottal pressure. According to him, as the low vowels are characterized by a smaller pharyngeal cavity, the supraglottal pressure increases and consequently the transglottal pressure gradient decreases leading to a lower fundamental frequency.

According to the "tongue pull" theory, high vowels have a higher pitch because when the tongue rises it pulls the larynx up via the hyoid bone causing an extra tension of the vocal folds, either vertically (Ladefoged's view /7/) or horizontally (Newelkowsky's view /8/).

All these theories, which were based on experimental data, have subsequently been confuted on the basis of further data.

Therefore, as none of these hypotheses can explain the phenomenon of the I.P., Silverman /9/ is led to conclude that "the various physiological, acoustical and mechanical mechanisms that have been proposed to account for the IFO [I.P.] are not mutually exclusive, and probably are all operative during speech production" (p.13).

However, it seems to us that such an explanation is quite obvious because speech acts are complex and it always happens that a single articulation is characterized by many factors. The point is that, as regards I.P., it is necessary to distinguish between cause and effect, that is between what we really command to the articulators to do in realizing a high or a low vowel and what is merely a consequence of it.

The aim of this experimental study is, on one hand, to verify whether the I.P. of vowels does exist in Italian; on the other hand, on the basis of acoustic data obtained from normal and oesophageal speech and from singing, to try to give an account for the phenomenon.

PROCEDURE

A list of about 200 meaningful Italian words has been prepared. Vowels /i/ /e/ /ɛ/ /a/ /ɔ/ /o/ /u/ occur in initial and medial stressed position. The list includes words differing in the vowel only. The list has been read three times in a randomized order in an anechoic room by a native Italian speaker. Of each word a wide band spectrogram has been made using a Voice Identification Sound Spectrograph by Electronic ApS in order to have the formant pattern of the vowel. In order to calculate the fundamental frequency, a narrow band spectrogram at a linear expanded scale and the Fo tracing given by an FFM by F-J have been made. As in almost all cases the vowel had a rising-falling Fo movement with a maximum occurring at about its midpoint, we have measured the Fo value at that point.

RESULTS

Fig. 1 shows the average Fo values of the three utterances in normal speech.

As we can see, /a/ has always the lowest pitch, whereas the other vowels undergo an increase in pitch going from 4 to 20 Hz. Furthermore, we have to notice that /i/ and /u/ show an Fo increase higher than that of /e/ /ɛ/ /o/ /ɔ/, the former being in a range of 15 - 20 Hz and the latter of 4 - 10 Hz.

As we can see the data confirm the existence of an intrinsic pitch of vowels also in Italian.

In order to verify which of the different theories can explain the phenomenon of I.P., it seems useful to make further experiments.

If the I.P. is due to the raising of the tongue that causes an extra tension of the vocal folds, as suggested by the tongue pull theory, the phenomenon should be nullified in oesophageal speech. In fact, with the total laryngectomy surgery the whole larynx with the hyoid bone and all associated muscles and ligaments are removed. The voicing source, so-called neoglottis, is given by the surgically altered pharyngeal oesophageal sphincter. Therefore, as there aren't any direct interconnections between the tongue and the neo-glottis, according to the tongue pull theory, in oesophageal speech differences in I.P. between high and low vowels would not be expected.

In order to verify the tongue pull theory the same speech material has been uttered by a laryngectomized speaker. However we have restricted the list of words to /a/ /i/ and /u/ vowels because, as we have said above, the difference in pitch is most remarkable for these vowels.

The data show that the mean Fo of both /i/ (84 Hz) and /u/ (91 Hz) of oesophageal speech is higher than that of /a/ (75 Hz).

As we can see, the I.P. persists also in oesophageal speech and consequently we can exclude both the horizontal and vertical versions of the tongue pull theory.

These conclusions agree with the results obtained on oesophageal speech by Gandour and Weinberg /10/. They are in favour of

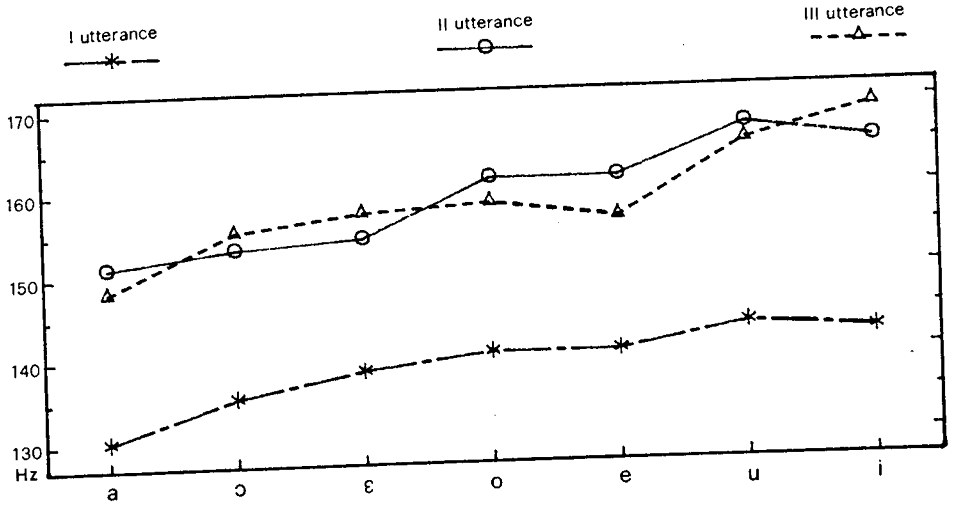


FIG. 1. Intrinsic pitch of Italian vowels.

the aerodynamic theory. In fact, according to them, the "impedance of the vocal tract is higher during the production of high versus low vowels. A natural response on the part of the speaker to this situation would be to increase respiratory drive or speech/vocal effort" (p.353),and in consequence of it high vowels would have a higher pitch.

However, it seems to us that the aerodynamic process is more complex and, therefore, it has to be reexamined in detail.

We know that glottal vibrations are determined by the difference between subglottal and supraglottal pressure : the lower supraglottal pressure is, the higher is the fundamental frequency. On the other hand as the supraglottal pressure depends on the opening degree of the constriction occurring along the vocal tract, there must be a close relationship between Fo and opening degree too.

In order to clarify this relationship, we have made an experiment on singing. We have analysed the Fo trend in monotone VCV sequences, where V was /a/ /i/ or /u/ and C was in turn a stop, a dental fricative, a lateral, a nasal or a trill.

Fig. 2 shows the average Fo values of the consonants.

As we can see, the data clearly demonstrate the existence of a direct relationship between Fo and opening degree of consonants. In fact we have the maximum pitch fall in stops and fricatives because of the high flow resistance at the articulatory constriction and it is nullified in nasals and laterals because of the free outscape of air through either the nasal cavities or the sides of the tongue. The clearest example of the existence of such a relationship is given by the Fo trend of the trill. In fact, in this case, Fo increases and decreases alternately with about 10 Hz and 40 Hz simultaneously with the dental openings and closings (fig. 3).

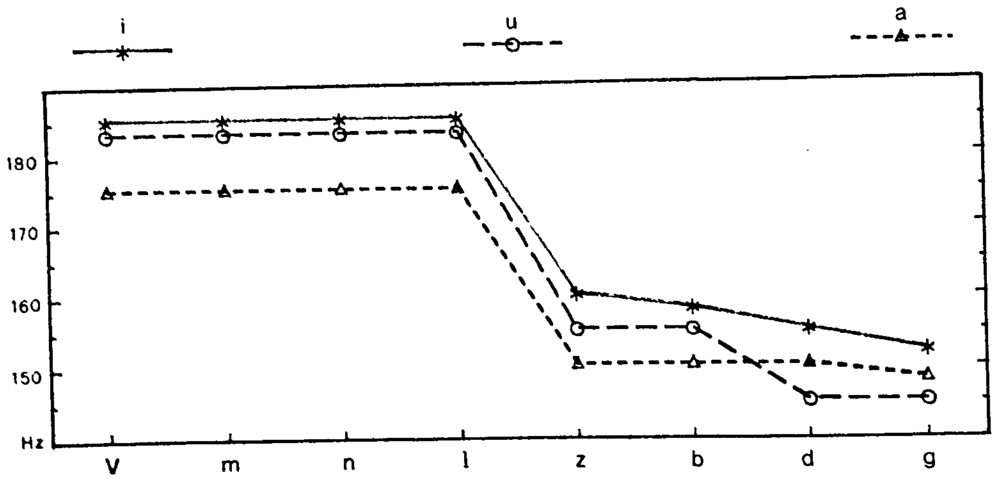


FIG. 2. Average values of Fo in singing.

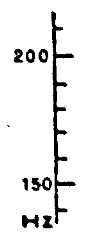


FIG. 3. Fo tracing of /arra/ in singing.

In the light of these considerations, we must conclude that the more open the consonant is, the higher is its pitch.

However, as regards the vowels, the data show that also in singing /i/ and /u/ have an increase in pitch of about 10 Hz respect to /a/.

At first sight the data seem to be contradictory because as regards the consonants the more narrow the constriction is the lower is Fo, whereas as regards the vowels it seems to happen the contrary, the more narrow the constriction is, the higher is Fo. The point is that when we classify the vowels as "high" and "low", or "close" and "open", we refer only to the oral cavity; conversely if we take the whole vocal tract into consideration, we realize how measleading is this kind of definition. In fact, as we can see in fig. 4, X-ray tracings of the Italian vowels /a/ /i/ and /u/ show that all these vowels are characterized by a same impedance occurring at different places along the vocal tract: at the pharyngeal cavity for /a/, at the soft palate for /u/ and at the hard palate for /i/.

From this point of view we must consider /a/ /i/ and /u/ as "close" vowels and consequently their different fundamental frequencies must be related to the point along the vocal tract where the maximum impedance occurs and not to the oral opening degree. From this point of view, we can easily understand why /a/ has an intrinsic pitch lower than that of /i/ and /u/. In fact a constriction in the pharyngeal cavity causes a sudden increase of the supraglottal pressure that leads to a drop of the transglottal pressure and consequently to a lowering of the fundamental frequency.

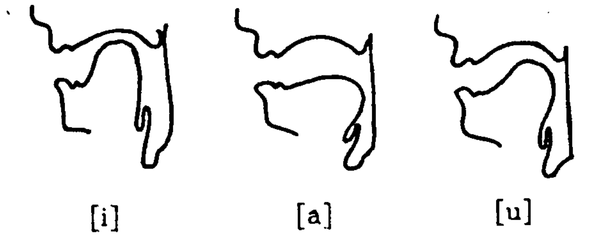


FIG. 4. X-ray tracings of /i/ /a/ /u/.

In light of this, we can give also an account for the acoustic coupling theory. According to this theory, a low F1 attracts Fo giving rise to a higher pitch. Now, we know that from an articulatory point of view a low F1 corresponds to a constriction occurring in the front half of the vocal tract and, therefore, to a wide pharyngeal cavity. So, once more it is the pharyngeal width to determine the higher pitch for /i/ and /u/, that is just the opposite of what happens for /a/.

CONCLUSIONS

The data gathered in this experimental research confirm that the phenomenon of the intrinsic pitch exists in Italian in normal speech as well as in singing and in oesophageal speech. Furthermore, the phenomenon .must be explained exclusively from an aerodynamic point of view, considering on one hand the configuration of the whole vocal tract during the production of the vowels and, on the other hand the pressure trend at glottal and supraglottal level.

As regards the tongue pull theory, even though many experimental studies have proved that there is a mechanical connection between the tongue and the larynx, our experiment on oesophageal speech clearly shows that it has nothing to do with the phenomenon of I.P.

As regards the acoustic coupling theory, suffice it to say that, as we have said above, the rising or lowering of a formant must always be seen as the effect of an articulatory gesture, even though we must admit that such an explanation is less evocative than the hypothesis according to which two frequencies attract each other because of their closeness.

REFERENCES

/1/ H.C. Taylor, "The Fundamental Pitch of English Vowels", Journal of Experimental Psychology, 16, 565-582, 1933.

/2/ N. Reinholt Petersen, "Intrinsic Fundamental Frequency of Danish Vowels", Journal of Phonetics, 6, 177-189,1979.

/3/ Z. Antoniadis, H.W. Strube, "Untersuchungen zum Intrinsic Pitch Deutscher Vokale", Phonetica, 38, 277-290, 1981.

/4/ J.E. Atkinson, "Aspects of Intonation in Speech: Implications from an Experimental Study of Fundamental Frequency", PhD Dissert., University of Connecticut, 1973.

/5/ J.L. Flanagan, "Speech Analysis, Synthesis and Perception", Springer, Berlin, 1965.

/6/ B. Mohr, "Intrinsic Variations in the Speech Signal", Phonetica, 23, 65-93, 1971.

/7/ P. Ladefoged, "A Phonetic Study of West African Languages", Cambridge Univ. Press, 1964.

/8/ G. Neweklowsky, "Spezifische Dauer und Tonhohe der Vokale", Phonetica, 32,38-60, 1975.

/9/ K. Silverman, "What causes Vowels to have Intrinsic Fundamental Frequency?" Cambridge Papers in Phonetics and Experimental Linguistics, 3, 1984.

/10/ J. Gandour, B. Weinberg, "On the Relationship between Vowel Height and Fundamental Frequency: Evidence from Esophageal Speech", Phonetica, 37, 344 - 354, 1980.

Grateful acknowledgment is made to Umberto Cinque for his precious technical collaboration.

VOWEL INTRINSIC PITCH IN STANDARD CHINESE

SHI BO and ZHANG JIALU

Institute of Acoustics
Academia Sinica
Beijing, China

ABSTRACT

We investigated whether an intrinsic pitch (IP) effect occurs in Standard Chinese and if it exists how IP and pitch level interact with each other. The fundamental frequencies (F0) of each 9 Chinese vowels at different tonal points were measured in three cases: (1) in a monosyllable, (2) in the word-initial and (3) the word-final position of a disyllabic word. The test items (400 monosyllables and 509 disyllabic words) were embedded in a frame sentence and uttered by 5 male and 5 female informants. The results show that the characteristics of IP are to be found in all four different tones of Standard Chinese in spite of the fact that those tones have different F0-patterns. Further, the higher the relative pitch value, the larger the difference in F0 among the vowels. The IP differences are reduced in word-final position. These results suggest a new hypothesis.

INTRODUCTION

Intrinsic pitch (or intrinsic F0) describes the influence of tongue height of vowels on the F0-value associated with them: high vowels have higher average F0-values than low vowels when other factors are kept constant. A great deal of research has been devoted to the analysis and quantification of intrinsic pitch in several languages: English, Italian, Danish, Japanese, French, German, Greece, Taiwanese Chinese, Yoruba, Serbo-Croatian, Itsekiri, and Chinese. IP has also been observed when vowels were sung at the same pitch. The reference list can be found in [1].

Various experimental conditions were applied in these studies. In the early experiments, isolated 'real' words as well as 'nonsense' words were used. The segmental environments (i.e. consonantal context) were carefully controlled. Later, the test words were embedded in a frame sentence. The effects of prosodic environment on IP had been took into account. Petersen [2] reported that the magnitude of IP in stressed syllables is larger than the one in unstressed syllables. Similar results were obtained for Italian accent/nonaccent words [3]. All of these studies generally showed similar results except Umeda's [4] which reported that there were no consistent IP effects in a 20-min reading by two speakers. In order to investigate whether IP effects occur in connected speech, Ladd and Silverman [5] compared test vowels (in German) in comparable segmental and prosodic environments under two different experimental conditions: (1) a typical laboratory task in which a carrier sentence served as a frame for test vowels; (2) a paragraph reading task in which test vowels occurred in a variety of prosodic environments. It was shown that the IP effect does occur in connected speech, but that the size of the IP differences is somewhat smaller than in carrier sentences. They pointed out that Umeda's finding was questionable because she apparently had not made any attempt to control for the prosodic environment of the vowels that were measured. In a recent study, Shadle [6] investigated the interaction of IP and intonation in running speech. She examined the F0 of the vowels [i,a,u] in four sentence positions. The results showed a large main effect of IP that lessened in sentence final position .

However, none of these studies were concerned with the roles of pitch level and the position in the word in affecting intrinsic pitch. The main goal of the present experiment was to get a general idea about the effect of intrinsic pitch in Standard Chinese. The effect was to be studied as a function of the following variables: (1) pitch level (in different tones); (2) position in disyllabic words (word-initial and word-final).

METHOD

The material consists of two parts, 400 monosyllables and 509 disyllabic words. All possible combinations of consonants and simple vowels in Standard Chinese were included in the monosyllable part, and each combination occurs four times with four different tone patterns. Among them there are 279 'real' monosyllabic words and 121 'nonsense' words. In the disyllabic word part, every word consist of one test syllable (a simple vowel preceded by an initial consonant) and one matched syllable. The matched syllable was chosen in such a way that the test vowels could be compared in a similar segmental environment and the same tonal surroundings. Examples are fǎhuà/fǔhuà; wēibā/wēibō/wēibī; tújǐng/tíxǐng, (the test syllables are underlined). Of the test syllables 273 were in word-initial and 236 in word-final position. As many combinations of two tones as possible were involved in this part.

In order to make all test items be in the same phonetic environment and to approach the situation of connected speech, all the monosyllables and disyllabic words were embedded in the frame sentence /Wǒ dú __ zì./ (I utter the character __.) and / Wǒ dú __ __ zhè gè cí./ (I utter the word __ __.) respectively.

Ten speakers (5 males and 5 females) of Standard Chinese were recorded. They had been trained for a short period before the recordings. A natural speech style was aimed at. The test materials were read once by each speaker in an acoustically treated room.

The recordings were fed into a Visi-Pitch (model 6087) for the extraction of F0. The counter on the Visi-Pitch provides a digital display of F0 for sustained vowels while the cursor allows the user to determine the F0 of any point on the pitch curve shown on the screen with ±1 Hz accuracy.

Fig.1 shows the measuring points of F0. They are: for high tone (T1) the middle point T1; for rising tone (T2) the lowest point T2-1 and the highest point T2-2; for dipping tone (T3) the starting point T3-1 and the lowest point T3-2; for falling tone (T4) the highest point T4-1 and the lowest point T4-2.

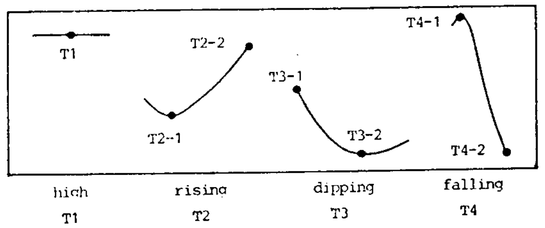


Fig.1 Measuring points of fundamental frequency

As a first step we only cared about average IP differences between vowels but ignored the differences between consonantal context and interspeaker variation. The statistical method was a one-way analysis of variance (with speakers and consonantal environments as a repeated measure).

RESULTS

1. Vowels Intrinsic Pitch In Four Tones

The data which will be analysed in this section were derived from 400 monosyllables. The intrinsic F0-values for each of 9 vowels and relative FO differences (ΔFO) between the vowel [a] and the remaining 8 vowels at different tonal points are given in Table 1. in which the data are mean values averaged across consonants, for 5 males and 5 females respectively. This is also shown graphically in Fig.2 (see •—•).

The data mentioned above permit us to make the following observations: 1) at points T1, T2-2, and T4-1, the F0-values of the vowels go from high to low as the tongue height of the associated vowel drops, and the F0 differences between high and low vowels are significant; 2) at points T2-1 and T3-2 a high vowel also has a higher F0 except that the F0-value of [o] of the males is a bit higher than that of [ʅ] and [i] and the FO-value of [a] of the females is higher than what is expected. The data at these five points show that Chinese, as a tone language, also exhibits the influence of intrinsic pitch.

The situation is more complex at points T3-1 and T4-2. We found considerable inter- and intra-speaker variability for F0-values at point T3-1. The main problem at point T4-2 is that the energy at the end of T4 is very low and the periodicity is not good enough to permit precision in measurements. As a result there is no consistent influence of IP at these two points.

Table 1. Mean intrinsic F0-value for each of the 9 Chinese vowels and relative FO differences (ΔFO) between the vowel [a] and the remaining 8 vowels at different tonal points, derived from 400 monosyllables, averaged across consonantal contexts, and for 5 males and 5 females respectively.

FO and ΔFO (Hz)

	T1		T2-1		T2-2		T3-1		T3-2		T4-1		T4-2	
	FO	ΔFO	FO	ΔFO	FO	ΔFO	FO	ΔFO	FO	ΔFO	FO	ΔFO	FO	ΔFO
(5 males)														
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
ɚ	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
ɚ	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

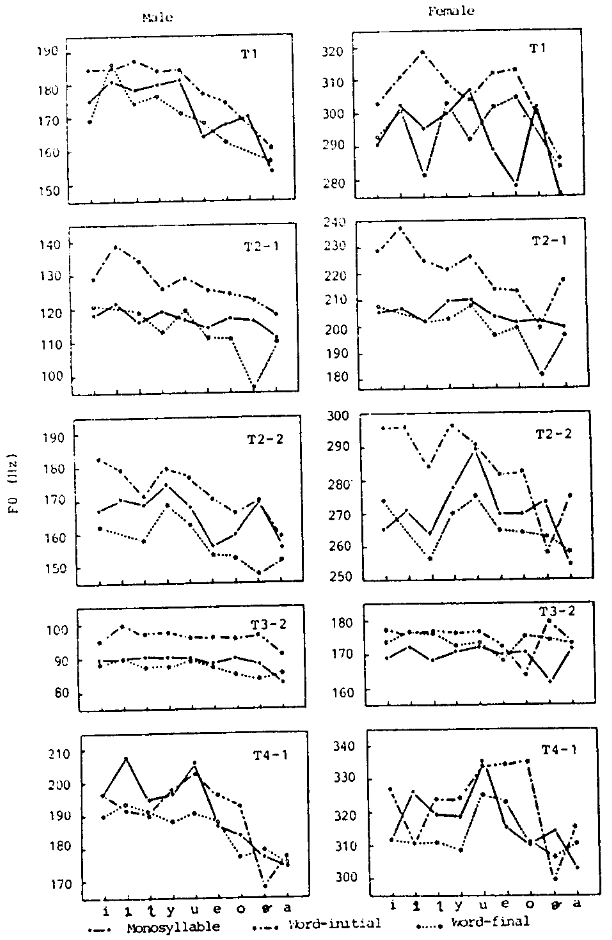


Fig.2 Mean F0 for the vowels, plotted as a function of tongue height, averaged across consonants and for 5 males and 5 females respectively

2. Effect of Word-position on IP

There are additional factors influencing F0-value of the vowels when the test syllables were in the disyllabic words. For instance, the F0-pattern of the test syllable could be modified by the adjacent tones as well as might vary with different stress pattern caused by different semantic meaning.

The data in Fig.2 (★and ☆) show that, though semantic meaning and tonal environment are not separated in the data, the effect of intrinsic pitch still occurs regardless of whether the test vowels were in word-initial or word-final position. So the variation of tonal characteristics due to intrinsic pitch is larger than the one due to semantic and tonal environment factors. However the magnitude of IP was reduced in word-final position. This reduction appears to be related to a lowering of F0 in this position (in Fig.2, the curves derived from the word-final position are the lowest ones in most cases).

3. Interaction of Intrinsic Pitch with Pitch Level

Fig.2 shows F0-values of 9 simple vowels as a function of the tongue height associated with them. Generally speaking, in each part of Fig.2, from left to right, the tongue height of the vowel goes from high to low and it is accompanied by a drop in F0, which reflects the effect of IP. But the curves in Fig.2 at different tonal points have different slopes, i.e. the differences of intrinsic F0 (ΔF0) across the vowels vary from point to point (also see Table 1.). The ΔF0 at points T1 and T4-1 (high F0) are obviously much larger than those at point T3-2 (lower F0).

Going a step further, there is little difference in ΔF0 between the males and the females in spite of the fact the F0 of the females is higher than that of the males. It indicates that the magnitude of ΔF0 is directly proportional to some kind of relative pitch value rather than to the absolute F0-value. In tone languages, 'tonal value' and 'tonal register' are often used to describe the relative relationships of pitch values. If we call the absolute F0-minimum as F0(min) and F0-maximum as F0(max), then the tonal value T(p) (in Oct.) for F0(p) (in Hz) is the binary logarithm of the quotient of F0(p) and F0(min). When F0(p) is equal to F0(max), the T(max) is the tonal register. Thus

Tonal value: $T(p)=\log_2(F0(p)/F0(min))$ Oct.

Tonal register:$T(max)=\log_2(F0(max)/F0(min)$ Oct.

The ΔF0 between i-a and between u-a are plotted as a function of the normalised tonal value (=T(p) divided by T(max)) in Fig.3. The '•' represents the averages over i-a and u-a across the males and the females. It is obvious that the higher the tonal value, the larger the ΔF0. In other words, the IP is more marked in the high frequency region of the tonal register than in the lower one.

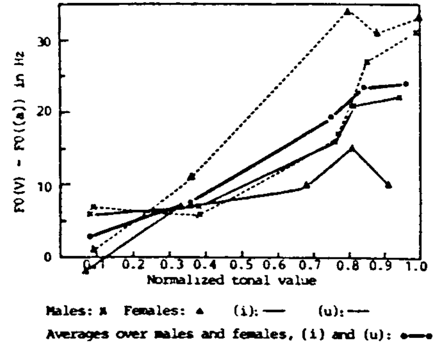


Fig.3 Mean F0 differences between (i,u) and (a) are directly proportional to the normalized tonal value

But the slopes of the two curves of the females turn negative when the normalised tonal value is bigger than 0.8. It seems that when the F0-value goes beyond certain limits, the direct proportional relation between ΔF0 and F0 will no longer be tenable. This suggests that is might be worth while to study intrinsic pitch in a larger F0 dynamic range such as in singing.

DISCUSSION

There have been various hypotheses for the cause of IP: dynamogenetic irradiation hypothesis[7], source/tract coupling hypothesis [8], pressure hypothesis [9], and tongue pull hypothesis.

Of the various hypotheses, it seems that the tongue pull theory has received the greatest attention. The early tongue pull hypothesis [10] supposed that the tongue, when raised to produce high vowels, pulls the hyoid bone and the larynx upwards, thus resulting in an increased vocal-fold tension which in turn leads to a higher F0. But this explanation is contradicted by the fact that the hyoid/larynx position always seems to be lower in [u] than in [a]. Ohala [11] modified the tongue pull hypothesis. He thought that the increased vertical tension in the vocal folds through the mucous membrane and other soft tissues without involving the hyoid bone and the hard tissues of the larynx. In support of this explanation, it appears that there is a positive correlation between ventricle size, which is assumed to reflect vertical tension in the vocal folds and tongue height and intrinsic F0 of vowels. The tongue pull hypothesis has been expanded further by Ewan [12]. Ewan suggests that the low F0 of low vowels, which are also assumed to involve a tongue retraction or pharyngeal constriction component, is caused by the soft tissues being pressed downwards in the direction of the larynx and thus increasing the vibrating mass of the vocal folds, which results in a decrease in F0.

But few of these hypotheses attempt an explanation of the 'nonlinearity' in IP. In Chinese the higher the tonal value, The larger the IP difference; in Italian, the accented syllables display greater IP than unaccented ones [3]; deaf speakers often exhibit a larger than normal IP which may be related to a higher than normal average F0 [13]. IP is reduced in final sentence position with a lowered F0 [6]. The common point of these results is that a larger IP difference seems always correlated to a higher F0. Moreover, the variation of tonal characteristics due to syntactic and semantic factors is much larger at the tonal roof than at the tonal floor [14]. So a larger variation of F0 always corresponds to a higher F0. And this sort of nonlinearity is relative to a within- subject variation (i.e. it does not mean the female should be expected to have a larger IP difference than the male because of a higher voice). There was a simpler explanation that general relaxation (as in an unaccented phrased-final position) may reduce intrinsic F0. But it is contradicted by the evidence against vowel neutralization in that 'relaxed' sentence position [6].

Here, we try to give a probable interpretation from the point of inherent nonlinearity 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 series of deformations in the mucous membrane and the soft tissues, and finally in the vocalis muscle itself thus causing increased tension. The relationship between the tension T and the elongation x of the vocalis muscle can be approximately expressed as:

$$T=ae^{bx}$$

The incremental tension per unit elongation, as given by $\partial T/\partial x(=abe^{bx})$ is obviously greater at larger values of x which generally correspond to higher F0-values. In other words, the same incremental elongation due to the tongue pull could cause a larger increase in tension T, thus leading to a larger F0 variance at high F0 than at low F0. However, it must be emphasized that this is only a probable conjecture. The reliable evidence for the interpretation should be based on physiological data. Last, we think that if this kind of nonlinearity in the production of speech could be confirmed, it would be helpful for a better understanding of the similar nonlinearity found in the perception of speech.

ACKNOWLEDGEMENTS

We would like to express our thanks to Prof. Eva Gårding and Prof. Maa Dah-You for their valuable suggestions and critical reading of the early versions.

REFERENCES

[1] Shi, Bo and Zhang Jialu. 1986. Vowel intrinsic pitch in Standard Chinese. Working Papers 29: 169-190. Dept. of Linguistics and Phonetics. lund Univ..

[2] Petersen,N.R. 1978. Intrinsic fundamental frequency of Danish vowels. J. of Phonetics (1978) 6: 177-189.

[3] Ferrero,F.E. et al. 1975. Some acoustic and perceptual characteristics of the Italian vowels. Paper presented at the VIIIth Int. Congr. Phon. sci. (Leeds), mimeographed.

[4] Umeda,N. 1981. Influence of segmental factors on fundamental frequency in fluent speech. J. Acoust. Soc. Am. 70: 350-355.

[5] Ladd,D.R. and Silverman,K.E.A. 1984. vowel intrinsic pitch in connected speech. Phonetica 41: 31-40.

[6] Shadle,C.H. 1985. Intrinsic fundamental frequency of vowels in sentence context. J. Acoust. Soc. Am. 78: 1562-1567.

[7] Taylor.H.C. 1933. The fundamental pitch of English vowels. J. Exp. Psychol., 16: 565-582.

[8] Flanagan,J. and Landgraf,L. 1968. Self-oscillating source for vocal-tract synthesizers. IEEE Transactions on Audio and Electroacoustics AU-16, 57-64.

[9] Mohr,B. 1971. Intrinsic variations in the speech signal. Phonetica, 23: 65-93.

[10] Ladefoged,P. 1964. A phonetic study of West African languages. In West African Language Monograph, Series I (Cambridge U.P., Cambridge).

[11] Ohala.J. 1973. Explanations for the intrinsic pitch of vowels. Monthly Internal Memorandum, Phonology Laboratory, Univ. of California, Berkeley (January), 9-26.

[12] Ewan,W.G. 1975. Explaining the intrinsic pitch of vowels. Paper presented at the Fifth California Linguistics Association Conference, San Jose, May 4, 1975, 1-9.

[13] Bush,M. 1981. Vowel articulation and laryngeal control in the speech of the deaf. PhD thesis, MIT (unpublished).

[14] Bannert,R. 1984. Towards a model for German prosody. Working Papers 27: 1-36. Dept. of Linguistics and Phonetics, Lund Univ..

MICROPROSODIC FUNDAMENTAL FREQUENCY VARIATIONS IN GERMAN

BERND MÖBIUS, ALICE ZIMMERMANN, WOLFGANG HESS

Institut für Kommunikationsforschung und Phonetik,
Universität Bonn, Poppelsdorfer Allee 47,
5300 Bonn 1, FRG

ABSTRACT

As a preliminary study in the analysis of German sentence intonation, this contribution deals with two types of segmentally conditioned fundamental frequency (F_0) variations: the influence of stop consonants on F_0 of following stressed vowels (coarticulatory F_0 variation; CF_0), and the differences in F_0 between high and low vowels (intrinsic fundamental frequency; IF_0). These microprosodic phenomena are recorded, evaluated and discussed in detail for one speaker. Apart from speaker-specific variations, the results are qualitatively quite consistent with data reported in the literature /1, 3, 8/. In intonation research, the CF_0 effect may be neglected, provided that the exact point of F_0 measurement is chosen appropriately, whereas IF_0 differences have to be evaluated for each speaker separately.

INTRODUCTION

The temporal course of voice fundamental frequency (F_0), as the pre-eminent representative of sentence intonation in German, may be considered as the result of several interacting factors. A thorough description and interpretation of fundamental frequency tracings must account for at least the following factors that affect the course of F_0 in different ways: microprosody, i.e., intrinsic fundamental frequency of vowels and coarticulatory F_0 variations; the distribution of word and sentence accents; sentence position; and modality /15/.

This contribution deals with the microprosodic factors, namely the influence of initial stop consonants on F_0 of stressed vowels (coarticulatory F_0 variation; CF_0), and the differences in F_0 between high and low vowels (intrinsic fundamental frequency; IF_0). These phenomena may be defined as variations of the speech signal which depend on the acoustical and physiological constraints of the human speech production system.

The aim of this study is to record and evaluate these microprosodic effects for one speaker and to eliminate them as factors disturbing the interpretation of intonation contours at the sentence level. The relevant procedures described in the literature /e.g. 2, 7, 11, 15/ generally imply an undesirable restriction in the choice of the test material. Considering the large inter-subject variation of the microprosodic effects /1, 6, 10/ the procedure described here seems to be fairly successful.

EXPERIMENTAL DATA ON GERMAN

Intrinsic Fundamental Frequency (IF_0) of Vowels: Test Material

In the first part of our investigation we evaluated the intrinsic fundamental frequency (IF_0) of the German vowels for one speaker. Within the key words vowel quality was varied as well as vowel quantity. According to the results of earlier studies of German microprosody /1, 9/ we expected systematic higher IF_0 values for high vowels compared to low vowels. The findings for the influence of vowel quantity on IF_0 are less clear-cut. It is true that the IF_0 differences between open and closed vowels tends to be more distinct in short vowels than in long ones; but a significant difference is solely stated by Antoniadis and Strube /1/. For our speaker the difference was insignificant.

The key words were embedded in a short carrier sentence of the form "Ich habe ... gesagt" ("I said ..."). The choice of a relatively short carrier sentence enabled us to control the intonation contour of the whole utterance /cf. 5/. The test material consisted of German words with the exception of the key words "Kir" and "Punk" which are borrowed from French and English, respectively. Nevertheless these two words may be considered as elements of present-day German; they were uttered with the usual German pronunciation, i.e., [ki:ɐ] and [paŋk]. Within the key words we examined and analysed the long and short stressed vowels of German: /a:/, /ɛ:/, /e:/, /i:/, /ø:/, /y:/, /o:/, /u:/, and /a/, /ɛ/, /I/, /œ/, /Y/, /ɔ/, /U/.

Coarticulatory F_0-Variations (CF_0): Test Material

In the second part of our investigation we evaluated the coarticulatory influences of initial stop consonants on the fundamental frequency of vowels by varying the place of articulation as well as the voicing of the plosives. The key words containing the initial German plosives /p/, /t/, /k/, /b/, /d/, /g/ were embedded in the aforementioned carrier sentence "Ich habe ... gesagt".

Procedure

Since we expected the microprosodic F_0 variations to be strongly dependent on the individual speaker, we decided to process the utterances of only one speaker in the first step of the experiment. The test sentences were spoken by a male subject (DL) three times each. The recording was carried out at three different days within two weeks. The test sentences were typed on cards and presented to the speaker in random order. The subject, a native speaker of Standard German, was instructed to pronounce the sentences successively with a few seconds' interval.

The material was recorded in an anechoic chamber using a professional microphone and tape recorder. Before digital processing we checked whether the test sentences were uttered with the same underlying intonation contour in all cases. A group of four listeners had to identify the sentence modality unanimously as declarative. The syllable containing the test vowel had to be realized with nuclear stress, i.e., rising F_0, controlled by means of a Frøkjaer-Jensen Transpitch Meter output. Utterances that did not meet these requirements were eliminated for the moment and recorded again; this proved to be necessary in a total of 14 cases.

The fundamental frequency extraction was carried out by an algorithm /13/ that represents the speech oscillogram in high temporal resolution. The program enables determining the duration of fundamental periods and calculating the actual F_0 values for each period. The results are presented and discussed in the following section.

RESULTS

The values of intrinsic fundamental frequency for the German long and short vowels presented in figure 1 were determined as follows. Presuming that our studies into microprosody are subordinate to intonation analysis on the sentence level, we looked for a way to condense the F_0 microstructure of a vowel in one single representative value. Two procedures with two measuring points seem to be particularly suitable:

a) Arithmetic mean $\bar{x}$: To begin with, we cut off the first and last third of the temporal course of the vowel. Then the arithmetic mean is calculated for the F_0 values of the remaining (second) third. This procedure is motivated by the fact that the influences of neighbouring segments may be considered relatively small in this quasi-stationary part of the vowel /1/.

b) 2/3 value: This is the momentary value of fundamental frequency at two thirds of the duration of a vowel /12/. This well-established method has been applied repeatedly in the literature /cf. 2/. Rossi defines a regularity which says that a pitch movement is not perceived by listeners as a whole; on the contrary, the perceived pitch of a vowel corresponds to a value of fundamental frequency that is measured at the boundary between the second and the last third of the temporal course of F_0. The 2/3 value may be regarded as representative for linearly rising or falling F_0 glissandos.

The results show - as we had expected- essentially higher IF_0 values for high vowels compared to those of low vowels, with the exception of the short /U/ which has even lower IF_0 values than /ɛ/. With the tongue height being equal, back vowels show higher IF_0 than front vowels, which is in accordance with data reported in the literature /1, 10/; /U/ is the exception here, too. The values for /i:/ and /I/ turned out somewhat lower than expected, a finding that may be caused by the structure of the test material containing several key words with final vocalized /r/, such as "Pier" or "Tier". This ought to be controlled by using other test words containing the vowels /i:/ or /I/, respectively, and final obstruents.

Two essential results of our investigation concerning the coarticulatory fundamental frequency variations are illustrated by figures 2 and 3. Figure 2 gives a detailed representation of the averaged F_0 tracings for the CV combinations "voiceless plosive + long vowel" and "voiceless plosive + short vowel". The figure shows that vowel quantity influences the actual F_0 values rather than the whole contour. Furthermore the influence of the initial plosive has decayed after at most 50 ms. This is important for further measurements. The influence of coarticulatory F_0 variations may be neglected when the point of measurement is chosen appropriately, that is, at least 50 ms after the vowel onset.

Figure 3 supports the hypothesis /1, 2/ that the place of articulation of the stop consonant has no significant influence on the course of fundamental frequency in the following vowel.

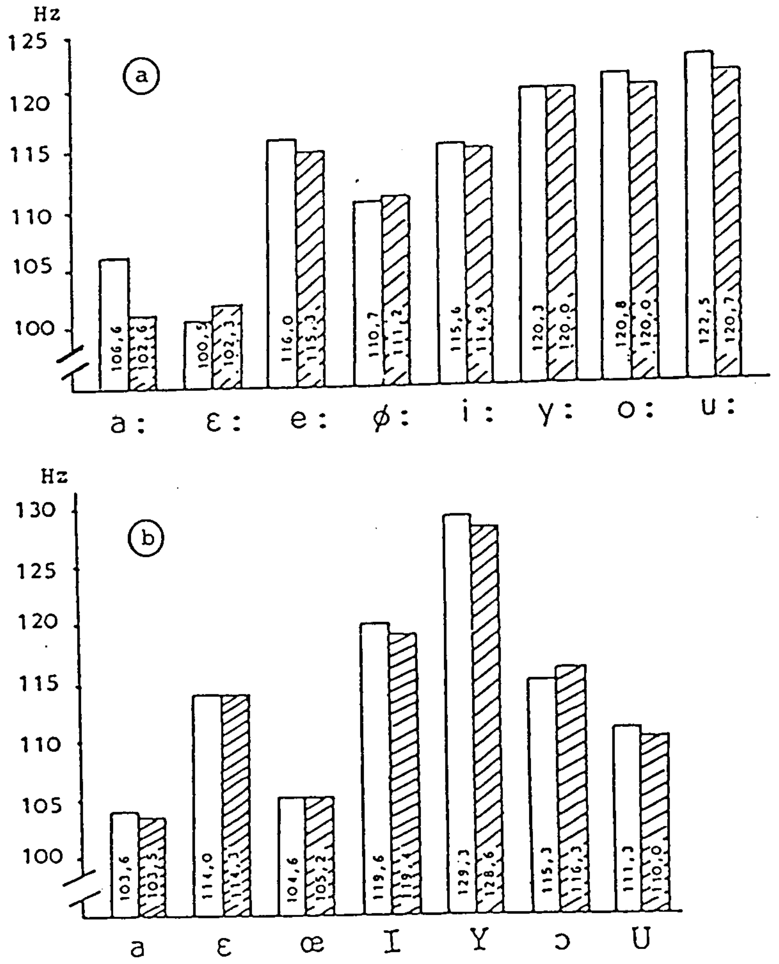


Figure 1a, b. Intrinsic fundamental frequency values of the German vowels. (a) Long vowels; (b) short vowels. F_o calculation for the 2/3 value (/12/; white surfaces) and for the arithmetic mean $\bar{x}$ (hatched surfaces). All utterances by one speaker (DL, male)

DISCUSSION

There remain at least two problems deserving discussion here, in particular for the part of our study that deals with the influence of initial stop consonants on the course of fundamental frequency of vowels.

The vowel onset after voiced and voiceless plosives is mostly reported in the literature to be rising or falling, respectively. We also share this observation. In a recent study Silverman /14/ argues that this so-called "rise-fall dichotomy" - falling F_o after voiceless stops and rising F_o after voiced ones - is an artifact brought about by the structure of the test sentences. In the great majority of investigations the key words are integrated within a carrier sentence in nuclear-stress position. In many languages, however, this position is marked by a rising underlying intonation contour. This constellation, applying to our data as well, is in Silverman's opinion and in accordance with the results of his experiments the ideal condition for an apparent dichotomy of rising and falling F_o contours.

Furthermore we proceeded from the assumption that the influence of a stop consonant on vowel F_o is purely progressive, i.e., only following vowels are affected /3, 6, 8/. In two recent studies, however, Kohler /4, 5/ showed that the F_o microstructure may also contribute to the discrimination of postvocalic lenis and fortis obstruents.

We hold the view that Silverman's arguments as well as Kohler's findings will have to be taken into account in the choice of test material in future studies

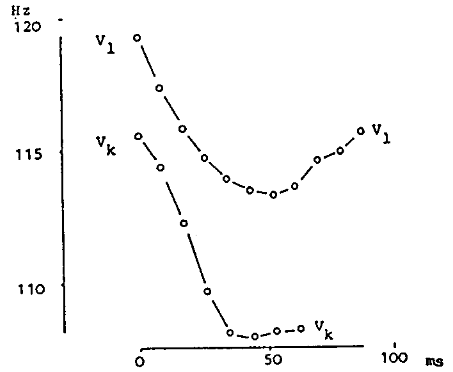


Figure 2. Coarticulatory F_o variations. The figure shows the F_o course of the CV combinations "voiceless plosive + short vowels" (V_k) and "voiceless plosive + long vowels" (V_l). Vowel onset at 0 ms. All utterances by one speaker (DL, male)

and in the interpretation of data reported in the literature.

CONCLUSION

The procedure presented here will allow us to study and analyse intonation on the sentence level without considering the interfering influences of the microprosody. The proposed points of measurement - arithmetic mean $\bar{x}$ and 2/3 value - are both found representative of vowel fundamental frequency. The actual variations of the intrinsic fundamental frequency of high and low vowels, however, will have to be determined for each speaker separately.

REFERENCES

/1/ Antoniadis Z., Strube H.W. (1981): "Untersuchungen zum 'intrinsic pitch' deutscher Vokale". Phonetica 38, 277-290

/2/ Di Cristo A., Hirst D.J. (1986): "Modelling French micromelody: Analysis and synthesis". Phonetica 43, 11-30

/3/ Jeel V. (1975): "An investigation of the fundamental frequency of vowels after various Danish consonants, in particular stop consonants". Annual Report of the Institute of Phonetics 9, 191-211 (Copenhagen)

/4/ Kohler K.J. (1982): "F_o in the production of lenis and fortis plosives". Phonetica 39, 199-218

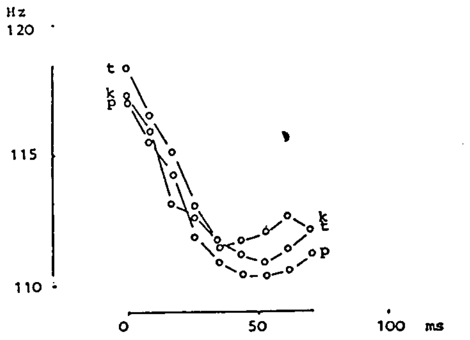


Figure 3. Coarticulatory F_o variations. The figure shows the F_o course of the combinations "(/p/, /t/, /k/) + vowel". Vowel onset at 0 ms

/5/ Kohler K.J. (1985): "F_o in the perception of lenis and fortis plosives". J.Acoust.Soc.Am. 78, 1 (1), 21-32

/6/ Lehiste I., Peterson G.E. (1961): "Some basic considerations in the analysis of intonation". J.Acoust. Soc.Am. 33, 419-425

/7/ Lyberg B. (1984): "Some fundamental frequency perturbations in a sentence context". J.Phonetics 12, 307-317

/8/ Mohr B. (1971): "Intrinsic variations in the speech signal". Phonetica 23, 65-93

/9/ Neweklowsky G. (1975): "Spezifische Dauer und spezifische Tonhöhe der Vokale". Phonetica 32, 38-60

/10/ Reinholt Petersen N. (1976): "Intrinsic fundamental frequency of Danish vowels". Annual Report of the Institute of Phonetics 10, 1-27 (Copenhagen)

/11/ Reinholt Petersen N. (1986): "Perceptual compensation for segmentally conditioned fundamental frequency perturbation". Phonetica 43, 31-42

/12/ Rossi M. (1971): "Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole". Phonetica 23, 1-33

/13/ Sendlmeier W.F., Stock D. (1983): "Ein Computerprogramm zur Manipulation digitaler Sprachsignale mit einer Anwendung zur Synthese nach dem Diphonkonzept". Fachberichte des Instituts für Phonetik und sprachliche Kommunikation 17, 1-16 (München)

/14/ Silverman K. (1986): "F_o segmental cues depend on intonation: The case of the rise after voiced stops". Phonetica 43, 76-91

/15/ Thorsen N. (1979): "Interpreting raw fundamental frequency tracings of Danish". Phonetica 36, 57-78

F_0 PERTURBATIONS IN HINDI

LIESELOTTE SCHIEFER

Institut fur Phonetik und Sprachliche Kommunikation
der Ludwig-Maximilians Universitat München, FRG

ABSTRACT

F_0 perturbations were measured for Hindi voiceless, voiceless aspirated, voiced, and breathy voiced stops combined with the phonemically long vowels /a e i o u/ in word-initial position and isolated word production. The analysis revealed: (i) significant differences between the stop manners of articulation for about 90 ms after release, (ii) significant influence of the tongue height of the vowel, and (ii) less influence of the place of articulation of the stop.

INTRODUCTION

F_0 perturbations after stops have been thought of as unavoidable by-products of the stop articulation and have been treated as a secondary cue for the perception of stop's manner of articulation. Ohde's studies [3] for example revealed a falling F_0 contour after voiceless and a lower F_0 onset but still falling contour after voiced stops. It was shown by Lea [2] and Umeda [5] that the influence of the preceding stop had disappeared in non-tone languages like English after 75 to 100 ms. These analyses of F_0 perturbations focused on voiced and voiceless aspirated stops; thus voiceless unaspirates and breathy voiced stops were rarely included in these studies. Moreover, little attention has been paid to the influence of the vowel on the F_0 perturbations. Hence, the aim of our investigation is threefold: (i) to provide data from a language with a four-way contrast within the stop categories, (ii) to present results from breathy voiced stops, (iii) to test the influence of either the place of articulation of the stop and the tongue height and tongue position of the vowel on the F_0 trajectory after stop release.

MATERIAL, INFORMANT AND PROCEDURE

A list of words was prepared containing the voiceless, voiceless aspirated, voiced, and breathy voiced stops in four places of articulation (labial, dental, retroflex, and velar) combined with the phonemically long vowels of Hindi in word-initial position in randomized order. There were some gaps in the material as only common words were chosen (cf. Table 3). One subject (female, 35 years) born in Simla (Himachal Pradesh) and raised in Simla and New Delhi served as informant. The material was tape-recorded in Munich (for further detail cf. Schiefer [4]), digitized on a PDP11/50 computer (sample rate 20 KHz) and filtered with a cut off frequency of 8 kHz. The periodic portions of the initial CV syllable were segmented manually into single pitch periods cf. [4], starting with the first visible period. The analysis was based on the first ten pitch periods after stop release. This covers a period of approximately 40-50 ms. F_0 was measured additionaly for pitch period 20, which is about 80 to 90 ms away from the burst. Separate two-factorial analyses of variance were conducted for the stop and vowel conditions for (i) all stops in general, (ii) the voiceless, (iii) aspirated, (iv) voiced, and (v) breathy voiced stops over the first ten pitch periods. The results for the 20th pitch period were calculated separately.

RESULTS

The influence of stop-manner. Fig. 1 displays the results for the stops in general. All F-values and p-values for the main effects of place (A), vowel (B) and interaction (I), P1 to P10, and P20 are listed in Table 1 for the stop conditions. The difference between the stops remains significant till pitch period 20. All stops differ in respect to the F_0 onset and F_0 contour. The onset is highest in voiceless stops (henceforth VL), lower in aspirated (ASP) and voiced stops (VD), and lowest in breathy voiced stops (BRE). VL stops have a falling, ASP a rising-falling, VD a falling, and BRE stops a rising pattern. The fall is steepest in VL, less steep in VD stops. All stops differ significantly from each other in respect to the onset (P1). At pitch period P20 the stops fall into two groups: VD and VL with lower F_0 and BRE and ASP with higher F_0 values.

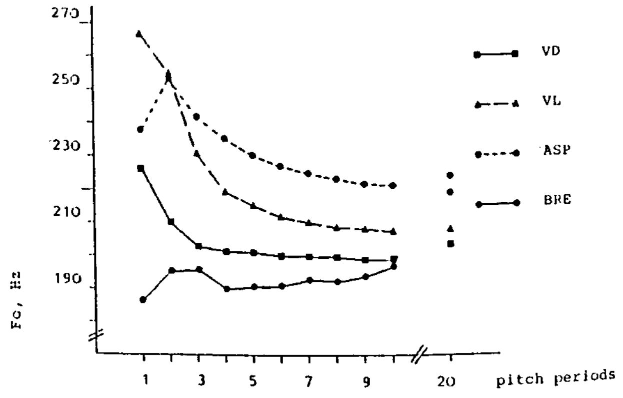


Fig.1: Fo values as a function of pitch periods and stop's manner of articulation

The influence of the place of articulation in VL stops. Figure 2 shows the results for the VL stops. The [+apic] stops /t/ and /ṭ/ show the highest F_0 onset, whereas the [-apic] stops /p/ and /k/ have lower F_0 frequencies at the onset with an overlap between both groups. The contour is falling for all stops except /k/ which shows a rising-falling pattern. Fig. 3 shows the results for the ASP stops. All stops have high F_0 values at vowel onset. F_0 increases from pitch period P1 to P2, and decreases continously from P2 to P20. Only /th/ differs significantly from the other stops. VD stops show a different pattern (cf. Fig. 4 and Table 1) from either the VL or ASP stops as F_0 is lower at the vowel onset, where the [-ant] stops /ḍ/ and /g/ have higher and the [+ant] stops /d/ and /b/ lower F_0 frequencies. But only /b/ differs significantly from the other stops. The difference remains significant till pitch period P10. At P20 the difference fails to reach significance. All stops except /ḍ/ show a slightly falling F_0 contour from P1 to P2/P3 and a slightly rising F_0 from P4 to P20. /ḍ/ has a steeper fall from P1 to P2 and a falling/level pattern towards the end of the trajectory. In breathy voiced stops (cf. Fig. 5 and Table 1) the F_0 is low at vowel onset and shows a rising pattern from P1 to P20. The differences at the onset are small; [+ant] stops have slightly higher F_0 frequencies than [-ant] stops, but they do not differ significantly from each other. The differences between the stops in pitch periods P4 to P20 are caused by the significantly higher F_0 values for the velar stop /gh/.

TABLE 1: Effect of place of articulation on F_0 as a function of the number of pitch periods. A represents the overall effect of place, B that of the pitch periods, and I the interaction.

	VL stop		ASP stop		VD stop		BRE stop	
	F	p	F	p	F	p	F	p
P1	77.0	***	1.2	n.s.	33.3	***	4.2	*
P2	13.7	***	3.1	*	13.8	***	9.4	***
P3	17.1	***	7.0	**	10.4	***	10.2	***
P4	2.7	*	8.7	***	12.4	***	8.4	***
P5	2.4	n.s.	6.3	**	12.6	***	12.8	***
P6	0.9	n.s.	5.4	**	12.1	***	19.7	***
P7	0.3	n.s.	5.6	**	15.2	***	22.8	***
P8	0.3	n.s.	5.1	**	13.4	***	29.7	***
P9	0.2	n.s.	4.4	*	14.7	***	26.9	***
P10	0.1	n.s.	3.5	*	16.3	***	28.7	***
P20	0.7	n.s.	1.9	n.s.	3.9	n.s.	2.7	*
A	5.0	**	23.2	***	27.5	***	44.1	***
B	89.4	***	10.3	***	4.2	***	9.4	***
I	3.7	***	0.2	n.s.	2.6	**	1.5	n.s.

The influence of the vowel on the F_0 contour. The results for the VL stops are shown in Fig. 6, the F-values and p-values for all vowel conditions are given in Table

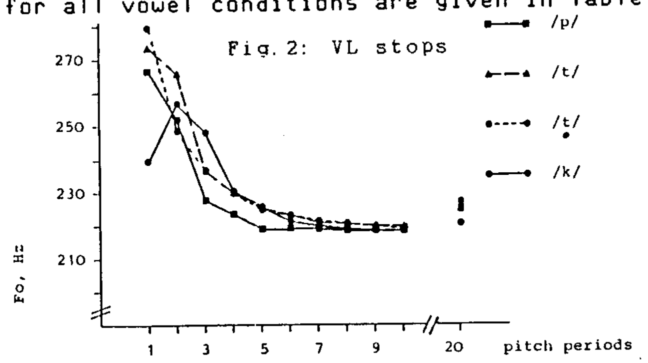


Fig. 2: VL stops

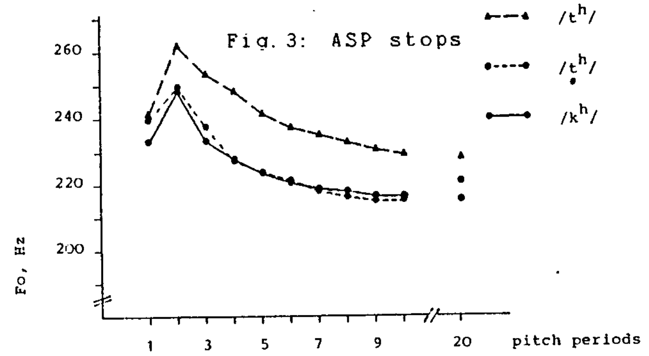


Fig. 3: ASP stops

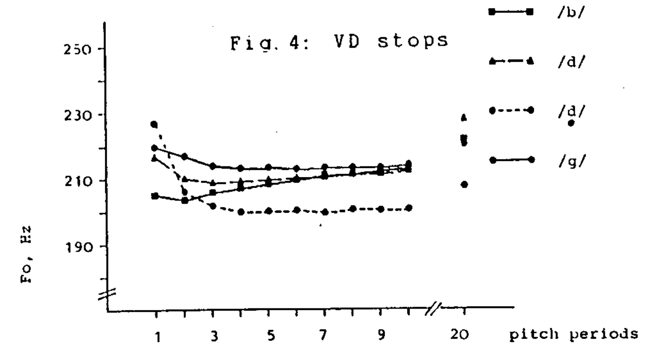


Fig. 4: VD stops

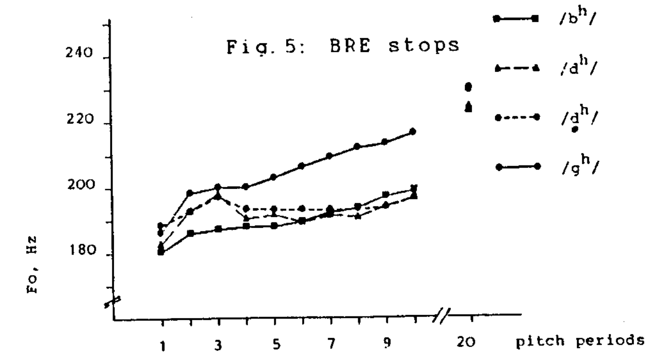


Fig. 5: BRE stops

Fig. 2 to 5: Fo values as a function of pitch period and place of articulation of the stop

2. The F. at vowel onset is a function of the tongue position of the vowel: central and back vowels have higher F. frequencies than front vowels. The F. trajectory for the vowels differs: /a/ shows a steep falling, the mid vowels /e o/ a falling, and the high vowels /i u/ a rising-falling pattern. The difference between F. onset and endpoint of the trajectory is again a function of tongue height: it is greatest for /a/, less for the mid vowels, and smallest for /i u/. These results thus reflect the well known effect of "intrinsic pitch". Fig. 7 displays the results for the ASP stops. The differences between the vowels at F. onset are greater in ASP stops compared to the VL ones. F. is obviously determined by the tongue position of the vowel: central and back vowels lead to higher F. onset than do front vowels. All vowels differ significantly from each other: /a/ shows a nearly level contour for the first two pitch periods, and a falling contour from P2 to P10, whereas the mid vowels have a rising-falling and /i/ a rising-falling-rising pattern. In VD stops F. at vowel onset (cf. Fig. 8) is determined by the tongue position of the vowel: F. ist lowest for mid, higher for high, and highest for low vowels. But only /e/ differs significantly from the other vowels. The vowels differ, too, in respect to the F. trajectory: /a/ shows a steep fall, /e o i/ a short fall followed by a rising contour, whereas /u/ has a rising pattern throughout. In BRE stops (cf.Fig.9), the F. onset is low. The influence of the vowel is smallest but significant at vowel onset. Back vowels show a slightly higher F. than non-back vowels. For detailed discussion cf. [4].

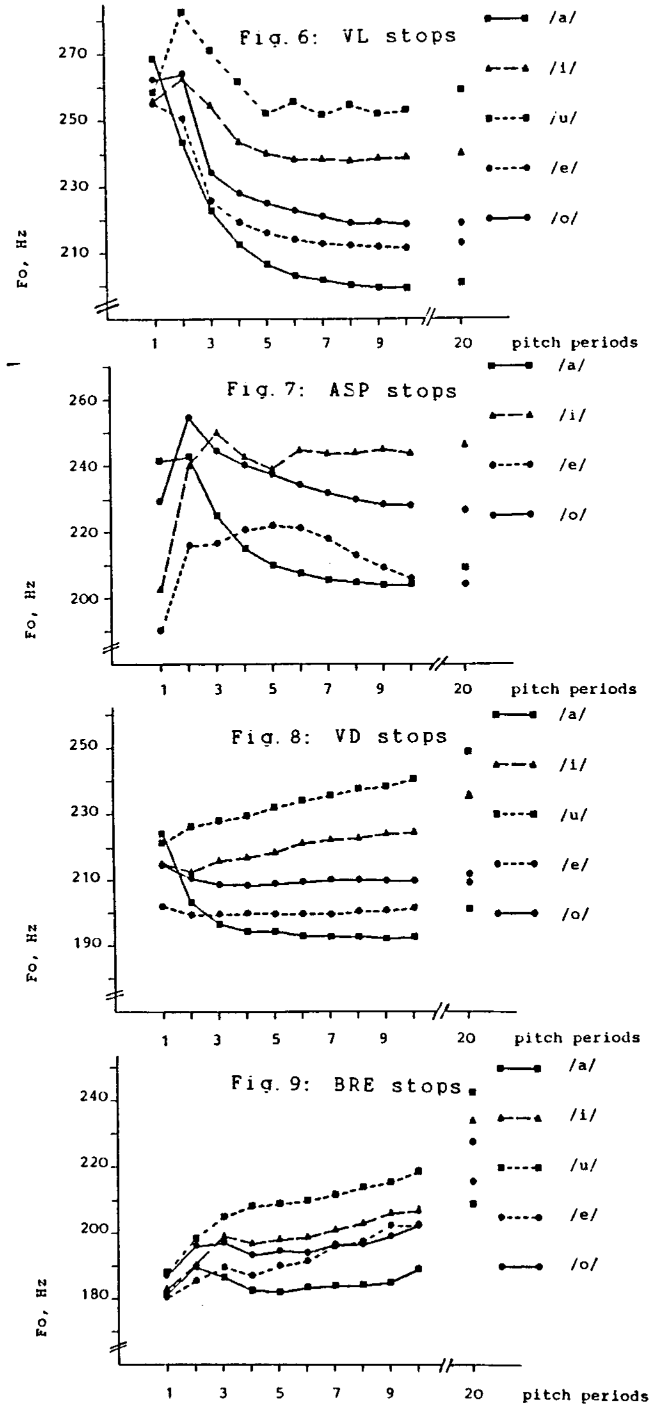


Fig. 6: VL stops

Fig. 7: ASP stops

Fig. 8: VD stops

Fig. 9: BRE stops

Fig.6 to 9: Fo values as a function of pitch period and the vowel

TABLE 2: Effect of the vowel as function of the pitch periods.

	VL stop		ASP stop		VD stop		BRE stop	
	F	p	F	p	F	p	F	p
P1	9.8	***	75.5	***	50.7	***	3.9	*
P2	71.3	***	35.3	***	13.8	***	9.4	***
P3	140.6	***	33.8	***	115.8	***	18.9	***
P4	130.7	***	26.1	***	135.2	***	34.0	***
P5	112.6	***	26.2	***	157.3	***	34.3	***
P6	141.7	***	35.7	***	190.4	***	32.8	***
P7	132.6	***	37.6	***	207.6	***	34.2	***
P8	153.6	***	41.7	***	219.3	***	40.0	***
P9	146.2	***	48.3	***	235.4	***	42.8	***
P10	155.1	***	49.0	***	252.0	***	40.0	***
P20	61.1	***	20.0	***	35.4	***	14.8	***
A	252.2	***	82.0	***	355.5	***	61.6	***
B	87.0	***	8.0	***	2.9	*	13.3	***
I	5.2	***	6.0	***	6.0	***	1.8	n.s.

Interaction between place of articulation and vowel. In order to compare our results with those from Ohde's studies we have calculated the F. difference between the first and second pitch period (cf. Tab. 3). Concerning the stop manners in general a F. fall was measured in VL and VD stops, whereas ASP and BRE stops show a rising pattern. This can be explained by the different timing and width pattern of the glottis for these two groups of stops (cf. [1]). The effect of place of articulation in general shows a falling pattern in VD and VL stops (with the exception of /k/). and again a rising pattern for the ASP and BRE stops. Concerning the interaction between place of articulation, vowel and stop's manner of articulation the results can be summarized as follows: (i) ASP stops cause a rise from P1 to P2. There is no interaction between place of articulation and vowel. (ii) The F. is rising after BRE stops with only a few exceptions which can be neglected due to the minimal differences between P1 and P2 in these examples. The results for /bho/ cannot be explained. (iii) VL and VD stops show similiar patterns and large interactions between place and vowel. In labial position [+back] vowels cause a rising F. contour, whereas in dental position the same effect is caused by [+front] vowels. In combination with the VL stop /u/ leads to a rising F. pattern, too. F. is falling in the retroflex position for both stop manners with the exception of /i/ after VL stops, which cause a rising F.. The differences between both stop manners are greatest in the velar position: F. rises after VL stops, whereas after VD a rise can be observed only with [+high] vowels.

GENERAL DISCUSSION

Concerning the stop manners of articulation in general our results verify those found by Lea and Umeda, as the differences remain significant for about 90 ms. and they are in good agreement with those found by Ohde: VL, ASP and VD stops cause a falling F. after vowel onset. The F. onset is higher for VL than for ASP stops. The overall F. is on the other hand higher after ASP than after VL stops. The places of articulation do not influence the F. onset pattern in a systematic way. The stop manners form two different sets: (i) ASP and BRE stops on the one side and VL and VD on the other side pattern differently with respect to the F. onset and contour. Whereas ASP and BRE stops cause a F. rise from P1 to P2 (with only few exceptions within the BRE category) the VL and VD stops reflect similiar interactions between place of articulation and the vowel. This can be explained by the underlying difference in the laryngeal behavior during the production of these stops. In VL as well as VD stops the glottis is almost closed during the moment of articulatory release of the closure, whereas the glottis is open in BRE as well as ASP stops. Thus the F. onset is less affected by articulatory movements during the production of ASP and BRE stops, but is subject to greater influences in VL and VD stops. Some general differences between our results and Ohde's are obvious: (i) ASP stops cause a rising-falling pattern across all place and vowel conditions. This difference is systematic without any exception. It can be assumed that the different methods applied in these studies may be responsible for this effect. (ii) VL and VD stops do not cause a falling F. pattern in all place or vowel conditions. The pattern is a function of the place of articulation and the tongue height and tongue position of the vowel. (iii) Concerning the interaction between place of articulation and vowel we found, in contrast to Ohde, a similiar pattern for the VD.and VL stops. This interaction is comparable to that of Ohde only in respect to the VD category.

TABLE 3: Interaction between place of articulation, vowel, and stop manner: difference between P1 and P2 in Hz. Positive values indicate a F. fall from P1 to P2, negative values a F. rise.

		VD	VL	ASP	BRE
means		7.9	8.7	-14.8	- 7.9
labial	/a/	11.2	50.5	---	-11.3
	/e/	3.3	31.0	---	- 6.4
	/o/	- 3.4	-14.6	---	6.9
	/i/	1.8	2.8	---	- 5.3
	/u/	- 4.2	-35.8	---	-11.5
	x	1.5	13.7	---	- 6.1
dental	/a/	13.8	49.0	- 5.0	0.4
	/e/	- 0.5	-14.1	-----	----
	/o/	11.4	9.3	-27.8	-15.3
	/i/	- 3.9	- 9.4	-----	-10.6
	/u/	1.7	-12.5	-35.0	-14.6
	x	6.6	7.7	-19.3	-10.1
retr.	/a/	42.7	51.7	- 0.8	- 5.4
	/e/	3.7	46.8	- 9.4	1.7
	/o/	8.4	36.7	-23.0	-13.6
	/i/	25.2	- 9.2	-24.0	2.1
	/u/	----	----	-----	- 6.7
	x	20.8	30.8	-10.0	- 4.1
velar	/a/	14.4	-20.5	- 1.5	-16.3
	/e/	4.3	-26.1	-37.2	- 4.4
	/o/	1.1	-31.1	-26.1	- 5.0
	/i/	- 2.2	-11.6	-42.9	-34.2
	/u/	-12.9	-17.4	-32.1	- 4.2
	x	2.9	-17.4	-15.0	- 3.0

REFERENCES

[1] Hirose, H.; Lisker, L.; Abramson; A.S.: Physiological aspects of certain laryngeal features in stops production. Haskins Lab. Status Rep. Speech Res., SR-31/32, pp.183-191 (Haskins Laboratories, New Haven 1972)

[2] Lea, W.A.: Segmental and suprasegmental influences on fundamental frequency contours; in Hyman, L. Consonant types and tones, pp.17-70 (Linguistic Program, University of Southern California, Los Angeles 1973)

[3] Ohde, R.N.: Fundamental frequency as an acoustic correlate of stop consonant voicing. J. acoust. Soc.Am. 75: 224-230 (1984)

[4] Schiefer,L.: F. in the production and perception of breathy stops: evidence from Hindi. Phonetica 43:43-69 (1986)

[5] Umeda, N.: Influence of segmental factors on fundamental frequency in fluent speech. J.acoust.Soc.Am. 70: 350-355 (1981)

INITIAL F_0-CONTOURS IN SHANGHAI CV-SYLLABLES - AN INTERACTIVE FUNCTION OF TONE, VOWEL HEIGHT, AND PLACE AND MANNER OF STOP ARTICULATION

LING KING - HARRY RAMMING - LIESELOTTE SCHIEFER - HANS G. TILLMANN

Institut für Phonetik und Sprachliche Kommunikation
der Ludwig-Maximilians Universität München, FRG

ABSTRACT

F_0 perturbations after voiceless and voiceless aspirated stops are analyzed in Shanghai, a tone language. It turned out that F_0 is always higher after voiceless than after aspirated stops, and this difference disappeares after 15 to 30 ms. The place of articulation of the stops does not contribute significantly to the F_0 difference, whereas the vowel does.

INTRODUCTION

For languages such as German and English it is well known that voiced stops cause an initial lowering of F_0 in CV-syllables and voiceless stops a relative raising. Umeda [6] and Lea [3] found that this effect remains evident during the first 75 to 100 ms of the vowel, whereas Hombert et al [1] showed that in a tone-language this effect disappeared after 40 to 60 ms. This raises the question whether this specific glottal behavior is language dependent or not. Most studies on F_0 perturbations after stops focused on the difference between voiced and voiceless aspirated stops rather than on the difference between the two voiceless categories (aspirated and unaspirated). On the other hand, those studies which examined that difference provided rather diverse results as in some of the languages voiceless aspirated stops caused higher F_0 values than the voiceless ones (Korean [2]), whereas other authors report the reverse (English [4]). A higher F_0 after aspirated stops was reported too for Cantonese by Zee [7], who measured the F_0 perturbations after [p] and [ph], respectively. The rather conflicting results cannot be explained easily as the studies differ in (i) number of speakers employed, (ii) material included (in most studies only a subset of either the stops or stop-vowel combinations is analyzed), and especially (iii) in method. With our present study we wanted to help contribute to a solution of the problem by employing further material from a tone language. Thus, the aim of our study is threefold: (i) to examine the F_0 perturbations caused by two voiceless stops in a tone-language, (ii) to measure the duration of these perturbations. and (iii) to analyze whether the F_0 perturbations interact either with tone, the stop's place of articulation, or the vowel.

MATERIAL AND INFORMANT

We constructed a list of words containing voiceless unaspirated (henceforth VL) and voiceless aspirated (henceforth ASP) stops in three places of articulation (labial, alveolar, and velar) followed by one of the vowels /a e i o u/ in word initial position and combined with one of the four tones high level (Tone 1), rising (Tone 2), mid level or dipping (Tone 3), and falling (Tone 4). As ASP stops do not occur with Tone 2, the difference between the stop's manner of articulation could be measured for Tone 1, Tone 3, and Tone 4 only. It should be mentioned here, that our analysis does not support the hypothesis of Zee-Maddieson [8] that Tone 2 is associated with voiced stops, as in the speech of our informant Tone 2 occurred (with few exceptions) only after VL, i.e., short lag stops. Not all possible combinations between stop, vowel, and tone occur in our material], as shown in Table 1.

TABLE 1: CV combinations as a function of manner of articulation and tone

		Tone 1 high level	Tone 2 rising	Tone 3 mid level	Tone 4 falling
		a e o i u	a e o i u	a e o i u	a e o i u
VL	lab	x - x x -	x x x x x	- x x x x	x x x x x
	alv	x - x x -	x x x x x	- x - - x	- x x x x
	vel	x - - - x	x x x - -	- x x - x	- x x - x
ASL	lab	x - x x -	- - - - -	x x x x x	x x x x x
	alv	x - x x -	- - - - -	x x x x x	x x x x x
	vel	x - x - -	- - - - -	x x x - x	x x x - x

Every word, containing one of the CV combinations, was written ten times on separate cards. The words were read by one informant (male, 34 years old), native speaker of Shanghai, but with imperfect knowledge of Mandarin. The recordings were made in our Institute on a Telefunken M15 studio tape recorder using a Neumann U87 microphone. The microphone was placed in front of the speaker at a distance of about 50 cm, who was seated comfortably in a chair. He was asked to read the words at a comfortable loudness and tempo. He was given the cards in randomized order and he had to read the words in the following way: after reading the first word (on the first card), he had to turn the card and put it aside before continuing with the next word. This procedure caused the speaker to read slowly and breathe after every word. We employed this method in order to avoid any kind of "list effects". The recordings were made in one session, interrupted by a pause of about 15 mins.

PROCEDURE

A preliminary analysis of the fundamental frequency was run with the Frokjer-Jensen F_0-Meter in order to check the realisation of the tones and to eliminate any mistake made by the speaker. The material was then digitized on a PDP11/50 with a sample rate of 20 kHz and filtered with a cut off frequency of 8 kHz. The first 15 pitch periods of the vowels in long syllables and the first ten periods in short syllables of Tone 1 were delimited manually with the help of a segmentation routine and stored for analysis (for detail cf. [5]). The F_0 was calculated separately for all CV conditions in all tones, averaged over all repetitions. Separate multivariate analyses of variance were applied for each tone condition.

RESULTS

The results of the statistical analysis are given in Table 2.

TABLE 2: Statistical results from the analysis of variance for Tone 1, Tone 2, Tone 3, and Tone 4, as well as the manners of articulation, places of articulation, and vowels included in the analyses. M=manner of articulation, P=place of articulation, V=vowel

	Tone 1	Tone 2	Tone 3	Tone 4
Manner	VL, ASP	VL	VL, ASP	VL, ASP
Place	l, a	l, a, v	l, al, v	l, a, v
Vowel	a, o, i	a, e, o, i, u	e, u	e, o ,u
Interactions				
M-P-V	p <.01	---	n. s.	n. s.
P-V	n. s.	n. s.	p <.05	n. s.
M-V	p <.01	---	p <.05	p <.001
M-P	n. s.	---	n. s.	p <.05
Main factors				
Manner	p <.001	---	p <.001	p <.001
Place	n. s.	n. s.	n. s.	p <.05
Vowel	p <.001	p <.001	p <.001	p <.01

Manner of articulation. The main effect of the stop's manner of articulation (cf. Fig. 1) is significant in all tone conditions (but cf. the interactions between manner of articulation and vowel). The F_0 onset is always higher after VL than after ASP stops. This effect disappears in Tone 1 after the third pitch period (P3), in Tone 3 after P6, and in Tone 4 after P5, respectively. This is equivalent to either 15, 30, or 25 ms.

Place of articulation. The effect of the stop's place of articulation is significant only in Tone 4 (p < .05), where the velar stop causes significantly higher F_0 values than the labial or alveolar stops. As there is an interaction between the manner and place of articulation, as plotted in Fig. 2, the main effect of the place cannot be interpreted by itself. It is apparent that the interaction is due to the velar VL stop /k/, which causes significantly higher F_0 values than the other stops /p/ and /t/, respectively.

Vowel. The main effect of the vowel is significant throughout. The difference between the vowels is usually greater after VL than after ASP stops and the results reflect the well known phenomenon of intrinsic pitch, where high vowels cause higher F_0 values than mid or low ones. As there is no interaction between the vowels and places of articulation, the results for Tone 2 and the VL stops are displayed in Fig. 3. In all vowel conditions F_0 falls from P1 to P2 and rises towards the end of the contour. The F_0 differences between the vowels are nearly the same at P2 as well as at P15. Tone 3 shows an interaction between place of articulation and the vowel (cf. Fig. 4), which obviously is due to a different behavior of the velar stop. Whereas the F_0 onset in /e/ is low after the labial and alveolar stop, it is higher after /k/, followed by a short fall instead of a rise. The differences are even greater for /u/, where the F_0 after /k/ is significantly higher than after the other stops.

Interaction between manner of articulation and vowel. In all tone conditions a significant interaction between the stops manner of articulation and the vowel can be observed. The results are plotted separately for the tones. Fig. 5 shows the results for Tone 1. The F_0 differences between the vowels are small after the ASP stops, greater after the VL ones; /i/ after the VL stops differs significantly from all other vowels. Fig. 6 displays the results for Tone 3. F_0 after the ASP stops is rising in /e/, level in /u/, whereas both vowels have a falling pattern after the VL stops; /u/ differs significantly from all other vowels. Tone 4 (cf. Fig 7.) shows a somewhat different pattern. This time, the interaction is caused by the ASP stops rather than the VL ones as the F_0 after VL stops is nearly the same for all vowels: the onset is high, followed by a F_0 fall. After ASP stops the F_0 onset is extremely low in /e/, which shows a rising-falling pattern, whereas the onset is high in /o/, followed by a quasi-linear F_0 fall towards the end of the contour. /u/ on the other hand, is associated with a high F_0 onset and a falling-rising pattern. To summarize these results it can be stated that (i) the interaction between manner and vowel is caused by a specific behavior of /i/ (Tone 1) and /u/ (Tone 3) after VL stops, and (ii) the different behavior of the ASP stops in Tone 4.

There is one higher level interaction

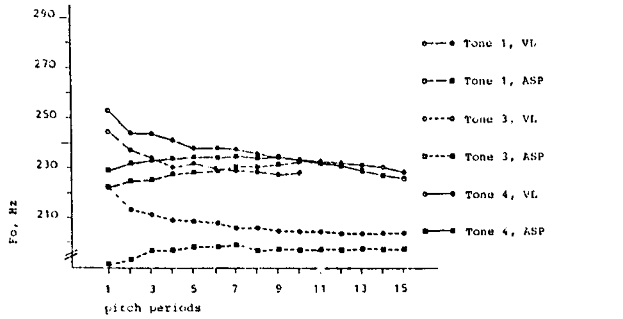


Fig. 1: Fo values in Hz for the manners of articulation as a function of pitch period and tone

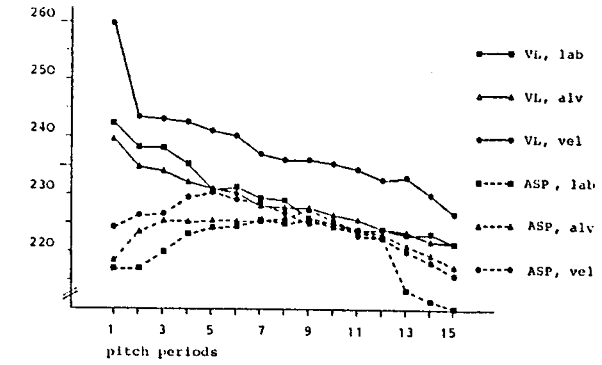


Fig. 2: Fo values in Hz for Tone 4 as a function of pitch period, manner and place of articulation

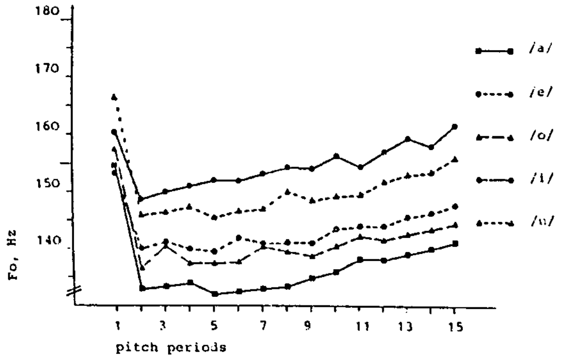


Fig. 3: Fo values in Hz for Tone 2 as a function of pitch period and vowel

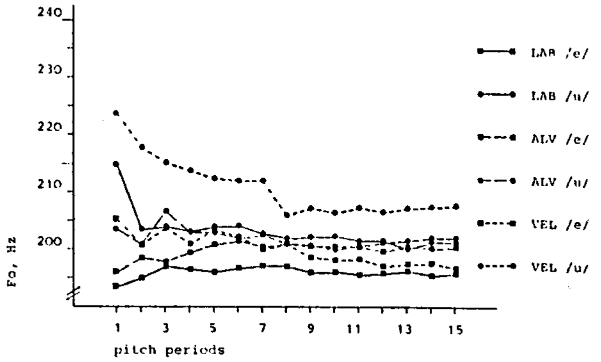


Fig. 4: Fo values in Hz for Tone 3 as a function of pitch period, place of articulation, and vowel

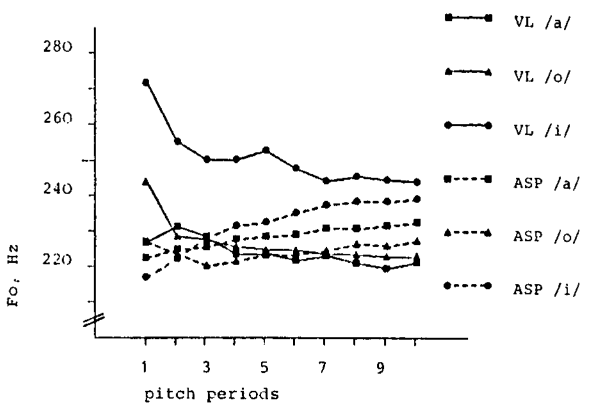


Fig. 5: Fo values in Hz for Tone 1 as a function of pitch period, manner of articulation, and vowel

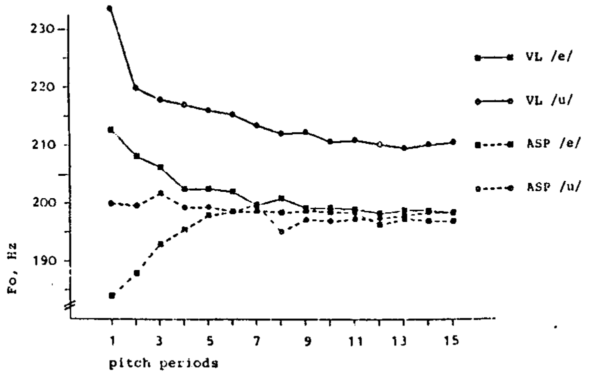


Fig. 6: Fo values in Hz for Tone 3 as a function of pitch period, manner of articulation, and vowel

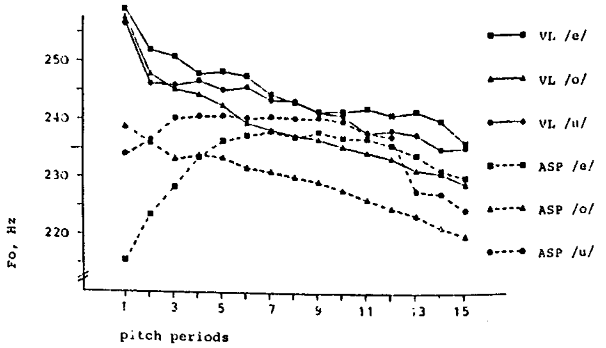


Fig. 7: Fo values in Hz for Tone 4 as a function of pitch period, manner of articulation, and vowel

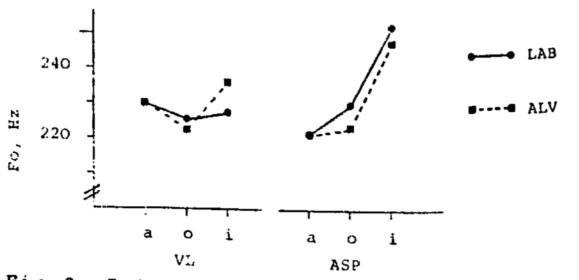


Fig. 8: Interaction between manner of articulation, place of articulation, and vowel in Tone 1. Fo values in Hz

between manner of articulation, place of articulation, and vowel in Tone 1. i.e. none of the factors examined contribute independently to the F_0 perturbations. The interaction (based on the mean F_0 values) is shown in Fig. 8. It is obvious that F_0 is higher after /p/ than after /t/; /a/ is associated with the lowest, /o/ with mid, and /i/ with the highest F_0. After /ph/ F_0 is highest in /a/, lower in /i/, and lowest in /o/, whereas after /th/ /i/ shows the highest F_0, /a/ mid, and /o/ lowest values. The differences for /a/ and /o/ are small, those for /i/ greater.
In order to have results comparable to those of Zee [7], who used the Cepstrum method to gain F_0 values and measured F_0 over the initial 78.0 ms of the vowels, we (i) give averaged F_0 values in Table 3 for Tone 1, Tone 3, and Tone 4, as well as the corresponding ms and (ii) analyzed the F_0 contour in /pey/ vs /phey/ in Tone 1, the results of which are displayed in Fig. 9.

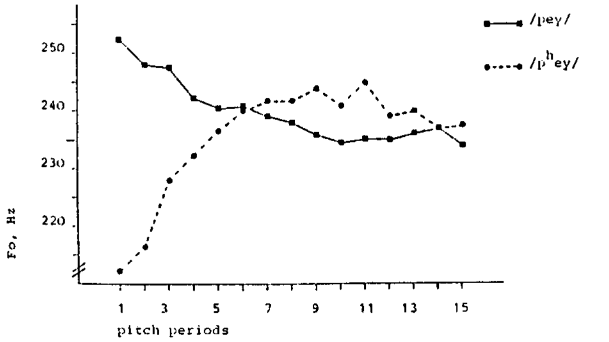


Fig. 9: Fo values in Hz as a function of pitch period and manner of articulation

It is clear from the averaged data that in our material the F_0 after VL stops exceeds those after ASP stops. On the other hand, the F_0 onset is high in /pey/ and falls towards the end of the contour, whereas it is low after /phey/ and rises till P6 where it exceeds the value of /pey/. The mean F_0 value for /pey/ averaged over about 65 ms (this corresponds to 15 pitch periods) is 239.4 Hz, that for /phey/ 235.6 ms.

DISCUSSION

To answer the question we have asked in the introduction it can be stated that there is a remarkable difference in F_0 after VL and ASP stops: F_0 is always higher after the VL than the ASP stops. This difference disappears after 15 to 30 ms. Our results thus are in agreement with those of Hombert et al [1] for Yoruba, as well as with those studies which reported higher F_0 values after VL stops [4], but disagree with the findings of Zee [7] for Cantonese. In the speech of our informant, the stops' places of articulation do not contribute significantly to the F_0 pattern. It is worth mentioning that the phenomenon of intrinsic pitch could be verified in a tone language, too. But the influence of the vowel is not independent of the stop's manner of articulation. The intrinsic pitch effect is greater after VL than after ASP stops and seems to interact with the tone too: the differences between the vowels are greater in Tone 2 than in Tone 1 or Tone 4. This factor cannot be discussed in detail here but will be dealt with in another paper. On the other hand, the interaction between the VL stop and high vowels seems to reflect a stronger coupling between the supra- and subglottal cavities after VL than after ASP stops.

TABLE 3: Mean F_0 values in Hz for the VL and ASP stops in Tone 1, Tone 3, and Tone 4, as well as the duration of the vowel portion in ms.

	Tone 1		Tone 3		Tone 4	
	F_0	duration	F_0	duration	F_0	duration
VL	232.3	43.0	209.5	47.7	240.1	41.7
ASP	228.5	43.8	196.8	50.8	233.2	42.9

REFERENCES

[1] Hombert, J.-M. - Ohala, J.J. - Ewan, W.G.: Phonetic explanations for the development of tones. Language 55: 37-58 (1979)

[2] Kagaya, R.: A fiberoptic and acoustic study of the Korean stops, affricates and fricatives. Journal of Phonetics 2: 161-180 (1974)

[3] Lea, W.A.: Segmental and suprasegmental influences on fundamental frequency contours; in Hyman, L. Consonant types and tones, pp. 17-70 (Linguistic Program, University of Southern California, Los Angeles 1973)

[4] Ohde, R.: Fundamental frequency as an acoustic correlate of stop consonant voicing. J. acoust. Soc. Am. 75: 224-230 (1984)

[5] Schiefer, L.: F_0 in the production and perception of breathy stops: evidence from Hindi. Phonetica 43: 43-69 (1986)

[6] Umeda, N.: Influence of segmental factors on fundamental frequency in fluent speech. J. acoust. Soc. Am. 70: 350-355 (1981)

[7] Zee, E.: The effect of aspiration on the F_0 of the following vowel in Cantonese. UCLA Working Papers in Phonetics 49: 90-97 (1980)

[8] Zee, E. - Maddieson, I.: Tones and tone sandhi in Shanghai: Phonetic evidence and phonological analysis. University of California Working Papers in Phonetics 45: 93-129 (1979)

ВСТРЕЧНЫЙ ДВУНАПРАВЛЕННЫЙ ОТСЧЕТ МОР В НГАНАСАНСКОМ ЯЗЫКЕ

ЕВГЕНИЙ ХЕЛИМСКИЙ

Институт славяноведения и балканистики АН СССР
Москва, 125040, Ленинградский проспект, д. 7

ABSTRACT

The principle of mora-counting is applied in Nganasan (Tawgy Samoyed) in two ways. A rule of consonant gradation posits the strong grades of consonants and consonant clusters before the even vocalic morae, and their weak grades before the odd vocalic morae. The morae in this case are counted from the beginning of a word. On the other hand, the stress is normally placed on the penultimate mora and the eventual additional stress on the pre-pre-penultimate mora of a word. This dual and counter-directed mora-counting must be related to different stages of the Nganasan linguistic history.

I. Среди самодийских и других уральских языков нганасанский выделяется сложностью и неординарностью своей морфонологии, наглядно демонстрируя, что такой типологический признак, как агглютинативная прозрачность морфемного состава слова, может свободно сочетаться с высокой развитостью фузионных явлений. Особенно заметную роль играет система чередования ступеней, которая распространяется на интервокальные шумные согласные и некоторые их сочетания: h (*p) : b, t : δ (*d), k : g, s : d́ (*j), ŋh (*mp) : h и ŋh : mb, nt : t и nt : nd, ŋk : k и ŋk : ŋg, ns : nd́, ʔt : t, ʔk : k, ʔs : s и др. (первой всюду указана сильная ступень, после двоеточия -слабая). Фонетическое качество отдельно взятого согласного, как можно видеть, не определяет его места в чередовании ступеней: сильная ступень одного чередования может совпадать со слабой ступенью другого, ср. k в соотношении с g и с ŋk или ʔk.

Уже в первом описании нганасанского языка у М.А.Кастрена /I/ было выделено два вида чередований ступеней.

С одной стороны, имеет место ослабление согласных в начале (исходно) закрытого слога, особенно часто и четко проявляющееся в двусложных основах, ср. kətu "ноготь, коготь" : Gen. kəδuŋ (> совр. kəδu; утрата ауслаутного -ŋ, происшедшая в нганасанском языке уже "после Кастрена", не привела к устранению чередования, а лишь превратила его из фонетически детерминированного в морфологически детерминированное), məku "спина" : Gen. məgu(ŋ), kintə "дым" : Gen. kində(ŋ), li̯ŋhi̯ "орел" : Gen. li̯mbi̯(ŋ), hoδür "письмо" : Gen. hotürə(ŋ), d́adî(ŋ) "силок" : Gen. d́asi̯nə(ŋ) и т.д.

С другой стороны, налицо определенная зависимость чередования согласных в начале суффиксальных слогов от числа и долготы предшествующих слогов. Это наблюдение М.А.Кастрена было отражено в подразделении именных и глагольных основ на фонетические классы, ср.: "Die erste Declination umfasst alle Nomina, die auf einen langen Vocal ausgehen und diejenigen auf einen kurzen Vocal auslautenden, die aus einer gleichen Zahl von Silben bestehen und eine kurze Penultima haben; zu den zweiten Declination gehören die auf einen kurzen Vocal ausgehenden Nomina, wenn das Wort entweder aus ungleichen Silben oder aus gleichen mit einer langen Penultima besteht oder wenn dem Endvocal m, n, ŋ vorangeht. Nach der dritten Declination werden die auf ein i oder einen Consonanten ausgehenden Nomina decliniert" /I:I56/ (ср. также пояснения и сходную формулировку для глагольных основ - /I:I58,I6I-I62,44I/). Но такая группировка (принадлежащая, возможно, не столько М.А.Кастрену, сколько его издателю А.Шифнеру) оказывается не только громоздкой, но и недостаточно точной: так, глагол h́āgimti- "улучшить", попадающий во второй класс основ, морфонологически ведет себя как основа первого класса (Praet. ńāgimti-d́iəmə, ср. для первого класса homəgimti- "заострить" : Praet. homəgimti-d́îəmə и для второго класса t́īmti- "заквасить" : Praet. t́īmti-śiəmə).Та же неполнота сохраняется в несколько более четких формулировках Г.Н.Прокофьева /2:59/. Это обстоятельство побудило П.Хайду в специальном исследовании, посвященном самодийскому чередованию ступеней /3/, оставить открытым вопрос о возможной зависимости выбора альтернантов суффиксальных морфем от четности/нечетности слога. Н.М.Терещенко в своем наиболее полном на сегодняшний день описании нганасанского языка /4/ иллюстрирует чередования большим и очень ценным материалом, но не ставит задачи вскрыть регулирующие их правила.

2. В то же время уже названная выше работа /3/ содержит ключ к пониманию принципов этих чередований. Анализ нганасанского ударения позволил П.Хайду открыть моросчитающий характер нганасанского языка: слог с долгим гласным или дифтонгом приравнивается к двум слогам с кратким гласным (/3:53/; см. также /5:397-400/ и наше исследование, где роль моры как фонологической единицы установлена также для энецкого языка /6:I3-I5/).

Систематизация имеющихся данных и их проверка путем полевого опроса информантов (Усть-Авам, I986 г.) дают возможность констатировать зависимость появления сильной/слабой ступени согласных и их сочетаний от длины предшествующей части слова в морах:

(R1)перед гласным четной от начала слова моры появляется сильная ступень, перед гласным нечетной моры - слабая ступень.

Сферу действия подсчета мор ограничивают

(R2)непосредственно после согласного (в том числе и исторически утраченного, но присутствующего на глубинно-фонологическом уровне) всегда выступает сильная ступень, если только само возникающее сочетание согласных не участвует в чередовании ступеней;

и (R3) непосредственно после долгого гласного или дифтонга всегда выступает слабая ступень.

При подсчете мор следует иметь в виду некоторые данные исторического вокализма нганасанского языка, а также особенности графики отдельных источников. Так, дифтонги ua и üa, развившиеся в непервых слогах из *ö, трактуются морфонологически как одноморные. Не создает дополнительной моры и дифтонгоидность произношения кратких гласных первого слога, отражаемая - хотя и непоследовательно - в записях М.А.Кастрена (ср. диграфы ea, oa на месте /e/, /o/) и, реже, других исследователей.

2.I. Примеры образования притяжательных форм 3Sg номинатива имен (суффикс -δu/-δü/-δi̯/-δi/-tu/-tü/-ti̯/-ti):

(R1) I мора: ni̯-ti̯ "его жена"; 2 моры: məku-δu "его спина", d́ütü-δü "его рука", h́ini̯-δi̯ "его старший брат", t́īmi-δi "его зуб"; 3 моры: bakunu-tu "его осетр", t́îrimi-ti "его икра", bārbə-tu "его хозяин", kə̄li̯-ti̯ "его слеза", śiəd́ə-ti̯ "его язык", holi̯ʔə-tu "его темя"; 4 моры: kə̄lüʔkü-δü "его короткий", kuəd́umu-δu "ее муж"; 5 мор: t́īpsinəδə-tu "его запястье".

(R2) tər-tu "его шерсть", ŋuət́u "его нога" (от ŋuəj "нога"; jt > t́), d́ebśiti "его пятка" (от ďebŝi(ŋ) "пятка"), hüəgaťü "его колено" (от hüəgaj "колено").

(R3) tā-δu "его олень", kəi-δi "его бок", latə̄-δu "его кость", biriə-δi̯ "сго рана", sūδə̄-δu "его лопатка", ŋəjbukə̄-δu "его шаманская шапка".

2.2. Примеры образования деспричастий с суффиксом -ďa/-ďi/-sa/-ŝa/-si̯/-ŝi (функционально близки инфинитивам):

(R1) 2 моры: bi̯ti̯-ďi "выпить", ďilə-ďi "поднять", ďorə-ďa "плакать", hotə-ďa "написать"; 3 моры: bi̯δi̯bti̯-si̯ "напоить", ďiləbti̯-si̯ "приподнять", ďorələ-sa "заплакать", hoδətə-sa "писать", bəδuatə-sa "расти" (-ua- = I мора!), buatə-si̯ "перешагивать", hīĺə-si̯ "раскрошить"; 4 моры: bi̯δi̯rnänti̯-ďi "хотеть пить", hotərubtu-ďa "заставить написать", buatəʔkə-ďi "начать перешагивать"; 5 мор: hoδətənantu-sa "хотеть писать", hīĺərəbtu-ŝa "обсыпать крошками".

(R2) iŝa "быть" (основа ij-), bəδuasa "вырасти" (основа bəδua(ŋ)-), bəriʔ-ŝi "порвать", bəuʔ-sa"перейти", tasəgim-ŝa "стать горьким".

2.3. Примеры образования причастий неосуществленного действия (суффикс -mətumaʔa/-məδumaʔa и др.), см. /4:270/:

(R1) 2 моры: tūj-məδumaʔa "не пришедший"; 3 моры: bū-mətumaʔa "не уехавший" (в записи Н.М.Терещенко - бӯмәтума"а, однако ср. bū-ďa "уехать"), ńili̯-mətimi̯ʔa "не живший", suə-mətumaʔa "не перекочевавший"; 4 моры: heδi̯ti̯-məδi̯mi̯ʔa "не шедший", hoδətə-məδumaʔa "не учащийся", hutərə-məδumaʔa "не запряженный", ŋəməkə-məδumaʔa "не начавший есть".

2.4. Примеры различий в историко-фонетической рефлексации, связанных с количеством предшествующих вокалических мор:

Самод. /7:I57-I58/ *tepə̂ "гвоздь" > нг. ťəhə, но самод. /7:24-25/ *i̯nəpə "тесть" > нг. ŋinəbə.

Сев.-самод. *putə̂tə̂ "туловище" (ср. ненецк. лесн. pi̯ttat, энецк. тундр. puδoδo) > нг. hütəδə.

Самод. /7:34/ *jə̂kə̂ "близнец"> нг. ďəkə, но эвенк. /8:628/ ńaka "хорошо"> нг. ńāgə "хороший" (в соответствии с R3, после долгого гласного k > g).

3. Рассмотренные выше случаи могут быть охарактеризованы как ритмическое чередование ступеней. С ним сосуществует уже упоминавшееся чередование по правилу

(R4) ритмически сильная ступень чередования подвергается ослаблению в начале исходно закрытого слога.

Показательно частичное несовпадение результатов замены сильной ступени на слабую и ослабления по правилу R4 - в частности, для кластера *nt.

3.I. Примеры образования форм 1Sg намеченного действия (суффикс -hantu-/-hatu-/-handu- и др.):

Сильная ступень, открытый слог: kəməhantumə "собираюсь поймать", ŋanabtāhantumə "собираюсь забыть".

Сильная ступень, закрытый слог: ŋatəhandum "собираюсь ждать", tolarhandum "собираюсь красть".

Слабая ступень, открытый слог: kəhərəhatumə "собираюсь оторвать", māruhatumə "собираюсь укрепить".

Слабая ступень, закрытый слог: mūntəhatum "собираюсь пасти", ŋəmurəhatum "собираюсь есть".

4. Отсчетом мор определяется одновременно и просодическая организация нганасанских слов. Ее ведущий принцип состоит в том, что главное ударение падает на тот слог, который включает предпоследнюю вокалическую мору: sə·mu "шапка", səmu·mə "моя шапка", kümā· "нож", kümā·mə "мой нож", ba·sa "железо", basā· "железный", ho·nsi̯ "иметь", honti̯ə̯· (фонетически также [hontē̜·]) "имеющий", hai̯·mu (фонетически также [ha·jmu]) "меховая обувь", ŝi̯ə·ďə (фонетически также [ŝē·ďə]) "язык".

Этот принцип, установленный в работах П.Хайду, может быть дополнен следующими более частными закономерностями:

- Факультативно имеет место оттяжка ударения на третий от конца слова слог, если предпоследней море соответствует краткий гласный верхнего подъема (особенно i̯) или ə: ńili̯·ti̯~ńi·li̯ti̯ "живет", satə·rə~sa·tərə "песец".

- Слова, состоящие из 4-5 вокалических мор, обычно имеют дополнительное ударение на слоге, который включает четвертую от конца вокалическую мору: bə:jŋübtü·ŝa "обидеть", tu:rkuťa·nə "грузовая нарта", tā:rā· "только олень", məlu̯ə̯·ďa·ʔa "олень с поломанным рогом".

- Слова, состоящие из большого числа вокалических мор - разумеется, таковыми могут быть только многоморфемные образования - обычно имеют дополнительное ударение на первой и/или третьей море основы, вне зависимости от того, попадает ли это ударение на четную или нечетную от конца слова мору: ho:təru(:)btuďu̯ə̯·m "я заставил написать", ho:təru(:)btuďu̯ə̯·mə "я заставил (его) написать". Как правило, акцентуация подобных слов варьируется в зависимости от темпа речи; не исключено, что относительная интенсивность ударения на корневой и суффиксальной частях может использоваться в стилистических целях.

5. Две модели отсчета мор в нганасанских словоформах независимы друг от друга. Чередование ступеней согласных регулируется отсчетом мор от начала слова, а выбор места ударения - отсчетом мор от конца слова. В результате и сильная, и слабая ступени консонантного чередования способны выступать в любой позиции по отношению к месту ударения, ср. satərəkü̯ə̯· "песец-то", koli̯gü̯ə̯· "рыба-то", satərə·ti̯ "его песец", koli̯·δi̯ "его рыба".

В предварительном порядке можно выдвинуть гипотезу о том, что эти две модели связаны с различными хронологическими стадиями в истории нганасанского языка. Система чередования ступеней, как установил уже А.Сотавалта /9/, обладает чрезвычайным сходством с системой, реконструируемой для прибалтийско-финского праязыка (в современных прибалтийско-финских языках она сохранилась реликтово), см. /I0; II: II9-I57/. Вероятно, эти системы могли возникнуть при такой акцентной организации слова, когда основное и дополнительные ударения падали на первый слог (мору) и все нечетные при отсчете от начала слова слоги (моры), т.е. слабая ступень систематически появлялась в предударной позиции. В таком случае закон чередования ступеней можно расценивать как аналог закона Вернера /I2:74-86/.

Однако в нганасанском языке данная система сумела сохраниться и после того, как система ударения модифицировалась (возможно, под иноязычным влиянием), вследствие чего акцентным центром слова вместо первой моры стала пенультима.

Литература

/1/ M.A.Castrén, Grammatik der samojedischen Sprachen, St. Petersburg, 1854.
/2/ Г.Н.Прокофьев, Нганасанский (тавгийский) диалект. - "Языки и письменность народов Севера", ч. I, Москва - Ленинград, 1937, 53-74.
/3/ P.Hajdú, Die Frage des Stufenwechsels in den samojedischen Sprachen. - Ural-Altaische Jahrbücher, Bd. 34, 1962, 41-54.
/4/ Н.М.Терещенко, Нганасанский язык, Ленинград, I979.
/5/ P.Hajdú, Samojedica. - Nyelvtudományi Közlemények, 66. köt., 1964, 397-405.
/6/ E.A.Helimski, Phonological and Morphonological Properties of Quantity in Samoyed. - "Studien zur phonologischen Beschreibung uralischer Sprachen", hrsg. von P. Hajdú und L.Honti, Budapest, 1984, 13-17.
/7/ J.Janhunen, Samojedischer Wortschatz: Gemeinsamojedische Etymologien, Helsinki, 1977.
/8/ Сравнительный словарь тунгусо-маньчжурских языков, т. I, Ленинград, I975.
/9/ A.Sotavalta, Astevaihtelua samojedikielissä. - Suomalaisen Tiedeakatemian Esitelmät ja Pöytäkirjat, 1, 1912, 47-52.
/10/ E.N.Setälä, Über Quantitätswechsel im Finnisch-Ugrischen. - Journal de la Société Finno-Ougrienne XIV:3, 1896.
/11/ Д.В.Бубрих, Историческая фонетика финского-суоми языка, Петрозаводск, I948.
/12/ L.Posti, From Pre-Finnic to Late Proto-Finnic. - Finnisch-Ugrische Forschungen, Bd. XXXI, H. 1-2, 1953, 1-91.

THE PERCEPTUAL CUES OF TONES IN STANDARD CHINESE

MAO-CAN LIN

Institute of Linguistics
Chinese Academy of Social Sciences, Beijing, China

ABSTRACT

The synthesized speech of /shi/ /tuo/ and /ai/ were utilized to investigate the perceptual cues for tones .

The result of this experiment indicated that the four tones can be generated alone by Fo pattern with the possibility of about 95% , whereas the four tones can not be distinguished by amplitude contours alone . It also showed that the effect of duration on the naturalness of tone-3 and tone-4 is greater than that on the rate of identification of tone-3 and tone-2

INTRODUCTION

In 1924 , Liu Fu discovered the important role of Fo in Chinese tone (1). It was found that the Fo curve in syllable not only has a " tone-section " , but also generally has a " onset-curving section " and " end-falling section " (2) . Chuang et al. made the Fo analysis and identification test for colloquial Standard Chinese (3) .

Howie demonstrated the primacy of Fo pattern in the identification of the four tones (4) . Wang talked about the role of Fo and amplitude in the four tones (5) . Lin and Wang discovered that the judgement of tone category of the first syllable in bisyllabic word is often influenced by pitch of the second syllable and duration of the first one (6) .

This experiment tried to investigate the role of Fo , amplitude and duration in the four tones by varying these parameters in the synthesized speech .

THE PHYSICAL MANIFESTATION OF TONES

We made an acoustical analysis of 138 monosyllables consisted of 38 different Initial and Final Combinations with tones spoken by two speakers (m and f) .

Fig. 1(m) and 1(f) were the Fo pattern of the two speakers . It can be seen from fig. 1 that each tone generally has its own peculiar Fo pattern .

Although the durations of the four tones did not show a regular relative relation , comparatively speaking , the duration of tone-3 were in most of the cases the longest .

Four different types of amplitude contour could roughly be drawn from the amplitude curves in 276 monosyllables, namely: mid-hump , back-hump , two-hump and front-hump . It can be seen that the amplitude contours in tone-3 spoken by m were all two-hump , but those spoken by f were two-hump only in 60% of the cases .

The peak of intensity in tone-3 showed in most of the cases the lowest .

PERCEPTUAL EXPERIMENT OF TONES

The syllables of /shi/ , /tuo/ and /ai/ were synthesized by a synthetic system (7) under five conditions shown in the left column of tables given below . All the speech sounds were randomized to make it impossible for the 14 subjects (as listeners) to predicate under which condition the speech sound were synthesized while he or she heard it . The average rate of identification of tone by subjects (14) was displayed in percentage in the right column of each table . The figures in parentheses in the tables represented the percentage of the speech sounds in good timbre judged by the subjects .

The data of the parameters in condition one roughly corresponded to the physical manifeslation of tones . A sonagram of /tuo/ synthesized by condition one was displayed in fig. 2 . Table 1 showed that the rate of correct identification of tone was 98.8% , and the speech sounds in good timbre amounted to 70.7% .

In condition two , the amplitude contours were only varied , e.g. , the amplitude contour of low-falling-rising of Fo varied to mid-hump from two-hump , but other parameters were the same as those in condition one . A sonagram of /tuo/ in this condition was displayed in fig. 3 . Table 2 showed that the rate of correct identification of tone was 97.6% , and the speech sounds in good timbre amounted to 67.1% .

In condition three , Fo patterns were all mid-level , and durations were the same as those in condition one , but the amplitude contours had the four different types of mid-hump , back-hump , two-hump and front-hump . A sonagram of /tuo/ in this condition was disnlayedin fig. 4. Table 3 showed that the subjects (14) identified the speech soundsas tone-1 about 90% . No one identified them as other tones , namely , no one identified the speech sounds with amnlitude contours of two-hump or front-hump and with mid-level of Fo as tone-3 or tone-4 . This result indicated that the four tones can not be distinguished by amnlitude contours alone .

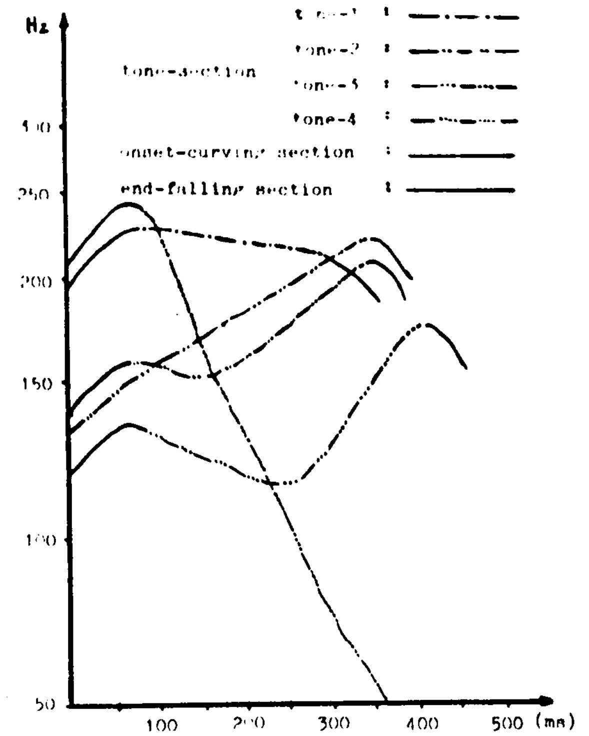


Fig. 1(m) average fundamental frequency curves in mono-syllables for Beijing speaker m

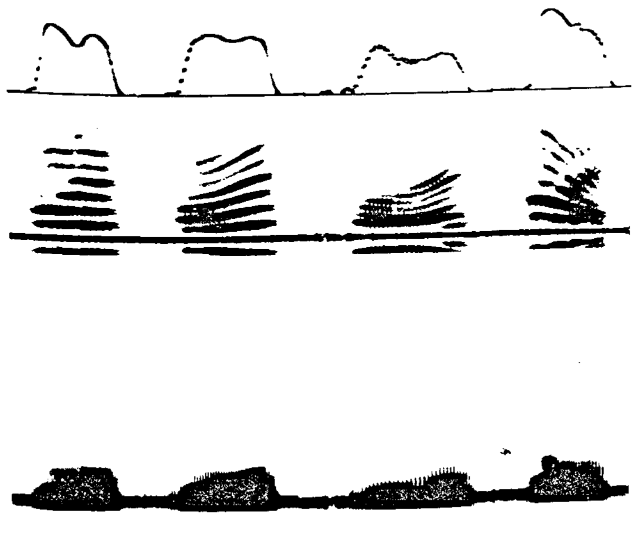


Fig. 2 sonagrams of /tuo/ synthesized in accordance with condition 1

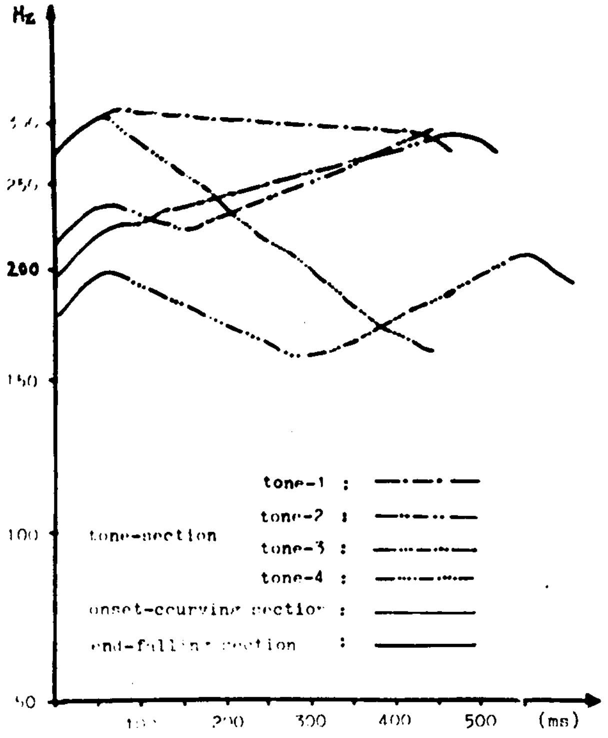


Fig. 1(f) average fundamental frequency curves in mono-sullables for Beijing speaker f

TABLE 1

Condition one	The rate of identificationof tones			
	tone-1	tone-2	tone-3	tone-4
1.1 Fo: high-level Amp.: front-hump back-hump T: 348ms	97.6 (75.8)			
1.2 Fo: mid-rising Amp.: back-hump front-hump T: 390ms		100 (76.2)		
1.3 Fo: low-falling-rising Amp.: two-hump T: 470ms	2.4 (2.4)		97.6 (61.9)	
1.4 Fo: high-falling Amp.: front-hump mid-hump T: 307ms				100 (69.0)

While the Fo patterns and the amplitude contours in condition four and five remained the same as those in condition one , the durations in condition four and five were different from those in condition one . In condition four , the durations of four different sounds were regulated as the same as those of tone-4 in condition one ; In condition five , the durations of four different sounds were done as same as those of tone-3 in condition one . The speech sounds synthesized by condition four were correctly identified as tone-1 , tone-2 and tone-4 95.8% in average , but as tone--3 90.5% , namely , the rate of correct identification of tone-3 decreased about 7% compared with that in condition one . This time , the number of speech sounds with tone-3 judged to be in good timbre decreased 22% from those in condition one.

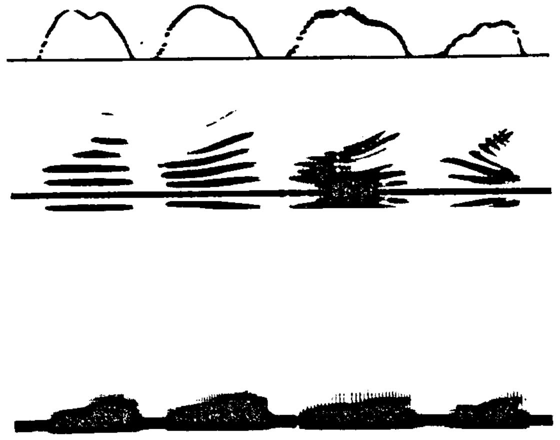


Fig. 3 sonagrams of /tuo/ synthesized in accordance with condition 2

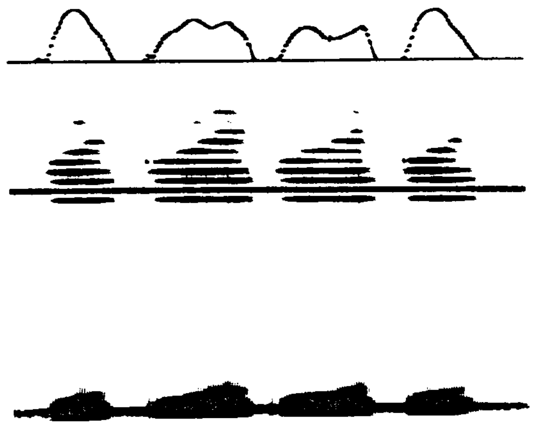


Fig. 4 sonagrams of /tuo/ synthesized in accordance with condition 3

And the speech sounds in condition five were correctly identified as tone-1 , tone--3 and tone-4 95.8% , but as tone-2 88.1% , namely , the rate of correct identification of tone-2 decreased 12% compared with those in condition one . This time , the number of speech sounds with tone-4 judged to be in good timbre decreased 19% from those in condition one . These two results indicated that the effect of duration on the naturalness of tone-3 and tone-4 was greater than that on the rate of identification of tone--3 and tone-2 .

CONCLUSION

We may conclude that the four tones can be generated by Fo pattern alone with the possibility of about 95% ; The effect of duration on the naturalness of tone-3 and tone-4 is greater than that on the rate of identification of tone-3 and tone-2 ; The four tones can not be distinguished by amplitude contour alone .

TABLE 2

Condition two	The rate of identification of tones (%)			
	tone-1	tone-2	tone-3	tone-4
2.1 F_0: high-level Amp.: mid-hump two-hump T: 382ms	100 (92.7)			
2.2 F_0: mid-rising Amp.: front-hump T: 412ms		97.6 (64.3)		
2.3 F_0: low-falling-rising Amp.: mid-hump T: 468ms			97.6 (57.1)	
2.4 F_0: high-falling Amp.: back-hump T: 370ms				95.2 (54.7)

TABLE 3

Condition three	The rate of identification of tones (%)			
	tone-1	tone-2	tone-3	tone-4
3.1 F_0: mid-level Amp.: mid-hump T: 332ms	92.2 (38.1)			
3.2 F_0: mid-level Amp.: back-hump two-hump T: 413ms	85.7 (42.7)			
3.3 F_0: mid-level Amp.: two-hump T: 443ms	85.7 (26.2)			
3.4 F_0: mid-level Amp.: front-hump T: 310ms	92.9 (57.1)			

TABLE 4

Condition four	The rate of identification of tones (%)			
	tone-1	tone-2	tone-3	tone-4
4.1 F_0: high-level Amp.: front-hump back-hump two-hump T: 310ms	100 (78.5)			
4.2 F_0: mid-rising Amp.: back-hump front-hump T: 310ms		95.2 (66.6)		
4.3 F_0: low-falling-rising Amp.: two-hump T: 310ms		4.7 (2.4)	90.5 (39.7)	
4.4 F_0: high-falling Amp.: front-hump mid-hump T: 310ms				97.6 (73.8)

TABLE 5

Conditin five	The rate of identification of tones (%)			
	tone-1	tone-2	tone-3	tone-4
5.1 F_0: high-level Amp.: front-hump back-hump two-hump T: 467ms	100 (83.3)			
5.2 F_0: mid-rising Amp.: front-hump back-hump T: 467ms	7.1 (7.1)	88.1 (66.7)		
5.3 F_0: low-falling-rising Amp.: two-hump T: 467ms			95.2 (69.0)	
5.4 F_0: high-falling Amp.: front-hump mid-hump T: 467ms				100 (50.0)

REFERENCES

(1) Liu Fu , An experimental record of Chinese tone , 1924 , Shanghai Qunyi book--store .

(2) Lin Mao-can , The pitch indicator and the pitch characteristics of tones in Standard Chinese , Acta Acustica , Vol. 2 No. 1 , 1965 .

(3) Chuang , c.k. , Hiki , s. , Sone , T. & Nimura , T. , The acoustical features and perceptual cues of the four tones of colloquial Standard Chinese , In Proceedings of 7th Inter . Cong. of Acoustics , Vol. 3 1972 .

(4) Howie , j.m. , Some experiment on the perception of Mandarin tones , In proceedings of the 7th Inter. Cong. of Phonetic Sciences , eds. by A. Rigault & R. Charbommeau , 1972 . The Hague : Mouton . PP. 900-904 .

(5) Wang , William S-Y. , Lectures on experimental Phonetics , in Collective Work of Linguistics , No. 14 , ed. by Beijing University , 1983 . Commercial Press .

(6) Lin Tao and Wang , William S-Y., The Problem of tonal perception , The Journal of Linguistic Society of China , No. 2 , 1984 .

(7) Yang , Shuen-an , The effect of the dynamic characterictis of voiced source upon the quality of systhesized speech , ZHONGGÓ YÜWEN , No. 3 , 1986 .

CHINESE TONES AND THEIR DURATION

M.K. RUMYANTSEV

COLLEGE OF ASIAN AND AFRICAN STUDIES, MOSCOW UNIVERSITY, USSR

ABSTRACT

In strictly phonological sense the parameter of duration can be regarded as redundant in so far as the system of Chinese tones is concerned. However, when their constitutive function is taken into account duration as well as intensity prove to be indispensable for producing "natural" Chinese speech.

In isolating languages tones play an important role: their primary - phonological or distinctive - function is to distinguish one morpheme from another. But it is not their only function. Various combinations and modifications of tones constitute the prosodic system of an isolating language as a whole. As a result of their interaction different rhythmic structures of words are created, different emotional and evaluative overtones are superimposed on information proper.

The main physical correlate of a tone is its fundamental frequency. Tones differ as to their register, pitch direction (contour - level, rising, falling, falling-rising) and the respective intervals. For a considerable period of time these parameters of tones have been in the focus of attention of experimental phonetics. The other physical correlates of tones - intensity and duration - have not yet received all the attention they deserve.

A considerable amount of data obtained during speech synthesis experiments [1] has convinced us that the parameter of duration cannot be ignored if our aim is "natural" Chinese speech. This feature is redundant only on the distinctive level: morphemes can be distinguished even if their duration is equal, provided, of course, that their registers and contours are different. Thus the simplest opposition is observed in the case of the first and the fourth Chinese tones: a level tone on a high note is opposed to a falling tone which starts quite high and then falls as low as possible. From the point of view of their constitutive function duration and intensity prove to be absolutely indispensable for making tones sound 'natural', really Chinese.

When we stress the importance of duration we do not mean to say that it is only the total duration of each tone that really matters. The parameter in question is highly relevant for determining the inner structure of complex tones. Thus the third (circumflex) tone is not only longer than any other basic tone, but is also characterized by a certain time relationship between its components: its rising part cannot exceed a certain limit, otherwise the listener may mistake it for the second (rising) tone.

The duration of tones as integral prosodic units of syllables has a functional load in words and through the latter in all the other prosodic layers of the language.

Our synthesis of two-syllable words has shown that the role of duration for producing natural (proper) Chinese sound cauls cannot be overestimated. There are 19 models of Chinese words and each model has its own time relationship of the constituting tones. If this or that time relationship is not observed the sound caul becomes unacceptable from the point of view of the language norms.

Under the influence of the higher prosodic levels tones may vary in length as well as in range, pitch direction and intervals. Thus different degrees of prominence - sentence stress, logical stress etc - can affect the duration of tones, but the time relationship typical of the respective words should not be distorted. Otherwise the listener may fail to identify the word as such.

Experimenting with syllables of different duration in speech synthesis enables us to solve at least two problems: 1) finding the optimal rhythmic models in each of the above mentioned 19 groups of words, 2) determining tolerance zones for each of these models.

The rhythmic function of duration is best seen in those models which are represented by combinations of tones of the same type. Thus, for instance, in the model constituted by a sequence of two first tones the second syllable is either longer than the first or at least is of the same length. The normative time relationship is equal to 1.3. The equal length is apparently the threshold because when the second syllable is shorter than the first the model is rejected as false, unnatural. Synthesis has so far not revealed the predominance threshold, beyond which the realisations of words are perceived as exaggerated or unacceptable. In the model of two second tones the second syllable is also longer than the first, but in contrast with the above mentioned model the equal length is not tolerated. At the same time the above limit of exceeding the length of the first syllable can be established somewhere in the region of 1.76. It should be noted, however, that although this realization of the model appears to be fairly acceptable, some auditors characterized the second syllable as unusually long.

The model of two fourth tones gives a completely different picture: the first syllable here is longer than the second. The variation zone is rather wide, ranging from 1.04 to 2.11.

If we try to correlate the so far discussed time relationships with the parameter of interval we shall come to the conclusion that the situation in different models is different. Thus in the model of two second tones we observe a direct proportion (the wider the interval the longer the syllable), whereas in the model of two fourth tones the reverse proportion is true: a wider interval is correlated with a shorter syllable.

As a rule, the duration relationship of tones in isolation is preserved when different tones are used in the modelled words. For instance, the third tone being originally the longest will remain to be so when it is part of this or that word. To what extent it will be longer than the other tones depends, however, on the concrete rhythmic model used. In the model of I+3 the duration of the second syllable (the third tone) exceeds the duration of the first syllable considerably (the proportion is I.67). There can be no question in this case of making the syllables equally long or reversing the proportion. That would be absolutely unacceptable. In the model 2+3 the original duration relationship of the tones is also preserved: the second syllable (the third tone) is longer than the first. To reverse the duration relationship would be out of the question, but the variation zone is rather wide (I.I - I.7). It remains to be seen whether making the syllables equal would be rejected by auditors or not.

If the third tone is used before the first, the second or the fourth it will be longer than the fourth but shorter than the first or the second. In the normative synthesized realization of the word kǎoyā the proportion is I.2 or it can be even increased. The first tone cannot be, however, shorter than the third tone. The synthesized word kǒuyīn is the case in point, the proportion between the second syllable and the first is I.03.

When the third tone is combined with the fourth, the former should be longer than the latter and the proportion varies from I.4I to I.66.

The fourth tone before the first, the second or the third tone is always shorter than any of them. In the model 4+I the first tone is longer than the fourth and the duration proportion is I.4I. It can be slightly increased, but when in the synthesized word - like unit xiàyī the proportion turned out to be I.93, this realization was rejected by the auditors. The longer duration of the first tone there went beyond the accepted norm. It should be pointed out that as far as the parameters of register and interval are concerned the constituent tones were within the norm and the reaction of the auditors cannot be accounted for by these parameters. The lessening of the duration of the first tone to the point when it becomes equal to the fourth tone or when the duration relationship shifts undermines the rhythmic characteristics of words. For example, in one of the programmes of the word àixī the first tone in the final of the syllable xī was shorter than the fourth tone in the syllable ài, which immediately caused a negative reaction from the auditor.

No less important is the time relationship in various models of the type "basic tone + neutral tone". Even the best programmes in so far as the parameters of register and interval were concerned were often rejected by the auditors because of some wrong duration proportions. The optimal proportion in the model "the first tone + the neutral tone" requires that the neutral tone should be twice as short or even shorter than the first tone. The proportions of I.36-I.38 formed the threshold. The time proportion was markedly improved in the realizations with the duration relationship equal to I.45.

The tolerance zone of the relationship between the second tone and the neutral one is about 2.II. The I.56 proportion was rejected. As far as the relationship between the fourth and the neutral tones is concerned, 2.51 appears to be within the norm. The I.75 proportion was rejected by the auditor, who insisted on the longer fourth tone.

Wrong time proportions interfere with the production of normative rhythmic characteristics in words even if the register and interval relations are correct. Nor can the correct time proportions alone, without the normative register and interval proportions, ensure good rhythmic characteristics, which are produced by the sum total of features. The close interconnection of duration and interval proportions and their combined effect are borne out by their combined interpretation by the auditors and the undifferentiated perception of the duration, register and interval parameters of words. Some realizations of the words tóufa and xífu were interpreted in precisely this way by the auditors. The rhythmic parameters of the synthesized word tóufa were deemed unsatisfactory. The auditor described the neutral tone in the beginning of the syllable fa as too high (tóu, 139-166, fa, 146-134) and suggested that the tone of the syllable tóu be prolonged, even though the time relationship in this case was quite normative: 2.24 (tóu - 415, fa - 185). The desired prolongation of the second tone in the first syllable was probably automatically associated in the auditor's mind with an increase of the interval, which would really improve the interval proportion between the end of the second tone and the beginning of the neutral tone. Merely to prolong the second tone without increasing its rising interval is not enough to improve the rhythmic parameters of the word. In one of the realizations of the word xífu the rhythmic parameters were also deemed unsatisfactory. The frequency interval between the end of the second tone in the syllable xí and the beginning of the neutral tone proved too small (I.08 against the normative I.23). The modelled time proportions (I.54 against the normative 2.II) were also unsatisfactory. The auditor insisted on lowering the neutral tone and on increasing the interval in the first syllable. The auditor failed to notice certain duration disproportions in the given word sample and sought to improve the rhythmic parameters by correcting only the register and interval proportions: both the lower register of the neutral tone and the increased rising interval of the tone in the syllable xí aim at one and the same thing, i.e., at increasing the interval between the end of the second tone and the beginning of the neutral tone.

Tones act in words as prosodic factors forming morphemes and words. Duration, as one of their constituents, is functionally important in words: time proportions in a word cannot be broken without distorting its prosodic make-up.

The most intimate time mechanisms of tone are manifest in the fine spectra of speech signals, responsible for their different quality. In different tones the finales of Chinese syllables are known to be perceived by ear as slightly differing in quality. In different tones and finales these distinctions are not the same but they are indisputably functionally important to the Chinese ear in the sense of the national specificity of sounds. In any case it is important to determine what spectral parameters account for this specificity.

Analysis of natural tones and their synthesis elucidate primarily the role of the frequency and amplitude parameters of the spectrum. For instance, in the natural realizations of syllables in our material a rise of fundamental frequency of the rising tone causes a progressive shift of the first formant. With a male speaker the shift proceeded as follows: at the beginning of the finale í of Tone 2 the first formant was 250 Hz, in the middle it became 300 Hz and at the end it was 350 Hz. With a female speaker the shift was even more pronounced: 350 Hz, 400 Hz and 500 Hz.

Analysis of the synthesized syllables with the finale í recognized by the auditors as "natural", that is undistinguishable from the natural sounding shows that their naturalness is accounted for precisely by this fine correction (correlation) between the fundamental frequency and amplitude values and different formants and by the coordinated function of all the spectral components. The measure and concrete proportions of that coordination are not universal and depend on the linguistic system, their main purpose being to ensure the normative quality of sounding. It is not by chnace that the attainment of this goal also has to do with making the synthesized signal natural or close to it. The impression of naturalness is produced by the absence of monotony (machine-like quality) in the spectrum of the synthesized vowel which, just like it is in natural speech, is not uniform, as far as its quality is concerned, at different segments of the sounding and evolves from the beginning to the middle and the end. For example, F_1, which is coordinated in frequency and amplitude with F_0 in keeping with the rules of the system, is represented by a set of coordinated values within the formant itself and among them rather than by one and the same value throughout the signal. In coordinating the values at every given segment the synthesized signal is in fact ascribed frequency and amplitude "microvariations". These variations are not universal or determined by the human organs of speech but systemic and linguistic, that is characteristic of a given phonetic norm. In our case we get the syllabic tone with all its inherent systemic characteristics. The latter are determined in the spectral structure not only by the corresponding frequency and amplitude coordination but also by time coordination: frequency and amplitude dynamics of the spectrum unfolds in its portions of time, which correspond to different segments of the sounding tone from its beginning to end.

The role of the coordination of frequency, amplitude and time in the spectrum of Chinese finales of different tones is well illustrated by the synthesized tones which were rejected as unsatisfactory.

Tone, as a phonological unit which distinguishes syllabic morpheme, in its model variant is a set of features functioning in unison: if within a certain period fundamental frequency values form an even contour, the envelope of amplitude values forms the same contour. The rise (fall) of fundamental frequency is accompanied by corresponding changes in amplitude values. Inadequate coordination of parameters often results in the inadequate synthesis of syllabic tone. However, the proportion of this correlation may differ, depending on the linguistic system. For example, the auditors rejected the realizations of the sharply falling Chinese (fourth) tone, whose programmes envisaged falling frequency intervals equal to 1.62 and 1.51. The intervals were not only within the norm but the optimal ones in fact. The amplitude values in principle changed in the same direction and, nevertheless, the tones were characterised by the auditors as "passive, inert" and the interval seemed to be inadequate (!). Consequently, the synthesized signals failed to reproduce the amplitude, frequency and time relationships that were worked out by the given linguistic system. The fall in the amplitude values at every given segment of the tone failed to fit the norm prescribed by the fall of the fundamental frequency values or to be synchronized with the time segments of the realization, during which these spectral changes took place.

When separate tone in which the fundamental frequency and amplitude parameters changed in different directions were given to the auditors for identification they were often confused or totally rejected as unacceptable within the given orthoepic norm. This is not to say that lack of coordination is always a defect. On the contrary, it is in many cases a normal phenomenon in connected speech. It is explained by the fact that the parameters coordinated in the units of speech pronounced separately or in the strong position are assigned different roles in connected speech. Thus at the level of syllable fundamental frequency always differentiates lexical meanings in the Chinese language system, i.e., acts as different tones, whereas the amplitude and times values of formants provide for other prosodic distinctions, such as the word's rhythm and intonation contrasts. It is necessary to learn to model this lack of coordination in simulated speech in order to get the needed sounding at every given point of speech continuum. This calls for great efforts on the part of linguists because the measure of this uncoordination, too, is being worked out within the language systems.

1. Chinese tones were synthesized in the laboratory of experimental phonetics at the College of Asian and African Studies (Moscow University) with the help of a formant synthesizer (SPPI-75).

RHYTHMIC STRUCTURE OF DISYLLABLES IN YORUBA

M.I. KAPLUN.

COLLEGE OF ASIAN AND AFRICAN STUDIES, MOSCOW UNIVERSITY, USSR

ABSTRACT

Studies of the tonal-rhythmic structure of Yoruba words (on the basis of disyllables) make it possible to suggest a general pattern of the rhythmic arrangement of tones and also to formulate the basic rules governing tones in this unit of speech.

The method of synthesis used in linguistic studies makes it possible to approach many linguistic problems at a qualitatively new level. In coping with these problems researchers nevertheless encounter considerable difficulties in view of the fact that, to meet the requirements of modelling speech, it is necessary to know the so called phonetic characteristics of speech which bring simulated speech closer to the linguistic prototype modelled in every particular case, alongside the features traditionally referred to as functional (phonological). Experiments in synthesis show, for example, that when modelling the syllabic tone (as well as the complex tonal-rhythmic mechanism of words), it is necessary to take into account all the multilevel functionally distinctive features.

In natural speech tones are known to be adjusted to each other and for this reason disyllables, the minimal lexical unit in which the basic tonal-rhythmic laws regulated by the rules of sandhi are manifest, have been chosen as the linguistic material in studying the tonal-rhythmic structure of Yoruba. In view of the fact that latest works on tones in Yoruba usually single out three tones, namely, medium, low and high (designated M for medium tone, H for high tone and L for low tone), nine tonal-rhythmic models of disyllables have been chosen for investigation: M+M, M+H, M+L, L+L, L+M, L+H, H+H, H+M, H+L.

In studying the tonal-rhythmic models of great interest are those phonation segments in which tones are joined together because it is precisely there that they are "adjusted" to each other, coordinating in a certain way the contour, register, time and amplitude characteristics. The junctures of tones manifest most graphically the parameters that organise the tonal-rhythmic model as a certain semantic unit of the given linguistic system. It should be stressed that for the tonal-rhythmic structure to be modelled successfully it is necessary to take into consideration not only the contour and register parameters but to equal measure amplitude, time and interval parameters which, interacting with each other, alone can ensure that the given simulated tonal-rhythmic model is fully or at least partially associated by native speakers with its natural prototype.

That is why we can hardly agree with J. M. Hombert[1] who concludes from his analysis of the perception of tonal-rhythmic structures in natural speech on the basis of bisyllabic nouns in Yoruba that acoustic parameters are informative in different degrees. He asserts that the native speakers of Yoruba use only two main acoustic characteristics which enable them to distinguish rhythmic structures in six tonal-rhythmic models. The first characteristic, according to Hombert, is connected primarily with F_0 in the vowel of the second syllable (V_2). The second characteristic is presented by him as a combination of three parameters -- the interval of the modification of F_0 in V_2, the medium value of F_0 in V_2 and the frequency breakage between the end of V_1 and the beginning of V_2. As is seen, Hombert includes only frequency characteristics in the set of functional features, treating amplitude and time characteristics as non-functional.

It is apparently indisputable that the native speakers of Yoruba are capable of distinguishing one tone from another in the final position by the features mentioned by Hombert. However, they are inadequate for the synthesis of rhythmic models normative from the point of view of prosody. This is borne out by the analysis of simulated M+L and M+H tonal-rhythmic models, which strictly reproduced all the rhythmically important contour, interval and time parameters and deliberately distorted only the amplitude parameters. After listening to the sounding of these models, the auditors observed that they could hardly be considered as normative.

Analysis of tonal-rhythmic models with the medium tone (M+M, M+L, M+H) shows that it is most stable rhythmically in all the combinations with other tones and retains the even contour, its own duration and medium register (with respect to other tones within the three-level register scale). The Yoruba register scale for male voices close to baritone tenor can be conventionally divided into three ranges:

low	medium	high
90 + 115 Hz	120 + 160 Hz	165 + 200 Hz

The specificity of the M+M tonal-rhythmic model consists in the fact that its contour is formed by two even tones, with the second syllable tone invariably beginning at the frequency of the end of the first syllable. The absence of a frequency interval between the syllables in the medium tone model is important from the point of view of rhythm, while any deliberately made interval between the syllable tones results in the broken rhythm of the model. Depending on the nature of the frequency breakage between the syllables, the auditors described the sounding they heard as a combination of the medium and high tones or the medium and low tones. For the M+M rhythmic model to sound naturally and be perceived unambiguously, its both syllables should be actualised in the medium range exactly. When the register was deliberately changed (with all the other acoustic parameters of the combination of these tones remaining unchanged) the auditors identified them, for example, as a combination of two high tones. The determination of the time relationship between the syllables, one of which is characterised as a medium tone, is a key functional acoustic feature in simulating these tonal-rhythmic models. The presence of the medium tone in disyllables suggests a certain strategy of synthesis, i.e., the creation of an exact time relationship between the medium tone, on the one hand, and the high and the low tone, on the other. Our experimental material included cases of the synthesised medium tone being identified as high. Analysis of this fact brought to our attention primarily the time relationship between the syllables of the disyllable, in which the first syllable (medium tone) is equal in its duration to the second (low tone), the time relationship between syllables characteristic of the combination of the high and the low tone. Consequently, when modelling the rhythm of a disyllable with the medium tone its duration should be longer than that of the syllables characterised by other tones.

One of the key rhythmic characteristics of bisyllabic models with the medium tone in the first syllable is the correlation of the frequency of the end (medium tone) and of the beginning (low or rising high tone), with its indispensable prerequisite being the absence of any frequency contrast at the juncture of tones. Several variants of (M+L and M+H) disyllables with different frequency contrasts and without these at the juncture of syllables have been synthesised to verify this hypothesis. The auditors identified only those of the synthesised (M+L and M+H) rhythmic models which had no frequency rupture between the syllables and also those with the minimal frequency rupture (not longer than a second). The rest of the synthesised models were identified incorrectly. The second-long interval of the frequency rupture between the syllables can, apparently, be considered admissible, whereas a longer one places these rhythmic models outsides the normative rhythm of disyllables.

The peculiarity of the M+H and M+L tonal-rhythmic models consists in the fact that they are organised by two types of equivalent tonal contours, every one of them worked out in the Yoruba language. Different acoustic characteristics become functionally important in their rhythmic organisation. All the programmes of synthesising the M+H and M+L bisyllabic models were recognised as normative when their tonal contour was either formed of two even tones belonging to different register levels or consisted of a combination of the even and the rising contour. Each type of the M+H and M+L models has its own rhythmic peculiarities. In the first variant the functionally important factor is the register contrast between syllables equal to a minor third. The cases of the frequency relationship between the registers exceeding that interval were described by the auditors as "rhythmically pronounced", "exaggerated", or too "robot-like", whereas those with a smaller interval were perceived as a combination of two medium tones. Apparently, an interval of a minor third can be considered as a sort of a crucial point, beyond which these models disintegrate.

Another important characteristic of the M+H rhythmic model is growing amplitude in the second syllable marked by a high tone with an even contour. Analysis of this type of sounding showed peak amplitude invariably at the second syllable.

A sizeable increase in amplitude in the second syllable (compared to the first one) apparently complements the even contour and a similar correlation of these acoustic parameters will produce the high tone effect in the given tonal-rhythmical model for the native speakers of Yoruba. The natural question arises about the functional importance of each of these parameters for the high tone to be perceived unambiguously. With this aim in view several models were synthesised, with only one parameter changing in every one of them and the rest remaining intact. For example, in one case the first and the second syllable of a disyllable had equal amplitudes, in another the amplitude of the second syllable was, on the contrary, increased but there was no register contrast between the medium and the high tone and in still another the first syllable had a bigger amplitude compared to the second syllable. The sounding was repeatedly recorded and offered for auditing at random. The results of the auditing analysis show that none of the acoustic parameters can be singled out as a factor determining the rhythm of a given model. It is rather the combination of these features that accounts for the certain stability of the rhythmic model in any contextual conditions.

Another type of tonal-rhythmic models (M+H and M+L) is formed by an even (in the first syllable) and rising M+H or falling M+L (in the second syllable) tonal contour. The factor important from the point of view of rhythm in this type of model is the rising or falling intervals, which begin at the frequency level of the end of the medium tone. In all the rhythmically normative models of this type the interval between the medium, the high and the low tone was within the range from a minor third to a fourth. It should be pointed out that the register frequency rupture in the first variant of the M+H and M+L models (_ - and - _) the rising and the falling tonal contour in the second syllable of the second variant of the same models (-/ and -\ ) has one and the same set of intervals, which should be not less than a minor third and not more than a fourth because the sounding with an interval less than a minor third does not give stable identification results, whereas that with an interval exceeding a fourth is described as "rhythmically accentuated", for this reason a strictly prescribed interval between the syllables of a bisyllabic structure can be viewed as a key rhythmic characteristic of this type of M+H and M+L models.

Analysis of the M+H and M+L tonal-rhythmic models makes it possible to see that the very term "high tone" and "low tone" accepted in Yoruba tradition and unambiguously defining their register does not exactly correspond to the real acoustic nature of these tones or at best corresponds only to one of the possible variants. This supposition is borne out by the results of the analysis and synthesis of the rhythm of bisyllabic structures, in which the high and the low tone was in the first syllable. The H+L and L+L tonal-rhythmic models, like any other with the high and the low tone in the first syllable, can have a double tonal contour, namely, rising and even (high tone) or falling and even (low tone), which is explained by the complex acoustic mechanism of these tones, which presupposes in each case a certain combination and correlation of register, time, interval and amplitude values.

In one of the programmes of the H+H tonal-rhythmic model recognised by the auditors as "unsuccessful" the tonal contour was formed by two rising tones. The auditors' judgment was, presumably, influenced by the fact that the rising contour of the high tone in both syllables of the disyllable was sounded in the high register, whereas the rhythmic arrangement of the high tone requires either the medium register (rising contour) or the high one (even contour). Each of these parameters in combination with the necessary set of other parameters is rhythmically important and functional in producing a normative sounding of one of the high tone variants. The synthesis of realisations with pronounced register and contour characterstics leads to a rhythmic disharmony, which violates the rhythmic stereotypes worked out in the Yoruba linguistic system and traditional among native speakers.

The low tone in the final position is of special interest in the H+L, M+L and L+L tonal-rhythmic models. To say that the low tone in the final position always has a falling contour will in no way be enough to cover all the peculiarities of that tone in the final position nor to show its effect on the rhythmic organisation of the entire disyllable, as the frequency interval of that tone and the speed of its formation remain outside the scope of research. Meanwhile, as is seen from the synthesis, these acoustic characteristics are functionally important not only for the rhythm of the model itself but also for contrasting other disyllables with the low tone in the final position. Orientation only to the falling tonal contour of the low tone in the final position is justified when this tone is contrasted with a different one, say, medium. Nevertheless, this criterion no longer works when two rhythmic models — H+L and M+L — with the low tone in the final position are set against each other. In this situation the interval and the speed of its formation rather than the falling nature of the low tone (it remains the same) become rhythmically significant, alongside other acoustic features involved in the differentiation of these models.

Studies of the tonal-rhythmic structure of Yoruba words (analysis and synthesis on the basis of disyllables) make it possible to suggest a general pattern of the rhythmic arrangement of tones and to formulate the main rules of sandhi governing tones in this unit of speech.

The interaction of tones in disyllables is based on three major rules. Two of them — the rule of register and register-contour oppositions -- operated in all the rhythmic models and their possible variants, while the third rule regulates the rhythm within only those models that fall under the rule of register-contour oppositions.

The specificity of the rhythmic organisation of bisyllabic Yoruba words consists in the fact that in one and the same model tones interact differently, depending on the context, for which reason one of the variants of the model can fall under the rule of register-contour oppositions, while another variant under that of register oppositions. For example, a variant of the "M+H" model with an even tonal contour in both syllables of a disyllable sounded in different registers — medium and high (_ -) — is governed by the rule of register oppositions, while its other variant falls under the rule of register-contour oppositions because the medium tone has an even contour in the medium register and the tone going after it has a rising contour (-/).

The rule of register oppositions regulates the interaction of tones in those tonal-rhythmic models which have tones with similar even tonal contours produced in different frequency bands, i.e., in different registers, which in their turn account for the opposition of tones in disyllables.

The rule of register-contour oppositions regulates the tonal-rhythmic models in which the tones combined have opposite contours and registers.

The third rule can conventionally be formulated as the rule of frequency correspondence or the equi-frequency correspondence of tones, which covers and regulates all the register-contour changes in the tones at all the segments of disyllables. The meaning of the main requirement of this rule by definition boils down to the equi-frequency correspondence between the beginning of the consequent tone in a disyllable and the end of the preceding one. Sometimes this equi-frequency correspondence can take place in the so called frequency correspondence zone, which allows for an insignificant frequency breakage (that is not perceived as a register contrast in Yoruba).

The phonological principle underlying the rule of the frequency correspondence of tones elucidates both the general mechanism of the interaction of tones in rhythmic models covered by the rule of register-contour oppositions and any particular manifestation of this general regularity. This ensures the necessary stability of tones and prevents their confusion in any contextual situation.

[I] J.M. Hombert. Perception of Bisyllabic Nouns in Yoruba. Studies in African Linguistics, 1976, Sup. 6.

RHYTHMIC ACCENT PATTERNS IN ALEUT

A.S.ASINOVSKY, Y.V.GOLOVKO

Institute for Linguistics
Leningrad, USSR, 199053

ABSTRACT

Aleut, being a language without a word-stress, forms its rhythmic structure by means of two main things: rhythmic accent and long vowels. There are several main rhythmic accent patterns in Aleut, long vowels "breaking ranks" - then, rhythmic accent is counted from that "break".

INTRODUCTION

Suprasegmental features of three now existing Aleut dialects have not been yet investigated in detail. It is obvious,however; that such information could give new material for typological studies, as well as for comparative Eskimo-Aleut linguistics (cf. the comparison of Eskimo rhythmic accent patterns in [1]).

The work of this kind on Eskimo material was begun some time ago[2]. The purpose of this report is to present rhythmic accent patterns of wordforms in Bering Island Aleut , which is, probably, a conservative form of Atka Island Aleut[3]. The material was obtained during two field trips to Bering Island in 1982 and 1985.

Long Vowels

The three short Aleut vowels /a/,/i/,/u/ have the correlative long vowels. Long vowels can be either included into morpheme signifiers or result from phonomorphological alternations: siching[s'íč:iŋ'] "four", sichiing [s'ič'i:ŋ'] "nine"; aalukux̂ [a:ł°uk°uχ°] "he laughs"; aaluusakux̂ [a:ł°u:sak°uχ°] "he laughs at sb.". The initial /a:/ in the last wordform is a part of the root-morpheme, the second long vowel /u:/ is a result of lengthening /u/ before the transitivising suffix -sa-. This is one of the suffixes provoking obligatory lengthening of the final stem-vowel. A long vowel in such a position can be treated as two vowels,with a morpheme-borderline between them.

Some word-combinations can also provide conditions for vowel-lengthening; for instance, some ordinal numerals do so: aalax hisix̂ [a:łaxis'iχ]"second", cf. alax [áł:ax] "two"; qaankus hisix̂ [qa:nk°usís'iχ] "third", cf. qankus[qánk°us°] "three".

The Correlation of the Morphemic- and Syllabic Structures

Phonomorphological rules of the syllabic structure of the wofdform depend, to a large extent, on the phonetic structure of the root-morpheme. The most typical root structures are CVCV and VCV - two-syllable roots with a final vowel: awa-l[áβ'ał] "work", asu-x̂[ás'uχ°] "pot", chachi-l[č'ač:ił]"cover". One-syllable and two-syllable roots are also possible: qa-x̂[qaχ]"fish", sasuli-l [sasúł'ił] "be annoyed". Roots of more than three syllables are rare, and none-syllabic roots is not found in Aleut.

Phonomorphological rules in Aleut are determined by two main principles: 1) all suffixes (numbering about 120) can only begin with a consonant; 2) the way of linking the stem is determined by the phonomorphological type of a suffix. We shall illustrate here three phonomorphological types of suffixes.

1. Suffixes linking the stem through a long epenthetic vowel, e.g. -ĝuta- "again" - imat-ii-ĝuta-ku-x̂ [imati:γutak°uχ°]"he is shouting again". Linking a vowel-stem, these suffixes lengthen the final vowel: adalu-u-ĝuta-ku-x̂ [aðału:γutak°uχ]"he is telling lies again".

2. Suffixes linking the stem through the epenthesis of normal length, e.g. -da "the imperative" -hum-i-da [humíða]"inflate!", -ĝi- "the objective resultative" - chx̂uuĝ-a-ĝi-ku-x̂ [čχu:γaγik°uχ°] "is washed". When linking a vowel-stem, these suffixes do not change the length of the final vowel, e.g. chachi-da "cover!"[č'ač'iða].

3. The suffixes whose way of linking depends on the final sound of the stem: in case of a vowel-stem the final vowel becomes long; the consonant-stem links these suffixes through the epenthesis of normal length. One of these suffixes is "the transitivisor" -sa-. A peculiar feature of the epenthesis of normal length is that its quality is not constant. It may be supposed that, in many cases, the choice between different variants of epenthetic vowels is influenced by the vowel structure of the wordform, e.g. chag-u-sa-ku-x̂[čáγ°usák°uχ°] "he is splitting sth. with sth."; ag-u-sa-ku-x̂ [áγusák°uχ°] "he is passing with"; chaqug-a-sa-ku-x̂[č'aq°úγasák°uχ°] "he is chewing sth."; iklug-a-sa-ku-x̂ "he has bumped against"[ikłúγasák°uχ] . The first two wordforms include the epenthesis /u/, the second two wordforms - /a/. This distribution is not obligatory but rather preferable, i.e. it is a tendency, not a rule. The choice between /u/ and /a/ depends on the preceding vowel: /u/ follows /a/, and /u/ is followed by /a/. This dissimilation according to the height of the raised part of the tongue) is typical of some suffixes which link the stem through the epenthesis of normal length.

Accent and Rhythmic Structure

Long vowels play an important role in the formation of the rhythmic structure of a wordform. Their distribution,however, does not enable us to treat them as correlating to a word or phrasal stress. The rhythmic structure of a wordform is formed not only due to long vowels but also according to a rule of distribution of accented and neutral (not accented) syllables.

In wordforms of CVCVC-structure (if they do not include long vowels), the first syllable is generally accented,e.g. tux̂ix̂ [t°úχiχ] "dot", hatix [hát:iχ] "lips", chalix̂ [č'ał:iχ] "fishline". Accented/neutral syllables do not coincide with the opposition long/short vowels: the vowel of an accented syllable is shorter than the corresponding long vowel.

A characteristic feature of accent is its influence on the quantity of the consonant following the accented syllable. If a consonant follows a long vowel in a wordform of similar structure, it does not change its quantity, e.g. taachix̂ [ta:č'iχ] "elbow bone", hachix̂ [háč:iχ]"back". The first word shows an intervocalic /č/ of normal length, the second one demonstrates the corresponding long consonant. This phenomenon can be seen most clearly if the consonant following the accented syllable is an obstruent or sonant.

In three-syllable isolated wordforms the second syllable is accented, e.g. hyutikux̂[çutík°uχ°] "he is pouring (water)", kidunax̂ [kiðúnaχ] "he helped sb.", samisis [samís'is'] "numeral". The consonant following the accented syllable is also lengthened but this is not so obvious as in two-syllable wordforms and is certain only for sonants. Consonant clusters appearing in none-syllabic positions do not influence the rhythmic structure of the wordform.

Four-syllable wordforms (with no long vowels) have two accented syllables - the first and the third, the third one being marked more distinctly, e.g. hax̂satikux̂ [háχsatík°uχ°]"he is getting ready", awazunax̂[áβ'az°ún:aχ] "he worked well". The accent on these syllables cause lengthening of the following sonant (at least, it is surely so in the position after the third syllable).

In multi-syllable wordforms a rule of rhythmic accent puts an accent on every second syllable, except in the cases when a long vowel appears and breaks the rhythmic structure; then, rhythmic accent is counted from that "rhythmic break".

The distribution of accented/neutral syllables given above is of probability character. Accent is closely connected with the syntactic context of wordforms, or , rather, with their syllabic structures , e.g. hlang haqakux̂ [łáŋaqák°uχ°]"my son is coming up". The one-syllable word hlang and three-syllable word haqakux give a four-syllable stretch(syntagm), which is accented as a four-syllable wordform. The rhythmic structures of the words do not contradict to the rhythmic structure of the stretch, and, so, they are not changed. When two "notional" ("independent") words form a three-syllable stretch, their own rhythmic structures inevitably contradict to the rhythmic structure of the stretch, e.g. hlang snukuq "I have sent my son"[łáŋsn°ukúq]. A three-syllable structure tends to have the second syllable accented - but not in this case (the first and third syllables are accented). It can be explained by a rule of obligatory accent of the only syllable in the first, "notional", word . Be - cause of this, the three-syllable stretch is accented as a multi-syllable structure (every second syllable). If a three-syllable stretch is formed of a "dependent " ("not notional") word and a "notional"word, the first syllable of the stretch is not accented, e.g. wan suda[β'ans°úða] "take this one!"; cf. qax̂ suda[qáχs'uða] "take the fish! In these examples the "pointing word" wan does not prevent the speaker from putting the "right" accent, and the "notional"word qax is "a starting point" for the rhythmic accent structure of the second stretch.

Let us take the multi-syllable stretch chiĝanax̂ qatukux̂[č'iγón:aχqat°úk°uχ°]"the river is rich with fish". The first wordform has a "right" accent on the second syllable. In the second wordform, our informants put an accent either to the first or to the second syllable. It depends on the type of pronunciation: the "full" type requires the second syllable to be accented, i.e. the rhythmic structures of both wordforms are preserved in the stretch. However, the second wordform can be uttered in a "reduced-type" pronunciation - then, the accent is put to the first syllable of the second wordform. It is important that if in the wordform the second syllable is accented, it produces a quasi-homonym, cf. qaatukux̂ "he wants to eat". Long and accented vowels differ enough the two words notto be mixed, but speakers often put an accent to the second syllable "to be on the safe side".

The rhythmic structure of a stretch is connected with the vocalic structures of the wordforms. It can be illustrated by wordforms with the epenthesis of normal length. For instance, there exist two variants of the imperative from the verbs with consonant-stems, e.g. aĝ-a-da "give!" and aĝ-da - same meaning. The choice between them depends on the rhythmic structure of syntactic context, e.g. ngus qax aĝ-ma [illegible]"give me the fish!", cf. qax aĝ-a-da [illegible]"give the fish!" In the first stretch the accent is put to the third syllable, and the epenthesis is deleted. In the second stretch the accent is put to the epenthesis, and it can, by no means. be deleted.

The rhythmic structure of a wordform is generally preserved in two cases: 1) if it does not contradict to the rhythmic structure of the syntactic context; 2) if it belongs to the first "notional" word of the stretch. The study of rhythmic accent structures of wordforms in different syntactic contexts shows that there are no prosodic means in Aleut which provide conditions for the wordform as an independent unit.

REFERENCES

1. Woodbury A.C. Eskimo and Aleut Languages. - In: Handbook of North American Indians, vol. 5. Washington: Smithsonian Institution, 1984, p. 49-63.

2. Yupik Eskimo Prosodic System: Descriptive and Comparative Studies./ Ed. by M. Krauss. Fairbanks: Univ. of Alaska, Alaska Native Languages Center, 1985.

3. Asinovskiy A.S., Vakhtin N.B., Golovko E.V. Etnolingvisticheskoye opisaniye komandorskikh aleutov (An ethnolinguistic description of the Aleuts of Commander Islands. - Voprosy yazykoznaniya, 1983, No.6, p. 108-116.

THE CHANGE FROM APICAL TO DORSAL R IN NORWEGIAN

ARNE KJELL FOLDVIK

Department of Linguistics
University of Trondheim
N-7055 Dragvoll, Norway

ABSTRACT

There has been a rapid spread of dorsal pronunciation of r in South-West Norway during this century, affecting more than 1/10 of the population of the country. The dynamics of the spread are described and reasons for it discussed.

In Norwegian the most common pronunciatjon of r is an alveolar tap, [ɾ]. In the Oslo area it is often palatalized, [ɾ̡]. A century ago r was pronounced as an apical trill, [r], in most parts of Norway. This pronunciation is now common only in a small area between the towns of Florø and Ålesund on the west coast. The apical trill in this area is a velarized one, [r̴], and the alveolar tap in the surrounding areas is also velarized, [ɾ̴]. In the three northernmost counties, Nordland, Troms and Finnmark an alveolar fricative or approximant is often used, [ɹ]. In South-West Norway where the dorsal pronunciation is common, a variety of pronunciations of r may be heard; from a palatal, velar or uvular fricative or approximant, [j, ɰ, ɣ, ʁ], to a uvular trill, [ʀ].

Information from the Norwegian Dialect Survey at the University of Oslo forms the basis of Map 1. It shows the towns of South-West Norway where a dorsal r pronunciation is common and areas where informants born about the turn of the century use the dorsal pronunciation.

Map 2 is based on several different sources of information. In the first instance it was based on information, some of it extremely detailed, that the Directors of Education in 81 towns and municipalities in South-West Norway supplied in 1978. [1] Secondly, it is based on information from colleagues, students, and local informants who have offered information about their own area after radio programs about the spread of dorsal r.

When the two maps are compared it becomes clear there has been a spread of dorsal r to big areas in the South-West, but no towns have been affected. Even if the striation of map 2 also covers fjords and thinly populated mountain areas, so that the spread may look greater than it actually has been, the area taken over by dorsal r has a population in excess of 400000, which is more than 1/10 of the population of Norway. This spread is the biggest change in the pronunciation in Norwegian during the last decades.

The dorsal r continues spreading quickly in some areas, notably round about the towns of Bergen and Florø, more slowly in other areas, for instance inland from the towns of Kristiansand and Arendal, and seems to have come to a halt near the town of Risor, an 'apical' town of the South-East coast.

The spread of dorsal r can not be accounted for purely by the motorically simpler movement that is needed to produce it. With the latitude in dorsal pronunciation it is not surprising that speech therapist reports from the schools in dorsal areas hardly ever show r-problems, while in the apical areas r-problems are very common indeed.

The spreading of dorsal r in Norwegian is facilitated by the fact that the dorsal and apical pronunciations are equally socially acceptable and are both used on Norwegian radio and TV, but also by the fact that most people do not experience the change in pronunciation as a change of dialect. On the whole Norwegians are dialect proud; pupils are by law encouraged to speak dialect and dialect is used by pop-artists as well as politicians.

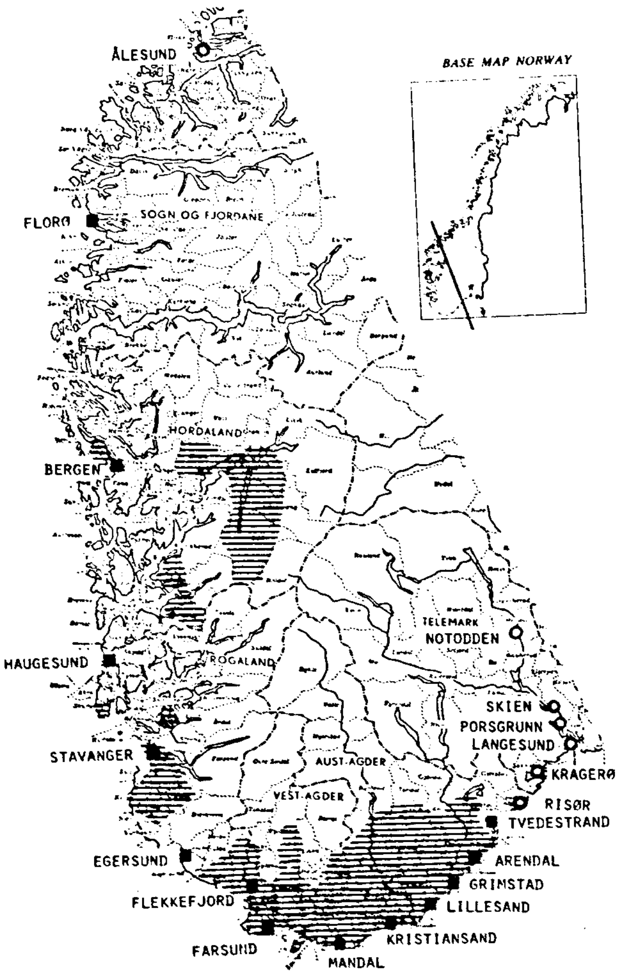
But the main reason for the spread of dorsal r would seem to be the prestige connected with dorsal towns and bigger settlements in the area and the linguistic influence that these centres excert on the rural districts. School centralization which leads to children often travelling long distances by bus to go to bigger schools more often than not situated in a town or bigger settlement, facilitates the spreading.

Even if the spread is fast in several areas at the moment there is reason to believe that it will not continue at its present rate. It seems likely that it will only continue as far as the linguistic influence of dorsal towns and settlements reaches. There are no signs of the dorsal r spreading to any of the apical towns. The dorsal r pronunciation in the capital of Oslo seems to be restricted to some upper class speakers only.

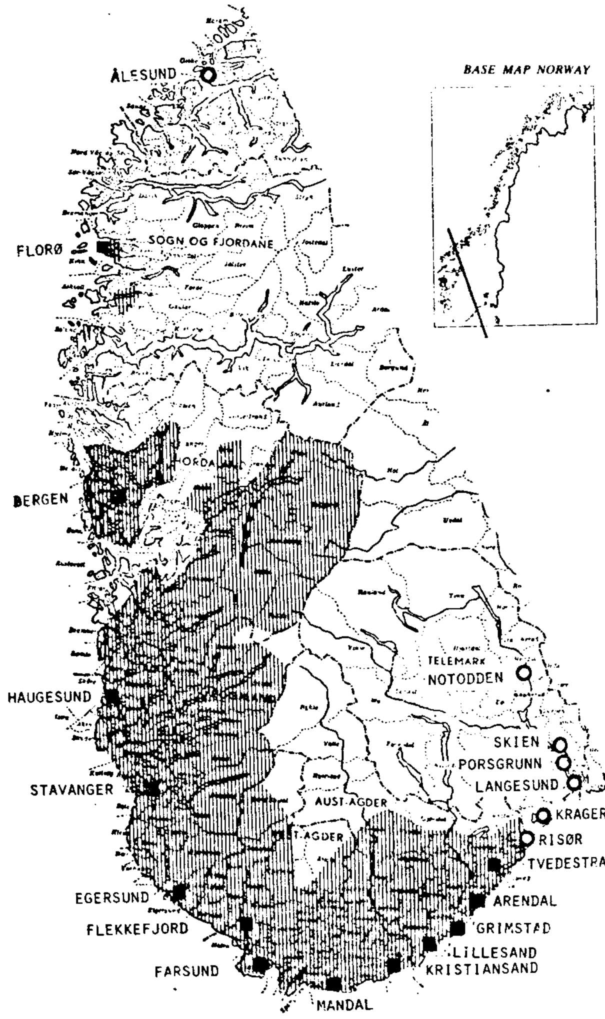
This change to dorsal pronunciation is easy to register; far easier than for instance minute changes in vowel pronunciation. Fieldwork and data collection concerning the spread of dorsal r will be continued.

Reference

[1] A.K. Foldvik, The pronunciation of r in Norwegian with special reference to the spread of dorsal r. Pp. 105-110. In (ed.) T. Pettersson: Papers from the 5th Scandinavian Conference of Linguistics. Acta Univ. Lundensis 30. Lund, 1979.



Map 1. Horizontal striation covers areas where informants born about 1900 use a dorsal r.
■ Towns with dorsal pronunciation of r.
○ Towns with apical pronunciation of r. From Foldvik (1979)
——— County boundary.
------ Municipal boundary.



Map 2. Vertical striation covers areas where informants born about 1960 use a dorsal r.
■ Towns with dorsal pronunciation of r.
○ Towns with apical pronunciation of r.
-·-·-·- County boundary.
------ Municipal boundary.

SOCIOPHONETIC ASPECTS OF DUTCH CLEFT-PALATE SPEECH

ALDA VAN ERP

Dept. of Applied Mathematics and Signal Processing,
Dr. Neher Laboratories Netherlands PTT,
Leidschendam, the Netherlands

ABSTRACT

The results of a rating experiment are presented and related to 18 phonetic variables obtained from a panel of phonetic experts. Recorded, non-emotional speech fragments spoken by ten adult male Dutch cleft-palate speakers were investigated. The verbal content was controlled in all speech fragments. In the rating experiment the cleft-palate speech fragments were rated by two groups of listeners (of whom 30 had had a training in speech therapy and 30 had not) on 19 speech scales and 15 scales pertaining to social status, social attractiveness and competence.
The results show which variables cause listener group effects. Additionally, the relation between speech ratings and personality/social ratings are displayed and compared for the two listeners' groups. Moreover, the listeners' ratings are related to the (expert) phonetic variables.

INTRODUCTION

Cleft-palate speech is the speech produced by someone who has (had) a cleft palate. It is pathological in the sense that it sounds obviously deviant from speech that falls within the range that is accepted as 'normal' by the speakers in a particular speech community. It deviates from 'normal' on a number of vocal aspects pertaining to articulation, phonation, resonance and prosody.

Vocal aspects may be used by listeners to infer information about characteristics of the speaker. For example, when you talk to someone whom you have just met for the first time, it may happen that you get a first impression of the other which is based on how the other speaks rather than what the other says. Although it may happen that such inferences are not in keeping with the reality, they play an important role in (first) impression formation.

It is not clear to what extent typical cleft-palate vocal characteristics contribute negatively to the impression people get of the speaker in question. The present paper attempts to answer this question. To this end perceptual descriptions of the relevant vocal characteristics were related to inferences about speaker characteristics that are based only on vocal information.

The process of attributing speaker characteristics from only vocal aspects appears to include not only conclusions about psychological and social aspects, but also about physical aspects of the speaker's identity such as his or her sex, age, height, weight, physique and state of health [2]. The study reported on here does not deal with inferences about physical aspects of the speakers. It deals with personality and social aspects on the basis of vocal aspects. Moreover, the personality and social judgments in this study are obtained from listeners only; they are not obtained by means of tests.

METHOD

At the base of the research reported on here are running speech fragments from a sample of Dutch cleft-palate speakers. Firstly, the vocal aspects of these fragments have been phonetically described. This was done analytically by a panel of experts. Secondly, the same vocal aspects were judged in an associative fashion. This was done in a rating experiment by various relatively large groups of listeners. Thirdly, the speech fragments were used for obtaining associative, inferential judgments about the speakers' personality and social aspects. This was done in a rating experiment as well, by the same listeners that also rated the vocal aspects in an associative fashion. Thus, the cleft-palate speech samples were described on three levels. Analogous to the lens model [1], these three levels of description can be referred to as 'distal', 'proximal', and 'attributional'. On the distal level is the phonetic description of the vocal aspects; on the proximal level the associative description of the vocal aspects; and on the attributional level the associative description of the personality and social aspects. The rating experiment was designed in such a way that the proximal and the attributional levels were distinguished.

Speech material

The data base consisted of recorded prose passage renderings by ten different cleft-palate speakers. The prose passage was an emotionally neutral reading text, and yielded more than one minute of running speech per speaker, the verbal content being controlled. The speakers were male, had slight South-Eastern Dutch accents, and were aged between 17 and 48 years.

The phonetic description

A combined approach was followed: Both a segmental and a nonsegmental description were made of the speech material (see above) by four experts. In addition, 12 linguistically trained judges were used for obtaining intelligibility scores. These were based on nonsense sentences read aloud by the ten cleft-palate speakers.

The segmental description indicated which phonemes were pronounced deviantly and how often each of the following typical errors were made: (1) fronting of the place of articulation,(2) backing of the place of articulation,(3) glottal stop, (4) nasal emission,(5) nasal explosion, and (6) denasality.

The nonsegmental description consisted of ratings on 33 vocal parameter scales. By means of scalar degrees it could be indicated for each parameter whether the deviation from a predefined neutral point was either 0, 1, 2, or 3. The ratings are a mix of quality and quantity. This means that if a particular vocal effect is very strong when it is present it would be rated as 3 if it occurred relatively often. However, if it occurred relatively rarely it would receive a lower score. The scales pertain to: (1) supralaryngeal features (concerning the lips, jaw, tongue tip, tongue body, the velopharyngeal mechanism, the pharynx, as well as supralaryngeal tension and precision of articulation), (2) laryngeal features (concerning phonation type and laryngeal tension), and (3) prosodic features (concerning pitch, loudness, and temporal structure). The intelligibility scores were expressed in percentages of syllables that were reported correctly, averaged over the 12 judges.

The associative description

For the associative description, the speech material that was phonetically described was presented to 60 female listeners. Their mean age was 22 years, ranging from 20 to 26 years. 30 listeners were students from the college of speech therapy training in Nijmegen ('TRAINED'). The other 30 were students enrolled in the Faculty of Arts of the University of Nijmegen, but not in language courses ('UNTRAINED'). The listeners were born and raised in the South-Eastern part of the Netherlands and were therefore accustomed to a South-Eastern Dutch accent. The rating experiment took place in a language laboratory with individual booths. The listeners were presented with the recorded speech material via headphones. They did not know any of the speakers nor did they know what the speakers looked like. The listeners were asked to rate each bi-polar (7-point)scale on the rating sheets they got in front of them. The meanings of the scale positions was explained to them in the written instructions. The instructions also encouraged them to give their first impressions. In fact, they were only given approximately 3 seconds per scale to respond. The rating scales were divided into two categories. One category contained scales pertaining to vocal aspects ('speech scales'), the other contained scales'), the other contained scales pertaining to personality and social aspects. The two types of scales were rated in separate sessions.

Twelve speech scales refer to more or less general vocal aspects that pertain to articulation, phonation type and prosody **(viz. standard, precise, intelligble, good reading performance; bright, creaky; high-pitched, varied, expressive, loud, quick, smooth).** Seven speech scales refer to pathological vocal aspects **(viz. nasal, with a blocked-up nose, snorting, snoring, glottalised, hoarse, lisping).** In addition, there was one question with a dichotomous response category, namely: "Do you think speech therapy is required? yes/no".

The personality/social scales refer not so much to the classic Evaluation, Potency and Activity dimensions, but rather to dimensions that were considered to be more relevant in connection with the social acceptability of people with a speech defect. Therefore, they refer to social status, (social) competence and social attractiveness **(viz. of high social status, highly ambitous, with qualities of leadership, self-confident, reliable, intelligent, suited for public speaking, strong-willed, careful, interested, friendly, warm-hearted, spontaneous, cheerful, modest).**

RESULTS AND DISCUSSION

Firstly, I will discuss the outcomes of t tests based on the associative ratings. This is done, separately for the speech ratings and the personality/social ratings, to determine whether there is any difference between the ratings of the two groups of listeners (viz. TRAINED versus UNTRAINED). Subsequently, the relations between associative speech ratings and personality/social ratings will be discussed and compared for the two groups of listenerss. Finally, the relations between the analytic speech description (esp. of the pathological aspects) will be related to the associative ratings. Again, the groups of listeners will be compared.

Before the associative ratings were subjected to t tests, interrater reliabilities (Cronbach's alpha) were computed, separately for each scale, and separately for the two groups of listeners. For the speech scales, these coefficients generally appeared to be satisfactorily high (>.80) and comparable for the two groups of listeners. The only conspicuous differences between the two listener groups occurred for a few pathological speech scales. Firstly for **not sniffing- sniffing, not snoring-snoring, not hoarse-hoarse, and not creaky-creaky,** where the reliabilities of the UNTRAINED listeners were comparatively low, ranging from .44 for **not sniffing-sniffing** to .74 for **not creaky-creaky** while the reliabilities of the TRAINED listeners were higher than .90. These differences were due to the fact that the TRAINED listeners indicated differences between the various speech fragments more clearly than the UNTRAINED listeners. And secondly there was a conspicuous difference for **not with a blocked-up nose-with a blocked-up nose,** where the reliability for the TRAINED listeners was lower than for the UNTRAINED listeners was lower than for the UNTRAINED listeners (viz. 71 versus .91). This was mainly due to the fact that the TRAINED listeners did not seem to agree much among themeselves with respect to the scale positions they assigned to individual speech fragments. As for the personality/social scales, there appeared to be no differences between the two listeners groups that are worth mentioning. Twelve scales were rated very reliably > .91). Three scales (viz. **unfriendly-friendly, conceited-modest,** and **unreliable-reliable)** were rated less reliably, with coefficients ranging from .69 to .80. In addition, it appeared that in general the personality/social ratings were less extreme than the speech ratings. This could mean that the listeners are rather careful in attributing personality/social characteristics to speakers on the basis of only their speech.

In subsequent analyses use was made of the mean of the ratings of 30 listeners on each of the 19 speech scales and 15 personality/social scales respectively, for each of the ten cleft-palate speech fragments.

The t test results revealed that statistically the ratings of the two groups of listeners (averaged over 30 listeners and 10 speakers) were equally high. Therefore, the conclusion is that there is no effect for groups of listeners, neither for the speech ratings nor for the personality/social ratings. With respect to the speech ratings this could mean that also in a normal social context the cleft-palate speech aspects are just as salient for laymen as for speech therapists. With respect to the personality/-social ratings this means that apparently listeners are rather consistent in the attribution of personality/social characteristics on the basis of someone's speech only.

In order to determine whether there are any relations between associative ratings of vocal aspects and associative ratings of personality/-social aspects, product-moment correlations were computed. Firstly this was done, separately for the two groups of listeners, for general ratings (i.e. ratings that are not only averaged over listeners but also over scales). This general correlation was .80 for the TRAINED listeners and .88 for the UNTRAINED listeners.

In order to be able to investigate the relation between associative speech ratings and associative personality/social ratings in more detail, correlations were subsequently computed between every speech scale and every personality/social scale, separately for the two groups of listeners. From the results it appeared that, for both groups of listeners, the more general (i.e. not-pathological) speech ratings correlate significantly with personality/social ratings far more often than do pathological speech ratings. Speakers that were judged to speak little **varied, expressive, precise, smooth, clear, loud,** and who were judged to have a **bad reading performance** were judged less positively on most personality/- social scales than were speaker that were judged to speak **varied, expressive, precise** etcetera. It should be noted, however, that this finding may be an artefact of the speech material that, was used (viz. a reading text). As for the pathological speech ratings, there were some remarkable differences between the two groups of listeners. For the UNTRAINED listeners there were significant correlations between ratings of a number of pathological speech aspects (viz. **nasal, snorting, snoring)** and judgments about certain personality/social aspects (viz. **unsuited for public speaking, without qualities of leadership, modest, wavering).** For the TRAINED listeners the judged presence of pathological speech aspects did not correlate significantly with personality/social judgments. This difference between the two groups of listeners is important insofar that it seems to indicate that a layman is more inclined to attribute certain less positive personality/-social characteristics to speakers with obvious cleft-palate speech defects than a speech therapist.

Before the analytic description of the vocal aspects were related to the associative ratings, statistical analyses were carried out - for the nonsegmental description and the intelligibility scores - in order to make sure that the ratings were reliable, and that the various parameters were relevant to the aim of the study. The reliability of the means of the scores was assessed by means of Cronbach's alpha.

It appeared that for the nonsegmental parameters this coefficient ranged from .06 for protrusion of the lower jaw to .88 for speech rate.

Only values higher than .75 were considered to be satisfactorily high. Consequently, only ten out of the 33 nonsegmental parameters were considered to have been reliably rated, namely: nasality, nasal emission, precision of articulation, whisperiness, creakiness, pitch mean, pitch range, loudness mean, interruptedness, and rate.

The parameters that were rated most severely, averaged over the ten speakers, were nasality, nasal emission and **whisperiness.** In order to assess whether the ratings on these scales varied as a function of the speakers, the ratings on each of these scales were subjected to separate analyses of variance with two fixed factors namely 'speakers' and 'raters' (level of significance= 5%). It appeared that the factor 'speakers' was significant for all ten scales. Inspection of the mean ratings (N=4) for each of these parameters revealed that this was not caused by just one or two speakers who had received extreme ratings while the other speakers were rated neutral.

For the average **intelligibility** scores (N=12) the reliability coefficient, averaged over ten speakers, was .85. The scores for the individual speakers ranged from 80% to 98%. However, for nine out of ten speakers the range was between 92% and 98%. This points to a ceiling effect. Additionally, in order to examine how the 18 different analytic variables (i.e. 7 segmental, 10 nonsegmental, and 1 intelligibility variable) were related, product-moment correlations were computed. There were nine significant correlations and in only five cases the correlation was so high that more than half of the variance in one variable was accounted for by another variable. This concerns the following variables: (nonsegmental) nasality and (nonsegmental) nasal emission (r= .71), (segmental) nasal emission and (nonsegmental) nasal emission (r= .79), (segmental) nasal explosion and (nonsegmental) loudness mean (r= .71), (segmental) glottal stops and intelligibility (r= -.87), and precision of articulation and pitch range (r= .71). Because neither of these correlations are extremely high, it was decided to relate all 18 phonetic variables to the associative ratings.

To determine the relations between the phonetic variables and the associative ratings, product-moment correlations were computed, separately for the two groups of listeners.

In the first place, this was done between the phonetic and the associative descriptions of the vocal aspects. Correlations between the following parameters were significant, for either one or both of the groups of listers. Correlations higher than |.63| are significant. The height of the correlations indicates the validity of the associative ratings of the speech aspects. The first correlation is of the UNTRAINED listeners; the second of the TRAINED listeners.

1. (Segmental) fronting of place of articulation with **lisping** (.84, .81)
2. (Segmental) nasal explosion with **snoring** (.48, .66)
3. (Nonsegmental) nasality with **nasal** (.92, .83)
4. (Segmental) nasal emission with **snorting** (.40, .81)
5. (Nonsegmental) nasal emission with **snorting** (.64, .90)
6. Intelligibility with **intelligible** (.54, .69)
7. (Nonsegmental) whisperiness with **hoarse**- (.88, .90)
8. (Nonsegmental) precise articulation with **precise** (.86, .83)
9. (Nonsegmental) rate with **quick** (.82, .80)

Apparently, for **snoring, snorting,** and **intelligible** the associative ratings of the TRAINED listeners are clearly more valid than those of the UNTRAINED listeners. For **lisping, nasal, hoarse,** and **precise** there is practically no difference.

In the second place, this was done between the phonetic description of the vocal aspects and the associative description of personality and social aspects. Correlations between the following parameters were significant for either one or both of the listener groups. Their height indicates the strength of the relationship between 'true' vocal characteristics (i.e. vocal characteristics as described analytically by experts who were trained to do the job) and inferred personality/-social characteristics. Again the first correlation is of the UNTRAINED listeners; the second is of the TRAINED listeners.

1. Nasality with **suited for public speaking** (-.61, -.66)
2. Whisperiness with **intelligent** (-.71, -.68)
3. Whisperiness with **of high social status** (-.66, -.69)
4. Whisperiness with **with qualities of leadership** (-.62, -.66)
5. Whisperiness with **suited for public speaking** (-.62, -.64)
6. Rate with **spontaneous** (.56, .64)
7. Rate with modest (-.71, -.76)

Apparently, these correlations are much the same for the two listener groups. In addition, it appeared from the results that precise articulation and pitch range correlate significantly with every personality and social scale except for **suited for public speaking, self-confident,** and **modest.** Their correlations ranged from .68 (e.g. for intelligent) to .91 (for careful) and were comparable for both groups of listeners.

CONCLUSION

From the results it appeared that there are obvious relations between some pathological vocal characteristics (viz. nasality and whisperiness) and negative ratings on some personality/social characteristics pertaining to social competence. These relations were more or less equally strong for TRAINED and UNTRAINED listeners. However, it also appeared that for more general (i.e. not-pathological) vocal characteristics (viz. rate, precision of articulation,- and pitch range) the relations with rating on personality and social characteristics are more pervasive. Moreover, the results strongly suggest that these five vocal characteristics mediated in the attribution process. Admittedly, whether these vocal characteristics are actually used for attributing personality and social characteristics of the speakers in a normal social context would have to be investigated in a more realistic setting.

ACKNOWLEDGEMENT

This research was financially supported by the research pool of the Catholic University of Nijmegen. In addition, partial financial support was given by the British Council. Financial support for attending this congress is given by the Netherlands PTT.

REFERENCES

[1] Brunswik, E. **Perception and the representative design of psychological experiments.** Berkeley and Los Angeles, California: University of California Press, 1956.

[2] Laver, J. & Trudgill, P. Phonetic and linguistic markers in speech. In: Scherer, K. & Giles, H. (eds.), **Social markers in speech.** Cambridge: Cambridge University Press, 1979.

The Flood of Japanised Foreign Words and These Futurity

Dr. William Akira Sakow, LL.D.

International Society of Phonetic Sciences

I want to discuss the flood of 'Japanised' Foreign Words which have appeared in the numerous advertising, social and educational fields, and to criticise from the point of Phonological and cultural viewpoint. Because, the making use of these strange words are very important from cultural standpoints, but, on the other hands, some serious flaws strikes my minds.

Explanation of Japanese pitch accent:-
/hato/ = ハト = dove, /hane/ = ハネ = feather, /kokoro/ = ココロ = mind. (2) Phonetic symbols, /š/ = /ʃ/, /ž/ = /ʒ/, /č/ = /tʃ/, /ǰ/ = /dʒ/, / / = phonetic notation.

For the last twenty or thirty years, and as an indirect result of the Second World War, we, even language teachers, often have been surprised at the large number of so called Japanised Foreign Words to be found in the field of advertising, and in many shops' names. Flood and utilization of foreign words into public and private informational fields, seems to me, as wonderful introducing knowlege of western civilization, given to the Japanese, but on the other hands, the thought to introduce things without critical consideration, is an abnormal thoughtlessness perhaps coming from the white-phobia of the Japanese during Meiji Restoration era. That is to say, most of Japanese advertisers may suppose without inquiry of the matter, that advertisement with so many foreign words (both original spelling and Japanised loanwords as well), may be interested and praised by many guests, whether the foreign vocabularies are understood or not by them. How hasty deed the Japanese do!

According to my these long belief, I'll select some hundreds of familiar Japanised loanwords from about 36,000, and try to estimate their future usage and longevity.

At present, the Japanised loanwords are said to be so many words including every fields, but as far as I studied the borrowed words, from educational point of view, I think about 10,000 will be enough, excluding personal names, geographical names and the names of novels and arts' works. About a half of the loanwords are English, in origin from Great Britain and U.S.A., and other minors are Chinese, French, Portuguese, Dutches, Russians, Itarians, Spanishes, Germans, Koreans, Sanscrits and Latins, etc.

1. The Japanised loanwords before World War II.

I was born in a countryside of Gifu Prefecture in 1912, but I REMEMBER VIVIDLY, even at present, that we, country boys, were using, in those days, twenty or more borrowed Japanese words without thinking of them as "loanwords." For example, my mother used to say to me, 'šappo to manto o wasure naide, Akira!' = 'Don't foget to wear chapeau and manteau, Akira!'

Other loanwords I can remember from my younger days were:- /šatsu/ - shirt, /botan/ - botão (P), /čju:bu/ - tube, /taija/ - tire, tyre, /meriken-ko/ - American flour, /pan/ - pão (P), /ko:çi:/ - coffee, /horumarin/ - Formalin (G), /aruko:ru/ - alchol (D), /doroppu/ - drop, /sandoitči/ - sandwich, /široppu/ - syrup, /so:se:ži/ - sausage, / kasutera / - castella (P), /katsuretsu/ - cutlet, /kare:raisu/ - curried rice, /omuretsu/ - omelette (F), /tomato/ - tomato, /korokke/ - croquette (F), /bisuketto/ - biscuit, /banana/ - banana, /bifuteki/ - bifiech (F).

2. Japanised loanwords from 1950 to 1982.

a) Among the above mentioned loanwords, pão, castella and some others were introduced from Portugues in 1774, and also coffee, syrup, beer and some others from Dutch.

Many words particularly since 1950 (but

also going back to the Meiji Period, an important fundamental period), have become mixed with Japanese words, and this has caused confusion and problems sometimes in national language education in Japan. Perhaps, if you refer to a copy of Japanese advertising, (Reference 2. p. iv) you will be surprised to see how many loanwords are used compared to the number of Japanese words.

Some Japanese scholars think that there is a danger that the flood of loanwords into our language may cause damage and confusion to our Japanese Mother tongue. But the majority of these loanwords surely enrich the vocabulary, and hence the culture of Japan. But we Japanese must pay important attention to the correct pronunciation of borrowed words and syllables. We must carefully compare the differences in pronunciation between the native and the Japanised pronunciation of it.

Some of these loanwords will die naturally, through lack of use or from being out of date, while others will live on and become a natural part of our language in the future.

b) Loanwords which were abbreviated into simple syllable by the Japanese:-

モガ/mo⌉ga/ = modern girl, モボ/mo⌉bo/ = modern boy, デフレ/de⌈fure/ = deflation, デモ/de⌉mo/ = demonstration, マスコミ/ma⌈sukomi/ = mass communication, プロ/pu⌉ro/ = professional, ハイテク/ha⌈iteku/ = high technology, アパマン/a⌈pa.man/ = apartment house+mansion, パソコン/pa⌈so.kon/ = personal computor, イメチェン/i⌈mečen/ = image change, just like a noun. モガ.ビヨーイン/mo⌉ga.bi⌈jo⌉:in/ = modern girl's beauty parlor.

3. Newly made Japanised Foreign Words.

a) We call these words "Wa⌈sei-Ei⌉go," which means, roughly, "English made by the Japanese." Unfortunately, however, many of these words or expressions are not understood by foreigners from whose country the words are derived! For example, the sound of some of these words resembles English, but their meaning is not understood by native English speakers. The Japanese seem to like to create new words in this way. It is quite an interesting and creative use of languages. Don't you think so? I can see though that other people are more critical and offended by what they see as a misuse of language. Some examples are:- ナイター/na⌉ita:/ = night game, オーバー.ナイター/o:ba:.na⌉ita:/ = overnighter, which means an accessory case for a short stay, シスター.ボーイ/ši⌉suta:.bo:i/ = sister boy, means a boy like a young woman. マルチタレント/ma⌈ručita⌉rento/ = multi talent, means a talent who can do everything well, ステッキ.ボーイ/su⌈tekki.bo⌉:i/ = stick boy, means a man who works and get money by doing a woman's walking mate, opp. stick girl, ライフ.ケア/ra⌈ifu.ke⌉ja/ = life care, means care for old men, メイン.スタンド/me⌈in.su⌈tando/ = main stand, means main seat at baseball stadium, ゴールデン.ウイーク/go⌈:ruden.ui⌉:ku/ = golden week, means a week with many public holidays, ウェザー.オール/u⌈eza:o⌉:ru/ = weather all coat, ノー.ストッキング/no:.su⌈to⌉kkingu/ = no-stocking, ノー.ゲーム/no:ge⌉:mu/ = no game, means the baseball game which can not be continued before fifth inning by rain or an accident, ロマンス.グレー/ro⌈mansu.gu re⌉:/ = romance grey, means a charming middle aged man, ガソリン スタンド/ga⌈sorin.suta⌉ndo/ = gasoline stand, means, gas station.

b) Some New Trends in making loanwords.

Recently I come across some of the following new and interesting ways of making more loanwords in Japanese:-

i) noun+suru (Verb) /su⌈ru/ = do.

Japanese may be able to make new expression easily by noun+suru (Verb). suru = do.

ex.:- Japanese noun+suru./rjo⌈ko:/(travel) +suru, = travel to, or make a journey.

ex.:- such a loanword + suru. /su⌈to⌉ppu/(stop) +suru, /sa⌉iensu/(science)+suru, /i⌈mečen/(image change)+suru. etc.

ii) a loanword + Japanese noun

ex.:- /ha⌈iteku.ša⌉kai/(ハイテク.シャカイ) = society with high technology.

/no:.hau kjo⌉:iku/(ノーハウ.キョーイク) = education which respect high technical knowledge.

/nu:do. e⌉:ga/(ヌード.エーガ) = nude movie.

/kon.⌈pju:ta:/(/KON/婚ピューター) = /kon/putor. Japanised loanword. = to find a mate for marriage by computor.

/ke⌈kkon/(結婚) = marriage.

c) Some important sound change for understanding these loanwords in Japanese.

i) When a word moves into another country, not only the pronunciation may be changed according to the phonemic system of the new linguistic environment, but also the spelling may be changed, too.

ex.:- パット/pa⌉tto/ - pad/ pæd/, バッチ/ba⌉tči/ - badge/bæǰ/, ホットドッグ/ho⌈ttodo⌉kku/ - hot.dog/hot.dog(winner), グロッキー /gu⌈ro kki/ - groggy/grogi/, etc.

According to the Japanese phonemic system, ッ sound (a choked semi-vowel) appears only a voiceless or a semi-voiced sound, and by the long custom, Japanese people tend to use voiceless or semi-voiced instead of voiced.

ex.:- /d/ → /t/, /ǰ/ → /č/, and /g/ → /k/, etc.

Semi-voiced sounds are /pa/ /pi/ /pu/ /pe/ /po/, /pja/ /pju/ /pjo/.

ii) The choked sound ッ is one kind of a stop or fricative consonant having Japanese one syllable value. I think this special consonant resembles a Glottal Stop in an English consonant.

ex.:- イッショ/i⌈ššo/ = together, /š/ = fricative, パット/pa⌈tto/ = suddenly, /t/ = stop consonant.

d) It is important for us to differentiate between the pronunciation of syllables in the Japanised loanwords, and in the country of origin of the loanwords.

i) We must take particular care about the changes and differences in pronunciations of vowels, consonants, syllable accent of pitch and stress, and closed and opened syllables.

ex.:-

スポークス.マン/su⌈po⌉:kusuman/ - 7 syllables.
spokes.man / spouksmən / - 2 syllables.

コスモポリタン/ko⌈sumopo⌉ritan/ - 7 syllables.
cosmopolitan/kozməpolitən/ - 5 syllables.

ii) A little explanation of Japanese Syllabary.

The syllabary (/o⌈nsetsu.moži/) is denoted with 'kana'-characters. Below is the systematic table of the fifty sounds of the kana syllabary in the Japanese language, which was thought to have been completed about 1695 A.D.

The kana syllabary (50 sounds).

ワ (wa)	ラ (ra)	ヤ (ya)	マ (ma)	ハ (ha)	ナ (na)	タ (ta)	サ (sa)	カ (ka)	ア (a)
ヰ (i)	リ (ri)	イ (i)	ミ (mi)	ヒ (hi)	ニ (ni)	チ (chi)	シ (shi)	キ (ki)	イ (i)
ウ (u)	ル (ru)	ユ (yu)	ム (mu)	フ (fu)	ヌ (nu)	ツ (tsu)	ス (su)	ク (ku)	ウ (u)
ヱ (e)	レ (re)	エ (e)	メ (me)	ヘ (he)	ネ (ne)	テ (te)	セ (se)	ケ (ke)	エ (e)
ヲ (o)	ロ (ro)	ヨ (yo)	モ (mo)	ホ (ho)	ノ (no)	ト (to)	ソ (so)	コ (ko)	オ (o)

This table does not always show whole syllables of Japanese, but vertically 5 letters, and horisontally 10 letters, make 50 syllables.

We can see from the table that Japanese syllables are constructed in the following:-

1) one vowel .. /a/ ア
2) one consonant+one vowel .. /ka/ カ
3) one consonant+one glide+one vowel .. /kyo/ キョ, /kwa/ クヮ, etc.

We know of course all syllables have always at least one vowel, therefore Japanese is an open-syllabled.

4. In conclusion.

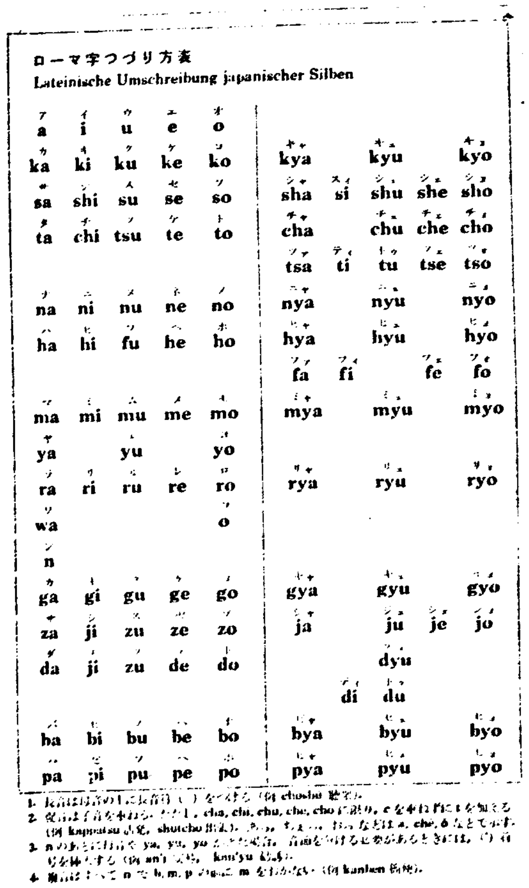
Within the limits of these four pages, I have tried to introduce you, my friends and fellow phoneticians from around the world, some of the many loanwords from different countries which have become a part of the Japanese language, and how these words have changed from their original pronunciation to become Japanised. There are so many of these words in Japanese now and it is not easy to make summary of them all in their address. But I have attempted to select a few appropriate, and I hope, interesting examples for you when I say 'bye-bye,' 'my home,' 'golden week,' and '/sayonara/home-run,' I think of these words now as Japanese words, not as Japanised foreign words. The Japanese people are now very familiar with such words.

In ending, I would like to say that language is one of the best tools we have, to enable us to become close friends with each other. It can and should strengthen mutual understanding and friendship betweem us. Language scholars can make a significant contribution to the welfare of different people and their communities.

So, let's increase exchanging programmes between our different countries. Now is the best time to implement these programmes.

Thank you for your attention!

Reference No. 1



ローマ字つづり方表
Lateinische Umschreibung japanischer Silben

a	i	u	e	o					
ka	ki	ku	ke	ko	kya		kyu		kyo
sa	shi	su	se	so	sha	si	shu	she	sho
ta	chi	tsu	te	to	cha		chu	che	cho
					tsa	ti	tu	tse	tso
na	ni	nu	ne	no	nya		nyu		nyo
ha	hi	fu	he	ho	hya		hyu		hyo
					fa	fi		fe	fo
ma	mi	mu	me	mo	mya		myu		myo
ya		yu		yo					
ra	ri	ru	re	ro	rya		ryu		ryo
wa				o					
n									
ga	gi	gu	ge	go	gya		gyu		gyo
za	ji	zu	ze	zo	ja		ju	je	jo
da	ji	zu	de	do			dyu		
						di	du		
ha	bi	bu	be	bo	bya		byu		byo
pa	pi	pu	pe	po	pya		pyu		pyo

1. [illegible] (例 chosha [illegible]).
2. [illegible] cha, chi, chu, che, cho [illegible] (例 kappatsu [illegible], shutcho [illegible]). [illegible]
3. n [illegible] ya, yu, yo [illegible] (例 an'i [illegible], kan'yu [illegible]).
4. [illegible] n [illegible] b, m, p [illegible] m [illegible] (例 kanben [illegible]).

Reference No. 2

朝日新聞　1987-1-21

松下電器の「MⅡフォーマット」新ビデオシステム
—NHKと共同で開発した方式—

National/Panasonic

時代のニーズに的確に応えた画期的ビデオ技術
ソウルオリンピックの国際中継へ向けていま実用開始

米国放送界を代表するテレビネットワークのひとつNBC放送では、情報化社会の高度化に伴ってニュースやドキュメントなど報道番組の重要性が高まるなかで、以前から小型・高性能で使い易く、画質・音質ともに優れた新しいビデオシステムを求めていました。そしてこの厳しい要求をみたす、画期的な放送用ビデオシステムとして、松下電器が開発した「MⅡフォーマット」新ビデオシステム機器を全面採用することに決定しました。

全システムを1/2インチカセットで統一

一般に放送局では、さまざまな方式のVTRを使いわけています。そして1インチや2インチのVTRはオープンリール方式ですから、操作などの際テープ操作が煩雑になり、熟練を要し、装置も大型になっていました。

「MⅡフォーマット」は、従来混在していたフォーマットを、カメラ機能から送出まで一挙に1/2インチカセットテープで統一。家庭用のVHSビデオカセットとほぼ同じ大きさですから、装置全体もこれまでの約1/2に小型化されます。カセット化でテープ操作も簡単になりました。

高画質と長時間録画を両立

「MⅡフォーマット」は、高密度記録が可能なメタルテープを採用しています。そしてダビング時の画質劣化が少ない、独自のコンポーネント記録方式を開発。これによって、1/2インチで1インチテープなみの画質を保ちながら、90分という長時間録画を可能にしました。

またポータブル業務用に小型カセットも開発し、ニュース取材用のポータブル式放送用カメラ一体型ビデオも、従来の約2/3の重さに小型化。さらに「MⅡフォーマット」は、従来の放送用ビデオのオーディオ2チャンネルに、FM2チャンネルを加え、音質を飛躍的に向上させました。この点もNBC放送が、「MⅡフォーマット」全面採用に踏み切った大きなポイントです。

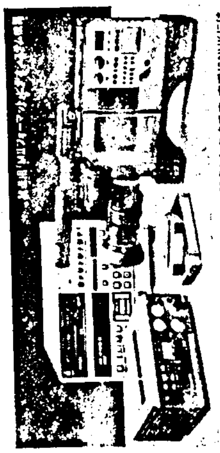
自動送出装置にも対応

放送システムとして欠かせない自動送出も重要なシステムのひとつです。「MⅡ自動プログラムシステム」への展開によって、コンピュータ制御のプログラムに沿って以前の記録システムとも連動し、ニュースやドラマ、コマーシャルなどを自動的に送出することが可能になります。

「MⅡフォーマット」は、まさに放送の将来を見すえたビデオシステムといえます。

総合ビデオ技術に高い評価

松下電器は、家庭用ビデオはもちろん、早くから業務用ビデオのあらゆるシステム機器の開発に取り組んできました。今回の「MⅡフォーマット」ばかりでなく、その総合ビデオ技術は、すでに国内外で高い信頼を得ています。



●MⅡフォーマットは、NHKと松下電器が共同で開発した方式です。すでにNHKには、200台におよぶ「MⅡフォーマット」によるスタジオ用ビデオ装置を納入。各地の放送局で番組制作・送出用機器として活躍しています。

Key to phonetic symbols

Vowels and diphthongs

1	iː	*us in*	see /siː/	11	ɜː	*us in*	fur /fɜː(r)/
2	ɪ	*us in*	sit /sɪt/	12	ə	*us in*	ago /ə'gəʊ/
3	e	*us in*	ten /ten/	13	eɪ	*us in*	page /peɪdʒ/
4	æ	*us in*	hat /hæt/	14	əʊ	*us in*	home /həʊm/
5	ɑː	*us in*	arm /ɑːm/	15	aɪ	*us in*	five /faɪv/
6	ɒ	*us in*	got /gɒt/	16	aʊ	*us in*	now /naʊ/
7	ɔː	*us in*	saw /sɔː/	17	ɔɪ	*us in*	join /dʒɔɪn/
8	ʊ	*us in*	put /pʊt/	18	ɪə	*us in*	near /nɪə(r)/
9	uː	*us in*	too /tuː/	19	eə	*us in*	hair /heə(r)/
10	ʌ	*us in*	cup /kʌp/	20	ʊə	*us in*	pure /pjʊə(r)/

Consonants

1	p	*us in*	pen /pen/	13	s	*us in*	so /səʊ/
2	b	*us in*	bad /bæd/	14	z	*us in*	zoo /zuː/
3	t	*us in*	tea /tiː/	15	ʃ	*us in*	she /ʃiː/
4	d	*us in*	did /dɪd/	16	ʒ	*us in*	vision /'vɪʒn/
5	k	*us in*	cat /kæt/	17	h	*us in*	how /haʊ/
6	g	*us in*	got /gɒt/	18	m	*us in*	man /mæn/
7	tʃ	*us in*	chin /tʃɪn/	19	n	*us in*	no /nəʊ/
8	dʒ	*us in*	June /dʒuːn/	20	ŋ	*us in*	sing /sɪŋ/
9	f	*us in*	fall /fɔːl/	21	l	*us in*	leg /leg/
10	v	*us in*	voice /vɔɪs/	22	r	*us in*	red /red/
11	θ	*us in*	thin /θɪn/	23	j	*us in*	yes /jes/
12	ð	*us in*	then /ðen/	24	w	*us in*	wet /wet/

/'/ represents *primary stress* as in about /ə'baʊt/
/ˌ/ represents *secondary stress* as in academic /ˌækə'demɪk/

(r) An 'r' in parentheses is heard in British pronunciation when it is immediately followed by a word, or a suffix, beginning with a vowel. Otherwise it is omitted. In American pronunciation no 'r' of the phonetic spelling or of the ordinary spelling is omitted.

/-/ Hyphens preceding and/or following parts of a repeated transcription indicate that only the repeated part changes.

⇨ the Introduction for a full explanation of the phonetic information.

PERCEPTION OF PARALINGUISTIC CUES OF AGE AND SEX IN MANIPULATED SPEECH: AN EXPLORATION

Leo W.A. van Herpt

Institute of Phonetic Sciences
University of Amsterdam

ABSTRACT

An experiment is performed to explore the possibility to eliminate or control different paralinguistic phenomena in spoken texts in such a way that fundamental perceptual dimensions can be judged more or less in isolation.
Specifically the effect of several acoustic manipulations of speech, coupled with different degrees of content masking, on the perception of voice and pronunciation qualities, and the attribution of age and sex are studied.

1.0 INTRODUCTION

A major problem in phonetics is the generally low correlation between perceptual ratings in speech characteristics and the supposed acoustic criteria of these attributes. To be able to identify the acoustic correlates of perceptual voice and pronunciation characteristics it is essential to have reliable and valid perceptual judgments.

Scaled values of Voice and Pronunciation (V&P) obtained through the use of the semantic differential method have been found to be satisfactorily reliable [6] but attempts to validate the measures against an external criterion are quite unsuccessful [3]. Judgments of V&P in connected speech are probably particularly contaminated by irrelevant factors due to halo-effects and stereotyping [8]. Both mechanisms produce systematic errors. Stereotyping transforms perception in the direction of expected behavior, and halo-effect biases judgments on the basis of one particular feature. Among possible irrelevant influences are content and intelligibility of the text and inferences concerning emotions, age and sex of the speaker. The result is a tendency to bias ratings on all scales if one of the attributes deviates from what is expected. The resulting semantic differential thus represents a general impression rather than a strict scaling of various (independent) V&P attributes.

Our work aims at increasing the objectivity of ratings of V&P cues by trying to eliminate information which gives rise to stereotyping and halo-effects. The term 'V&P cues' refers to content-free measures of speech, especially those components of paralanguage which Trager [14] calls voice qualities and qualifiers which include such things as pitch height, pitch range, glottis control, resonance, intensity, tempo, rhythm control and articulation control. To have listeners respond optimally to those vocal qualities of speech the verbal meaning has been masked in the present experiment. This has been done in several ways and degrees to eliminate or mask also several paralinguistic phenomena in order to give more prominence to other paralinguistic dimensions that have to be judged.

Specifically the present exploration was designed to study the effect of manipulations of several paralinguistic speech components coupled with different degrees of content masking on the perception of fundamental perceptual dimensions of V&P and on the perception of vocal age and sex cues.

2.0 EXPERIMENT

The experiment consists of an evaluative rating by 47 listeners based on one minute oral reading from six Dutch native speakers. Each speech sample is manipulated in seven different ways. These 42 fragments and the unmanipulated text twice are rated on fifteen seven-point bipolar semantic scales. In addition the judges indicated in each condition the supposed age and sex of the speaker.

2.1 Speech material

The speech material consists of an identical text of about one minute duration read aloud by the speakers: three men - 28, 32 and 36 of age - and three women - 28, 28 and 32 of age. An oral reading text was chosen to control for between-speaker differences in lexicon and syntax. These six fragments have been analysed, manipulated and resynthesised on a Data General computer (Eclipse S/200) at a sample frequency of 10 kHz. In order of presentation the following manipulations of stimuli are performed.

1. **Reverse.** This procedure is comparable to playing backwards a tape recording at normal speed. The method eliminates the necessity to control the general effectiveness in conveying meaning [12], the content is fully unintelligible, but overall pitch spectrum is preserved.

2.**Splice.** Scherer's randomized splicing procedure basically "consists of cutting a stretch of recording tape into pieces and splicing them back together in random order" [11]. We randomized stretches of 50 milliseconds which results in total unintelligibility. Random splicing preserves the voice spectrum and eliminates or masks voice dynamics such as rhythm and intonation. Since many of the pauses are broken up and since the parts are randomly distributed the tempo impression is more distorted than under the 'reverse' condition.

3. Scramble. This procedure divides the text in segments of equal length - in our case of 1 millisecond - which are alternatedly multiplied by +1 and -1. The resulting discontinuity causes a perceptually unpleasant tone which we reduced by filtering the signal low-pass at 1000 Hz (Butterworth 4th order). The content of the text is completely lost, and along with the high frequencies also a good deal of the voice quality, though indications of suprasegmental phenomena can still be heard.

4. Speech Babble. In this procedure a text is, so to say, several times piled on itself. The number of 'echoes', the distance between the starting points of the echoes and the damping factor are parameters in the program. Our stimuli are synthesized with five echoes at a distance of one second and without damping. We surmise that perceptual judgments of this type of stimuli may be related to long-term average spectra because this condition focusses the attention of the listener on the 'average' frequency spectrum instead of on pitch variation which is commonly related to intonation.

5. Normal. The fifth and the the ninth condition are identical and consist of the original, unmanipulated recordings.

6. Filter. The three female voices have been low-pass filtered at 250 Hz; the three male voices at 150 and at 250 Hz. (This manipulation is only partly effective because after filtering the signals are amplified again.) From the ratings of judges it appears that also in this condition almost all content is lost. According to Kramer [9] and Starkweather [13] along with the high frequencies most of what is usually called voice quality is filtered out.

7. Whisper. The signal is resynthesized with noise as source. Because of the spectral roll-off of the noise source, the frequency components in the vicinity of the very low frequencies are relatively strong, which results in an impression of roughness in the intensity domain and poor intelligibility.

8. Vocoder. A seven channel vocoder analysis has been performed. For resynthesis of the fragments the seven spectral envelopes are used again. Noise with a spectrum identical to the average speech spectrum served as a carrier. The result is a whisper-like speech of reasonable intelligibility.

2.2 Stimulus tape

We had in view to present the stimuli in order of intelligibility. A pilot study with three expert listeners showed that the intelligibility of the conditions Reverse, Splice, Scramble and Speech Babble are respectively nil to minimal. Filter, Whisper and Vocoder - in that order - are less content-masked which might cause the listener to concentrate on the content of the stimuli. Hence, a Normal condition is inserted between Speech Babble and Filter. The ninth condition is again Normal which enables us to determine the reliability of this measurement.

In each condition the speech samples of the six speakers are presented in random order, with exception of the Filter condition in which first the three male voices with cut-off point at 250 Hz are presented, next the female also at 250 Hz and after that again the male voices, now at 150 Hz.

Each sample lasts 50 seconds and there is an interstimulus interval of 5 seconds in which the next speaker is announced.

In order to give the listeners an impression of the type of stimulus to be judged each condition is preceded by an example of the relevant manipulation. These examples consist of 15 seconds of another woman's and 15 seconds of anothers man's voice, relevantly manipulated.

2.3 The rating instrument

The scales. The rating instrument which is used to acquire the perceptual paralinguistic measures is constructed for the description of V&P quality of Dutch speakers. The instrument originated from a master pool of some 800 adjectives referring to attributes of speech, from which 85 scales existing of bipolar items were composed [2]. Scales which are semantically redundant or statistically inappropriate were removed [1]. Factorial studies [1],[5] led to further elimination of scales and showed that a reasonably stable perceptual space can be spanned on the basis of seven times two bipolar scales. Each pair is selected on account of their similarity in meaning as expressed by their closeness in semantic space. The seven pairs of scales and their clusternames are shown in Table 1. (Scale 15 is added on behalf of the present experiment.) The scorings of scale 8:'soft-loud' should in the present case not be interpreted in a literal (physical) sense because all fragments on the stimulus tape are amplified to approximately the same intensity.

Table 1. Clusters1) and their scales2) with Values of Ideal Voice and Pronunciation3)

Ia.	Voice Appreciation: Melodiousness		
	01. monotonous	- melodious	(6.16)
	02. expressionless	- expressive	(6.32)
Ib.	Voice Appreciation: Evaluation		
	13. ugly	- beautiful	(6.26)
	14. unpleasant	- pleasant	(6.73)
II.	Articulation Quality		
	03. slovenly	- polished	(5.95)
	04. broad	- cultured	(6.09)
IIIa.	Voice Quality: Clarity		
	05. dull	- clear	(5.92)
	06. husky	- not husky	(5.63)
IIIb.	Voice Quality: Subjective Strength		
	07. weak	- powerful	(5.42)
	08. soft	- loud	(4.04)
IV.	Pitch		
	09. shrill	- deep	(5.04)
	10. high	- low	(4.18)
V.	Tempo		
	11. dragging	- brisk	(5.63)
	12. slow	- quick	(4.69)
-	Intelligibility		
	15. good	- bad	

1) Cluster are indicated by Roman numerals, Scales by Arabic numerals.
2) To facilitate readibility and statistical treatment all scales are polarized with the scale term that, according to the Ideal Voice value, is the more desirable one to the right.
3) Values of Ideal Voice & Pronunciation on seven-point rating scales from [12].

The scoring form. The two scales of each cluster are separated, which brings about two parallel testhalves. The two testhalves are presented on separate pages of the scoring form. To control for sequence effects the scales are presented in two different orders. In table 1 all scales are oriented with the positive pole to the right; on the scoring form the polarity of every second scale is reversed. The stimuli are scaled through an application of the method of equal appearing intervals, employing a seven-point scale.

2.4 Procedure

The experiment was performed in the language laboratory of the University of Amsterdam. For this purpose the original tape was copied on cassettes, so the raters could work individually. After the instruction, and prior to the presentation of the tapes, the listeners were familiarized with the scales and the rating procedure by scoring the semantic differential of their own voice. Then the listeners gave their ratings while listening to the speech samples. When the listener had finished his ratings of a speech sample, he indicated perceived age and sex of the speaker. At the end of the session the listeners scored the semantic differential for the typical V&P of both a man and a woman at the age of 30. The listening task in all required approximately an hour and a half.

2.5 Raters

The rating experiment was carried out with 25 female and 22 male students of the Faculty of Arts of the University of Amsterdam. The raters are 21 - 34 years of age; mean age of women being 25.4 years, mean age of men 24.3 years. Since it is known that sex of rater influences their judgments [4],[8], a sample size of 25 raters in each sex group was planned - in general that suffices for an effective reliability >.90 of the scales [6],[7].

Raters and speakers are chosen from the same age category because of a possible interaction between listener's age and attributed speaker's age which might bias by way of halo-effects or stereotyping the scoring on other scales too.

3.0 DATA TREATMENT

Combination of Normal Conditions. Comparison of the two Normal conditions (5 and 9) shows a high reliability of the scales. All mean differences between the two conditions over all scales for different partitions of the speaker and rater samples are smaller than a fourth of a scale unit and none of the differences is significant, which implies that it is allowed to combine the ratings of the two conditions to a single set of scores.

Combination of scale pairs. The correlation matrix of the fifteen scales in condition 5+9 (Normal) with all speakers and raters shows eight correlation coefficients >.60. Scale pair 5 and 6 excepted the scale pairs of each cluster (1-2, 3-4,...13-14) prove to be highly correlated. Factor analyses of several partitions of the sample show the same picture, hence we consider it justified to combine the scores of those scale pairs occasionally.

Combination of raters or speakers. The raters clearly differentiate between female and male speakers. Frequency distributions on all scales except 15:Intelligibility in condition 5+9 (Normal) differ significantly (P<.001). Against our expectations [8] sex of rater did not influence the judgments in this condition significantly. So it is allowed to combine the scores of the raters, but it is imperative to give separate descriptions of female and male speakers.

4.0 RESULTS

4.1 Intelligibility

The intelligibility of the eight conditions, as perceptually judged by the raters, does not completely match the ranking of the experts (see 2.2). The Filtered stimuli were considered fairly intelligible by the experts and therefore presented after the first Normal condition. Our listeners judged quite differently and ranked Filter about equal in intelligibility to Splice. The disagreement may be explained by the use of trained versus naive listeners. According to Kramer [10] band-pass filtering tends to enable the listener to pick up gradually more of the content after repeated exposure - a circumstance more applicable to our experts.

The low-judged intelligibility of Scramble in relation to Splice is ascribed to the unpleasant impression of the stimuli. Scramble is scored very low on on the two Evaluation scales, whereas Splice is second best after Normal (Table 2).

Table 2. Intelligibility (scaled 0-10) for female (♀) and male voices (♂) voices in eight conditions.

Conditions	(♀)	(♂)	Rank
1. Reverse	1.5	1.4	2
2. Splice	3.0	2.1	3
3. Scramble	1.7	1.1	1
4. Sp.babble	4.3	3.3	5
5+9 Normal	8.5	8.5	8
6. Filter	3.1	2.5	4
7. Whisper	5.3	5.4	6
8. Vocoder	6.2	5.0	7

4.2 Effect of conditions

The perceptual dimensions are influenced differently by the various conditions.

Splicing leads to a higher Appreciation (cluster Ia+Ib) of the female voice, whereas conditions in which the fundamental frequency is manipulated (Whisper, Vocoder) lower it. Male voices, however, are rated highest on Appreciation in the Normal and lowest in the Scramble condition. Articulation Quality (II) is not systematically influenced by the different conditions. Clarity (IIIa) is - for men and women - lowered by Whisper and, less anticipated, heightened by Splice. The impression of Subjective Strength (IIIb) is weakest for both sexes in the Vocoder condition and, more interestingly since all voices are amplified to about the same intensity, strongest when the female voice is 'Babbled' and when the male voice is 'Scrambled'. The male voice is deeper and lower (IV) when the speech is Reversed and higher when Scrambled. The same applies for the female voice as far as the scale 'shrill-deep' is concerned; it is higher when Babbling and lower when Whispering. The perception of Tempo (V) of female and male speech is influenced differentially by 'pitch' manipulations: Whisper and Filter slow down female speech, Vocoder the male Tempo. Speech Babble is considered brisk and quick for both sexes.

A tentative conclusion from the foregoing is that listeners accentuate, contingent on (e.g. sex or age) characteristics of the speaker, specific qualities of the (speech)signal when describing speakers on a semantic scale.

4.3 Incorrect attributions of age and sex

Some conditions give rise to more incorrect attributions of age and sex than others. Table 3 summarizes for female and male speakers separately, the number of cases in which the speaker was, wrongly, estimated 50 years or older and the number of false sex attributions in all conditions. (Maximum score: 3 x 47 = 141.)

Table 3. Number of false age and sex attributions of female (♀) and male (♂) speakers, in all conditions.

Condition	Age			Sex		
	(♀)	(♂)	nR1)	(♀)	(♂)	nR1)
1. Reverse	8	22	4	7	3	3
2. Splice	6	4	6	0	0	6
3. Scramble	22	36	4	32	22	3
4. Sp.Babble	4	26	0	3	0	0
5+9 Normal	1	11	0	4	0	0
6. Filter	6	47	1	1	12	1
7. Whisper	27	56	0	111	0	0
8. Vocoder	19	29	0	105	6	0

1) Number of non-Responses

The manipulations of the stimuli gave rise to about 600 false attributions of age (estimations of 50 years and older) and sex. The relation between the attributions and the perceptual dimensions find a simple expression in the frequency distributions of the ratings. For sake of brevity we will restrict ourselves to a descriptive summary of those distributions in relation to sex attributions. The conclusions given below are based on the scorings averaged over all manipulations; the conditions that create the most salient differences are mentioned.

On the dimension of Melodiousness 'false' attributed female and 'false' attributed male are rated similarly and positioned between the 'good' attributed men and women. This shows clearest in Vocoder and Scramble. On the Clarity dimension the distribution concerning 'false' women is almost the reverse of that of the 'good' women and resembles the 'good' male curve. The 'false' men are less conspicuous except in the Scrambled condition which shows a higher clarity for 'good' women over 'good' against concordant ratings for 'false' groups. Keeping in mind that all voices are of about the same intensity, it is striking that the raters consider the female voices to be louder and more powerful than the male voices and rate the 'false' attributions somewhere in between. In the Pitch dimension the curves of 'false' attributions are again positioned between the 'good' curves, but in this case more in the direction of the 'good' male curve. This shows clearest in Scramble and Whisper. In the Tempo dimension the same picture arises, however this time with the 'false' curves more aligned with the 'good' female scores. Filter and Whisper are most indicative. Articulation Quality, Intelligibility, and Evaluation do not differentiate between good and false attributions. Concluding, the general picture that arises from this section, is that the distributions of ratings of correct attributions of the two sexes mirror each other whereas both 'false' distributions are quite identical with values between those of the correctly attributed sexes. The general direction of the 'false' curves seems to indicate whether the quality concerned is weighted heavier for men or women.

ACKNOWLEDGEMENT

I wish to thank R.J.J.H. van Son (Institute of Phonetic Sciences, Amsterdam) who wrote the software to produce the stimuli in condition 1-4, H.J.M. Steeneken (Institute for Perception TNO, Soesterberg) who manipulated the speech fragments of condition 8, and I.M. de Leeuw and H.M. Nederlof who assisted in the experiment.

REFERENCES

[1] Blom, J.G. & Herpt, L.W.A. van, "The evaluation of jury judgments on pronunciation quality", Proceedings Inst. Phon. Sciences, Univ. of Amsterdam, 4 (1976), 31-47.

[2] Blom, J.G. & Koopmans-van Beinum, F.J., "An investigation concerning the judgment criteria for the pronunciation of Dutch", Proceedings Inst. Phon. Sciences, Univ. of Amsterdam, 3 (1973), 1-24.

[3] Boves, L., "The phonetic basis of perceptual ratings of running speech", Dordrecht/ Cinnaminson, Foris, 1984.

[4] Boves, L., Fagel, W.P.F & Herpt, L.W.A. van, "Conceptions of women and men concerning the speech of men and women" (in Dutch), De Nieuwe Taalgids 75 (1982), 1-23.

[5] Fagel, W.P.F. & Herpt, L.W.A. van, "Analysis of the perceptual qualities of voice and pronunciation", Proceedings Inst. Phon. Sciences, Univ. of Amsterdam, 7 (1982), 1-25.

[6] Fagel, W.P.F., Herpt, L.W.A. van & Boves, L., "Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation", Speech Communication 2 (1983), 315-326.

[7] Herpt, L.W.A. van & Hoebe T., "Attribution of age from perceived speech", Proc. Inst. Phon. Sciences, Univ. of Amsterdam 9 (1985), 1-22.

[8] Herpt, L.W.A. van, "Influence of rater's sex on voice and pronunciation assessment", Proceedings Inst. Phon. Sciences, Univ. of Amsterdam, 10 (1986), 19-39.

[9] Kramer, E., "Judgements of personal characteristics and emotions from non-verbal properties of speech", Psychological Bulletin 60 (1963), 408-420.

[10] Kramer, E., "Elimination of verbal cues in judgments of emotion from voice", J. Abnormal and Social Psychology 68 (1964), 390-396.

[11] Scherer, K.R.,"Randomized splicing: A note on a simple technique for masking speech content", J. Exp. Res. in Personality 5 (1971), 155-159.

[12] Sherman, D.,"The merits of backward playing of connected speech in the scaling of voice qualities disorders", J. of Speech and Hearing Disorders 19 (1954), 312-321.

[13] Starkweather, J.A., "Content-free speech as a source of information about the speaker", J. Abnormal and Social Psych. 52 (1956), 394-402.

[14] Trager, G.L., "Paralanguage: A first approximation", Studies in Linguistics 13 (1958), 1-12.

ОСОБЕННОСТИ МУЖСКОЙ И ЖЕНСКОЙ РЕЧИ В СОВРЕМЕННОМ РУССКОМ ЯЗЫКЕ

Е.А.ЗЕМСКАЯ М.В.КИТАЙГОРОДСКАЯ Н.Н.РОЗАНОВА

Институт русского языка АН СССР Москва I2I0I9 СССР

ABSTRACT

The study of the peculiarities of male/female speech has until now not been undertaken for the Russian language. The contrast male/female speech manifests itself differently on different linguistic levels and in different spheres of language. Modern female/male speech is not linguistically homogenious. The parameter male/female speech is superimposed upon other social characteristics of speech. The difference between male and female speech is not an absolute one. It cannot be formulated in terms of stricts rules based on a yes/no principle. The difference between male and female speech manifests itself as certain tendencies.

Социальная дифференциация пронизывает язык по многим параметрам:возрастным,профессиональным,образовательным,локальным (место рождения,место жизни).Между тем одно из основных членений,разделяющее весь род людской на два класса - мужчина vs женщина,до сих пор обычно остается за пределами внимания языковедов.По крайней мере по отношению к русскому языку работ,изучающих противопоставление мужская/женская речь, не существует. Вместе с тем это основное противопоставление человеческих личностей,находящее отражение во всех ипостасях существования человека - от сферы физиологии и психологии до сферы социальной и культурной - получает своеобразное преломление в языке.

Исследование особенностей мужской и женской речи,которые обнаруживаются на разных уровнях языка,проводится в Институте русского языка АН СССР в рамках общей темы"Русский язык в его функционировании.Оно осуществляется на материале современной литературной устной речи.Основным источником материала являются непосредственные наблюдения и фиксация устной речи.В ряде случаев применяется опрос информантов.Для выработки общей концепции представляется плодотворным использование результатов исследований в смежных гуманитарных областях (этнография,социология,психология, психолингвистика,педагогика).

В таких языках,как русский,различие между мужской и женской речью не накладывает запретов на использование тех или иных грамматических форм (за исключением грамматических показателей принадлежности речи мужчине или женщине -глагольные флексии и пр.) или звуковых единиц.Однако оно проявляется на разных уровнях языкового членения,прежде всего в сфере фонетики и лексики.Противопоставление мужская/женская речь касается также коммуникативно-прагматических условий, регулирующих использование языка.Некоторые различия между мужской и женской речью определяются речевым этикетом,правилами общественного поведения.

Существенные различия между мужчинами и женщинами наблюдаются в стратегии и тактике речевого поведения.Мужчины и женщины по-разному строят свою речь,по-разному обращаются к знакомым и незнакомым,лицам своего и иного пола.Выявление конкретных особенностей речевого поведения партнеров коммуникации,состоящей из лиц одного пола или разных полов (мужчина-мужчина,женщина-женщина,женщина-мужчина) -необходимый этап изучения различий между мужской и женской речью.В настоящее время можно утверждать,что стратегия речевого поведения мужчин и женщин касается не только выбора лексики и фразеологии.Так,существенно различны особенности строения текста,связанные с переходами к новой теме.В женской речи выше степень ассоциативных переходов от темы к теме,мена темы больше зависит от влияния чисто ситуативных факторов.

Проблема выявления особенностей мужской и женской речи теснейшим образом связана с действием в языке механизма так называемой категории вежливости /7;I0/. При этом остро встает вопрос о том,какие из обнаруженных особенностей мужской и женской речи имеют универсальный (или,по крайней мере,-общеевропейский,общий для цивилизованных языков)характер,а какие свойственны отдельным языкам или отдельным группам языков/I5/.

Различия между мужской и женской речью соотносимы с существующими в данном обществе наборами и иерархией социальных ролей,включая профессиональные роли.Несмотря на равноправие мужчин и женщин в СССРперед законом и конституцией,социальные и семейные роли по-разному распределены между ними, что находит проявление в использовании мужчинами и женщинами языка,в их речевом поведении.

"Ролевой"фактор необходимо учитывать при исследовании преобладающей и специфической тематики бесед(женщины- мода, кулинария,дети и т.д.;мужчины- спорт,техника,политика и т.д./2/).Он оказывается значимым и для выявления распределенности типовой тематики при реализации фатической функции языка(погода,здоровье,дети, мода Vs спорт,политика).Следствием подобных различий является разная степень владения лексикой ряда тематических групп (непосредственные наблюдения за речью мужчин и женщин подкрепляются анализом материалов "Вопросника по лексике современного русского языка"и путем опроса информантов/3/).Некоторые общие тенденции в различении мужской и женской речи на уровне лексики обнаруживают сильную зависимость от двух параметров -образование и профессия.

Мужская/женская речь не представляют собой единого,целостного в лингвистическом отношении объекта.Социальное расслоение касается мужской/женской речи так же, как и речи вообще.

Факт различия между речью мужчин и женщин не является особенностью лишь нашего времени.Хотя противопоставление мужская/женская речь не исследовалось подробно по отношению к истории русского языка,в отдельных работах оно рассматривалось/1/.

Обратимся к фонетике.Фонетика представляет собой тот уровень языка,в котором различие между мужской и женской речью предстает иначе,чем на других уровнях,что объясняется как незнаковым характером этого уровня,так и тем,что такие фонетические особенности речи,как тембр,мелодика, экспрессивные модуляции голоса и т.п.,тесно связаны не только с социальным поведением мужчин и женщин, но и со строением мужского и женского организма.В то же время при рассмотрении ряда фонетических явлений(особенно в области интонации)обнаруживается изоморфизм данного уровня с другими уровнями языка(с лексикой,синтаксисом).

На сегментном уровне обращают на себя внимание следующие различия.

I. В области вокализма можно отметить ряд особенностей в тембральной окраске гласных,связанных с тем, что для многих мужчин характерен меньший раствор рта при производстве звуков,чем для женщин. Это приводит к образованию более "узких" гласных,менее богатых по тембру.Типичными можно считать,например,такие особенности:

1.Растяжка I предударного [а] .В женской речи в позиции I предударного слога после твердых согласных на месте ⟨а⟩ ,⟨о⟩ частовозможно произношение широкого открытого [а:] ,длительность которого равна ударному или превышает его: [мъга:з'и́н], [ръска:за́лъ] , [пръда:jу́т]. Эта особенность интересна тем, что раньше она характеризовала старомосковское произношение,и до сих пор встречается в речи старых москвичей (растяжка предударного [а] была свойственна речи Д.Н.Ушакова,Б.Л.Пастернала).

В современном мужском произношении в этой же позиции встречается часто гласный более узкий,приближающийся по тембру к [ъ] . Ср.,например,такие факты, записанные от мужчин-москвичей,носителей литературной нормы:

[ръзга^{ъ}во́ры], это [п'ътра^{ъ}гра́цкъь] сторона, [па^{ъ}го́дь].

Эти особенности определяют различия в ритмике слова. Наряду с общелитературной моделью тътʌта́ в современном литературном произношении встречаются две другие: тъта:та́ и тъта^{ъ}та́.Если первая из них особенно частотна в женской речи, то вторая более широко распространена в мужском произношении.

Распространенность растяжки предударного [а] в современном произношении женщин - это случай перераспределения старых произносительных вариантов по другим социальным признакам.Интересно,что старомосковская растяжка [а] наблюдается и в речи коренных ленинградок /I2,I74/.Подчеркнем, что произношение широкого долгого [а:] в этой позиции,как свидетельствуют данные научной литературы,никогда не было свойственно речи петербуржцев.

2.Для женского произношения также характерна бо́льшая дифтонгичность ударных [о] и [э].Неоднородность этих гласных особенно заметна,если на них находится фразовый акцент:

Нас в санат[^{у}о]рий отправляют//; А на [л'^{и}э]то куда́ собираетесь?

В мужской речи в этих же фразовых условиях обычно произносятся более однородные гласные.

П.В области консонантизма можно отметить одну общую тенденцию,свойственную мужскому произношению:меньшая степень напряженности артикуляции согласных.Эта тенденция обусловила распространение следующих особенностей в мужской и женской речи:

1. В женской речи наблюдается аффрикатизация [т'] , [д'] (цеканье-дзеканье),менее характерная для мужской речи /4/.

2. Меньшая степень напряженности согласных в мужском произношении обусловила ряд звуковых изменений в потоке речи,более частотных у мужчин:

ослабление смычки у согласных - [п]очему, [д]авайте, [ш']еловек;

озвончение согласных (ассимиляция по звонкости под действием соседних гласных и сонантов) - [пъта^{ъ}му́штъ]. Эта особенность наиболее ярко обнаруживается в заударной части слова в слабой фразовой позиции:

Ну а почему и не́т допустим? Почему не вы- [п'ьт'] рюмку коньяка́ допустим [дыпъ]?

Однако при эмфатическом произношении в ударных или предударных слогах напряженность и длительность согласных возрастает:

[з:]а[р:]аза такая!; Ух [ж:]арко!; Вот [д:]ура! /II,I49/.

Обращает на себя внимание также бо́льшая консонантная насыщенность мужской речи,обусловленная тем, что для мужчин характерна более сильная деформация гласных в потоке речи, их количественная и качественная редукция, выпадение гласных.Ср. следующий фрагмент мужской речи:

(о театральной постановке) А.Вы мха́товский вариант видели в театре? - [въ мха́тъ фск'и вър'а́нт в'и̂д'л'ь фт'а́тр'ь] Б. Угу// А. Ну почему́ кошмар? Кто́ там играет? - [ну пъш'му́ ка^{ъ}шма́р/ кто́ тъм ъгрът].

Таким образом, при сопоставлении сегментных характеристик можно отметить,что особенности женского произношения наиболее ярко проявляются в сфере вокализма,а мужского - в сфере консонантизма.

Рассмотренное противопоставление тесно связано с просодическими различиями в мужской и женской речи. Остановимся на некоторых из них.

1. При акцентном выделении слов во фразе обнаруживаются различия в фонетическом оформлении акцентно выделенных слов.(О коммуникативных функциях фразовых акцентов в устной речи см./9/). Так, в женской речи широко представлена растяжка ударного гласного. Причем этот способ акцентного выделения обнаруживается в разных жанрах устной речи:

(из научного доклада) Я думаю что перед нами вообще/ очень увлека́-ательная/ очень интере́-есная проблема//; (из телевизионной передачи) Это был такой доморо́-ощенный орке́-естр//; (из разговорной речи) Я ей поме́-ерила только/ я же ей по-ме-ри-ла/ а она говорит ба́-абушка/ не снима́-ай//.

В мужской речи в акцентно выделенных словах шире используется растяжка согласного:

(из научного доклада)На русском языке/говорит русский языковой индив-ви́д//; (из спортивного телерепортажа) Эта дистанция опять показала свой к-кова́рный нр-ра́в//; (из разговорной речи; о нарисованной в детстве картине) Притом масляными красками/ и/ п-па́льцем рисовал//

Существенно отметить,что эта предпочтительность в выборе фонетических средств наблюдается и в устных научных выступлениях,где различия между мужской и женской речью наиболее сглажены. Конечно,невозможно утверждать,что в речи мужчин совсем не допускается растяжка гласных.При выражении некоторых значений она вполне естественна и даже необходима (например,при перечислении /II/). Однако высокая встречаемость растяжки гласного в словах с акцентным выделением вносит в мужскую речь оттенок назидательности и может вызвать отрицательную реакцию слушателей.

2. Широкое использование растяжки ударных гласных создает условия для более яркого выражения мелодики на этих гласных. Так, по наблюдениям С.В.Кодзасова,при выражении несогласия в отрицательном ответе нормативным является использование сочетания падающего тона или положительного акцента с модуляцией:

Это Ваня приехал? - Не́т.Пе́тя./5/

В женской речи модулированные тоны и акценты выражены особенно рельефно за счет удлинения гласного:

А. Лёни не было// Б. Не́-ет/ Лёня бы́-ил! Да ну ка...ка́-ак же!; А. Это не твоя ручка? Б. Не́-ет/ у меня япо́-онская//.

Модуляция голоса при растяжке гласного широко используется женщинами и в других речевых ситуациях,например,при обращении к детям или животным:

(разговаривает с попугаем по кличке Бони) Ну давай поговори́-им// Давай поговори́'-им/ поговори́'-им конечно Бонечка//.(Отметим, что во 2-ой и 3-ей фразах модуляция сочетается с гортанной смычкой в середине гласного).

3. По нашим наблюдениям, женщины чаще используют интонационные средства для выражения многих значений,тогда как мужчины в этих же речевых ситуациях обычно прибегают к средствам лексики и грамматики.Например,в женской речи широко распространены оценочные высказывания типа:

Он та́-ако́й симпати́чный!; Она та́-ака́я проти́вная!, имеющие специфическое просодическое оформление (положительный акцент на слове такой, часто сочетающийся с растяжкой предударного гласного и восходящим тоном на этом гласном+восходящий тон на оценочном прилагательном). Мужчины для выражения экспрессивной оценки чаще используют лексические средства (отличный,отлично,здорово и др.).Ср.,например,фрагмент из рассказа супругов о просмотренном фильме:

Жена: Это та-акой фильм! Муж. Да/ отличная картина!

Таким образом,женской эмоциональной речи свойственна просодическая эксплицитность,тогда как для мужчин характерна эксплицитность лексическая.

Чуткость женщин к интонационному рисунку речи неоднократно отмечалась в художественной литературе. Ср.такое наблюдение: "...-Везет же некоторым! Так судачили женщины,и вопрос,кому именно везет,мог толковаться как угодно: женщины объясняются чаще всего не словами,а интонацией, как птицы".(Борис Васильев."Рослик пропал...")

Выводы

I.Анализ современной спонтанной литературной речи показал,что противопоставление мужская/женская речь на фонетическом уровне является весьма существенным и лингвистически значимым для современного русского языка.

2.Современная женская/мужская речь лингвистически неоднородна.Она дифференцирована по ряду признаков:возрастной,образовательный,профессиональный,территориальный,индивидуально-психологический.Таким образом,можно сказать,что компонент "женская речь"/"мужская речь" накладывается на другие социальные характеристики речи.

3.Различие между мужской и женской речью не является абсолютным.Оно не может быть сформулировано в виде строгих правил, строящихся по принципу да/нет.Различие между мужской и женской речью проявляется в виде тенденций (женской речи свойственно более..,мужской речи свойственно более...).Эти тенденции в принципе допускают нарушения,т.е.имеются мужчины,говорящие как это характерно для женщин,и женщины,говорящие как это характерно для мужчин.Однако нетипичность такого говорения очевидна.Она обнаруживается в том,что женская манера речи у мужчин всегда обращает на себя внимание и нередко ведет к передразниванию.Отрицательное отношение к такой манере закреплено в выражениях типа: "бабья интонация", "бабья манера".Отметим, что мужская манера речи у женщин,по нашим наблюдениям,хотя и может вызывать негативную оценку,но обычно не ведет к передразниванию.Возникает вопрос: можно ли на этом основании сделать вывод,что в оппозиции мужская/женская речь первый член является немаркированным,а второй маркированным? Ответ на этот вопрос может быть получен лишь при дальнейшем изучении специфических особенностей мужской и женской речи.

4.Противопоставление мужская/женская речь по-разному обнаруживается в разных сферах языка.Оно выражено наименее резко в кодифицированном литературном языке. В некодифицированных сферах языка это противопоставление проявляется более отчетливо. Сравнение литературной разговорной речи и городского просторечия показывает,что в просторечии некоторые специфические черты женской и мужской речи,выявленные по отношению к литературному языку, обнаруживаются более ярко.

5.Речевой механизм подвержен влиянию ряда социальных и психологических факторов,действующих под знаком "+" или "-" на проявление специфических особенностей мужской и женской речи.Иными словами,помимо социальных (см.п.2) и индивидуально-психологических черт говорящего значимыми оказываются тип ситуации общения,гомогенность/гетерогенность собеседников по признаку "пол",жанр речи,речевая интенция и др.

Литература

/I/ Алексеев А.А.Язык светских дам и развитие языковой нормы ХУШ в. - Функциональные и социальные разновидности русского литературного языка ХУШ в. Л.,1984.

/2/ Бернштам Т.А. Будни и праздники: поведение взрослых в русской крестьянской среде.- В кн.:Этнические стереотипы поведения.Л.,1985.

/3/ Вопросник по лексике современного русского языка.М.,1964.

/4/ Воронина С.Б. Экспериментально-фонетическое исследование явления аффрикатизации палатализованных [т'] , [д'] в современном русском литературном произношении.Автореф.канд.дис.М.,1984.

/5/Кодзасов С.В. Интонация вопросительных предложений: форма и функции.-Вкн.: Диалоговое взаимодействие и представление знаний.Новосибирск,1985.

/6/ Геодакян В.А.Генетико-экологическая трактовка латерализации мозга и половых различий.- Материалы Всесоюзн.конференции "Теория,методология и практика системных исследований".М.,1985.

/7/ Лендел Ж.Обращения,приветствия и прощания в речевом этикете современных венгров.- В кн.:Национально-культурная специфика речевого поведения.М.,1977.

/8/ Мартынюк А.П. Об отражении социальных ролей и психологических особенностей женщин в языке.-Вестник Харьковского ун-та, 1986,№290.

/9/ Николаева Т.М.Семантика акцентного выделения.М.,1982.

/I0/ Папп Ф.Паралингвистические факты. Этикет и язык.- В кн.:Новое в зарубежной лингвистике.Вып.ХУ.М.,1985.

/II/ Русская разговорная речь.М.,1973.

/I2/ Силина О.И.Роль социолингвистических факторов в формировании современной произносительной нормы.-В кн.:Лингвистика и модели речевого поведения.Л.,1984.

/13/ Eakins B.W., Eakins R.G. Sex differences in human communication. Boston,1978.

/14/ Philip M.Smith. Language,the sexes and society.Oxford.Basil Blackwell,1985.

/15/ Wierzbicka A. Different cultures, different languages,different speech actes (English vs Polish).- In: Journal of Pragmatics, 1985.

DIE VOKALE KOMMUNIKATIONSFÄHIGKEIT IM SYSTEM DER SPRECHSPRACHLICHEN KOMMUNIKATIONSFÄHIGKEIT

HANS-JÜRGEN BASTIAN

Sektion GKM, WB Sprechwissenschaft
Universität Greifswald
Greifswald, DDR, 2200

ABSTRACT

Die sprechsprachliche Kommunikationsfähigkeit umfaßt die linguale, die vokale und die paralinguale Kommunikationsfähigkeit. Ausgehend von den phoniatrischen Kriterien der Stimmgesundheit und der physiologischen Norm der Stimmfunktion, explizieren wir die vokale Kommunikationsfähigkeit als das sprachwissenschaftliche Kriterium der Stimmgesundheit. Der menschliche Stimmgebrauch ist ein biopsychosoziales Phänomen.

PHONIATRISCHE KRITERIEN DER STIMMGESUNDHEIT

Die organische und funktionelle Intaktheit der biologischen Schallquelle und im Verbund damit ein normales akustisches Feedback sowie eine normale taktil-kinaesthetische und propriozeptiv-vibratorische Rückkopplung der Sprechorgane sind notwendige Bedingungen für die vokale Kommunikationsfähigkeit. Die genannten Faktoren garantieren die optimale Unwandlung der Strömungsenergie des Luftstroms in Schallenergie und damit die uneingeschränkte Leistungsfähigkeit des Senders. Da die Lautbildung als Modifizierung einer Anregungsfunktion durch die Hohlräume des Ansatzrohres erfolgt, ist vorrangig nach der Leistungsfähigkeit des Glottisgenerators zu fragen, der den Anregungsklang zur Verfügung stellt.

Stimmgesundheit bzw. physiologische Norm des Stimmgebrauchs ist ein Bereich stimmlicher Leistungsfähigkeit, der durch folgende Kriterien charakterisiert ist: Anamnese, indirekte Laryngoskopie, Stroboskopie, auditiv-visuell-palpatorischer Stimmfunktionsstatus. Zentrales Bewertungskriterium ist der klare, allenfalls gering behauchte, resonanzreiche Stimmklang in alters- und geschlechtsadäquater, mittlerer Sprechstimmlage. Ausdauerfähigkeit der Stimme, Intonationsfähigkeit im allgemeinen und Lautstärkesteigerungsfähigkeit im besonderen sind nicht reduziert. Die Erholungszeit der Stimmfunktion ist nicht verlängert. Vereinzelte Funktionsfehler aus den leistungsbereichen Atmung, Einsatz und Ansatz dürfen nicht diagnostisch überbewertet werden. Einzelne Funktionsfehler konstituieren noch keine Dysphonie.

Dieses Bewertungsschema versagt zwangsläufig immer wieder bei den -häufig gutachterlich relevanten- Patienten, die bei relativ unauffälligem Stimmklang über mangelnde stimmliche Ausdauer klagen.

Hauptmangel dieses traditionellen Konzepts stimmlicher Leistungsbewertung ist das Fehlen eines exakten, praxisrelevanten Ausdauerbegriffs. So avanciert der Stimmklang zum ausschlaggebenden Bewertungskriterium der stimmlichen Leistungsbewertung, obwohl von ihm nicht auf die Ausdauerfähigkeit extrapoliert werden kann. Tatsächlich jedoch existieren zwei unterschiedliche Leistungsbereiche des Stimmgebrauchs der Art Homo sapiens, die sich in der Phylogenese entsprechend den Anforderungen der sprechsprachlich-kommunikativen Praxis entwickelt haben: Ausdauerfähigkeit und intonatorische Variabilität. Diese Leistungsbereiche unterscheiden sich fundamental hinsichtlich ihrer konstitutiven Parameter wie hinsichtlich ihrer biochemischen Anpassungsmechanismen an die verschiedenen Belastungsmuster bzw. Anforderungssituationen. Ihr sprecherzieherisches und gesangspädagogisches Training fordert dementsprechend spezifische Belastungsreize, je nachdem, ob die Vorgänge im Bereich der Muskelzelle oder -wie im Koloraturgesang- die zentralen Koordinationsvorgänge im Vordergrund stehen.

Ausdauer ist die Widerstandsfähigkeit gegenüber Ermüdung bei spezifischer Belastung. Gegenüber der intonatorischen Variabilität stellt die Ausdauer die Basisfähigkeit dar. Die Ausdauerfähigkeit schlägt als gesteigerte Belastbarkeit (optimierte Störungskompensation), als Vergrößerung der Reserven sowie als beschleunigte und vertiefte Erholung zu Buche. Das bedeutet eine Verminderung der Störanfällig- und eine Vergrößerung der Zuverlässigkeit. Unter dem Ausdaueraspekt unterliegt die Funktion ausschließlich der physiologischen Ermüdung. Ermüdungsvorgänge sind in der Regel spätestens nach 24 Std. abgeklungen. Die Ausdauerfähigkeit basiert auf den Elementen des oxydativen Stoffwechsels. Ihr Niveau wird vor allem von der Funktionstüchtigkeit des Glottiswiderstandes kardio-pulmonalen

Systems sowie von psychischen Faktoren wie Motivation, Einstellung, Wille, Kommunikationsstrategie bestimmt. Eine wichtige Rolle spielt dabei die Ökonomisierung der Organfunktionen. Die Ausdauerfähigkeit ist die Grundlage für die intonatorische Variabilität. Sofern die intonatorische Variabilität die Intensität betrifft, basiert sie auf einer Steigerung der glykolytischen Kapazität und einer Zunahme der Muskelmasse, der sog. Arbeitshypertrophie /1, 2/.

ZUR PHYSIOLOGISCHEN NORM DER STIMMFUNKTION

Die physiologische Stimmfunktion, die sog. Orthophonie, ist bis heute nicht eindeutig auf der Grundlage der objektiven Funktionsparameter des Stimmapparates definiert. In der Diagnostischen Praxis der Phoniatrie gibt es daher in der Beurteilung der Stimmgesundheit einen subjektiven Ermessensbereich. Stimmgesundheit geht, wie Gesundheit überhaupt, fließend vom Physiologischen ins Pathologische über.

DER MENSCHLICHE STIMMGEBRAUCH ALS BIOPSYCHOSOZIALES PHÄNOMEN

Die menschliche Stimme ist ein biopsychosoziales Phänomen. Nicht nur die biologischen Leistungsvoraussetzungen einschließlich der leistungsbegrenzenden organismischen Faktoren - man denke an die konstitutionell "kleine" Stimme oder an morphologische Organminderwertigkeiten des Kehlkopfes in Form von Asymmetrien /3/ - determinieren die menschliche Stimme, sondern auch die situativen Anforderungen an die Sprech- und an die Singstimme sind gruppenspezifisch normiert und damit determiniert. Die sprechsprachliche Kommunikation hat nicht nur natürlich-materielle, morphologisch-organische, sondern auch gesellschaftliche Bedingungen. Die menschliche Stimme ist nicht nur eine biologische Funktion. Ihr Erscheinungsbild ist immer sozial und kulturell überformt. Das gilt bereits für solche grundlegenden Sachverhalte wie die Artikulationsbasis oder den Lautstärkepegel in der konkreten Kommunikationssituation, der wesentlich von der räumlichen Distanz der Kommunikationspartner (Kanallänge!) und damit auch von sozialen Faktoren sowie vom ubiquitären Lärmpegel (Industriegesellschaft!) bestimmt wird. Überdies ist die Stimme immer mit der Artikulation gekoppelt, so daß über die situationsgerechte und damit normadäquate Lautungsstufe (Artikulationspräzision! Sprechspannung!) situationsrelevante Determinanten soziokultureller Art in den Kommunikationsprozeß eingehen.

VOKALE KOMMUNIKATIONSFÄHIGKEIT - DAS SPRECHWISSENSCHAFTLICHE KRITERIUM DER STIMMGESUNDHEIT

Die vokale Kommunikationsfähigkeit ist Bestandteil der individuellen Leistungsdispositionen des Individuums. Das Individuum ist ein gesellschaftliches Wesen. Damit sind körperliche Leistungen zugleich soziale Phänomene. Diese haben nicht nur eine Gesellschafts-, sondern auch eine Naturgeschichte.
Die vielen heiseren Rock-, Pop- und Beatsänger repräsentieren spezifische Stimmoden. Auf Grund unterschiedlicher, speziell gruppenspezifischer Wertsysteme kann die Kategorie "schöne Stimme" sowohl Gesundes als auch Pathologisches beinhalten. Stimmbildungsideale sind immer auf ihre soziokulturelle Gruppenspezifik hin zu befragen.
Vokale Kommunikationsfähigkeit des Sprechers ist die -lern- und trainingsabhängige-Fähigkeit, entsprechend den Normen des Stimmgebrauchs seiner Sprechgemeinschaft im allgemeinen (Registerwahl! Mittlere Sprechstimmlage! Oralität bzw. Nasalität! Stimmeinsätze! Intonation!) und seiner sozialen Schicht und Gruppe im besonderen die Variablen Stimmklang,Klangfarbe sowie melodische, dynamische und rhythmische Akzentuierung einschl. Pausengestaltung im komplexhaften Zusammenwirken mit der lingualen und der paralingualen Kommunikationsfähigkeit ausdauernd zu produzieren und auf diese Weise

- den lingualen (semantisch-denotativen) Nachrichtengehalt entsprechend den Erfordernissen der Kommunikationssituation und damit des Kommunikationsgegenstandes und der Kommunikationsintention
- sowie die vom Sprecher situativ intendierte Menge an nichtlingualen (ektosemantischen, konnotativen) Informationen zu übermitteln und so die vokale Selbstdarstellung des Sprechers und die Sprecheridentifikation zu ermöglichen
- sowie die Monosemierung der sprechsprachlichen Zeichen zu bewirken, und zwar im Verbund mit Kontext (sprachlichem Zusammenhang) und Vorwissen des Hörers,

so daß sprecherseitig die Realisierung der Kommunikationsintention gewährleistet ist, d.h., mit stimmlichen Mitteln gefördert wird oder zumindest nicht durch sprecherseitig unbeabsichtigte, emotional negative Nebenwirkungen gestört wird. Bei Vorliegen dieser Bedingungen ist eine Stimmfunktion als kommunikativ zu klassifizieren. Kommunikativität nennen wir die Fähigkeit einer Stimme zur Realisierung der hier explizierten sprechsprachlich-kommunikativen Notwendigkeiten. Kommunikative und physiologische Stimmfunktion sind identisch. Die gute Stimme im absoluten Sinne gibt es nicht. Eine Stimme ist immer nur gut in bezug auf ihre Leistungsfähigkeit in konkreten, gruppenspezifischen Kommunikationssituationen.

REFERENCES

/1/ D.G.R.Findeisen, P.G.Linke, L.Pickenhain (Eds.), "Grundlagen der Sportmedizin für Studenten, Sportlehrer", 2. Aufl., Leipzig, 1980.

/2/ Sektion Sportwissenschaft (Ed.), "Sportmedizinische Grundlagen sportlichen Trainings", Leipzig, 1967.

/3/ E.Unger, H.-J.Bastian, "Phoniatrische Kriterien der Tauglichkeit von Studienbewerbern". Dt. Gesundh.-Wesen 31: 2000 - 2003 (1976).

DEVELOPMENT OF PHONOLOGICAL OPPOSITION VOWEL/CONSONANT WITH NORMAL CHILDREN AND CHILDREN WITH ANARTHRIA

NATALYA LEPSKAYA

Dept. Of Applied Linguistics
Moscow University
Moscow, USSR, 119899
ГСП-3, B-234

TATYANA BAZZHINA

Dept. of Applied Linguistics
Moscow University
Moscow, USSR, 119899
ГСП-3, B-234

ABSTRACT

The difference in the development of speech sounds with normal children and children with anarthria is revealed at the babbling stage. In this period normal children's vocalizations exhibit "syllable-likes" in which we can find segments with max-contrast and min-contrast between vocal and consonant elements.

By contrast to a normal child a child with anarthria is capable of producing only mid-contrast segments. This proves that mid-contrast syllable-likes are stipulated by the functioning of the speech mechanism, but production of max-contrast units is the major requisition for establishing the first phonological opposition: vowel/consonant.

The phonological system of the patient is destroyed on the level of speech production but is kept intact on the level of speech perception. This means that there are the two systems of distinctive features: one of them is connected with speech production while the other - with speech perception.

INTRODUCTION

Most authors writing on language acquisition and analysing the pre-speech stage of normal speech development discuss their data in such linguistic terms as phonemes (vowel and consonant), prosodic features etc. |1|. It seems to be more correct if we analyse these facts in terms of "syllable-like", "vowel-like", "consonant-like", because such vocalizations neither motorically, nor functionally are speech sounds and even less so-phonemes. Vowel-likes and consonant-likes are examined here within the syllable-likes, since, according to N.I. Žinkin's data, it is the syllable which is actually the unit of speech production |2|, and at the pre-speech stage it is respectively, the syllable-like.

The purpose of this investigation is to compare data of normal development (pre-speech stage) and data of anarthria since this comparison may be helpful in revealing some typological features of language as such and, in particular, they can throw light on the process, preceding formation of phonological oppositions.

THE METHOD OF INVESTIGATION AND THE MATERIAL

For the purpose of this investigation the pre-speech stage has been divided into some sub-stages, i.e.: crying (0-0.2); cooing (0.2-0.4); pre-babbling (0.4-0.6) and babbling as such (0.6-1.1). We have studied vocalizations of 38 normal babies from 0.3-0.10 and vocalizations of one child with anarthria who was 7 years old. All these vocalizations have been recorded and then treated by the oscillograph and separator. The majority of the normal children remained under observation for a period of some months; others were observed at certain points of their life (for instance, at the age of 3 or 8 months etc.). The patient with anarthria was under close observation for more than a year. For the purpose of the our investigation of the patient's vocalizations we worked out some special experiments. We asked the patient to analyse the sound structure of words like |pȃpa|, |tȃta|, |t'ȍt'a|, |hudȍi|, |bal'sȍi| using alphabet; to repeat definite types of the syllables given by the examiner |pa|, |p'i|, |n'e|, |du|, |bo|, |n'i|, |mu|, |o|, |u|, |i|, |a|, |ma|. The patient's vocalizations when he was alone with his toys were also recorded.

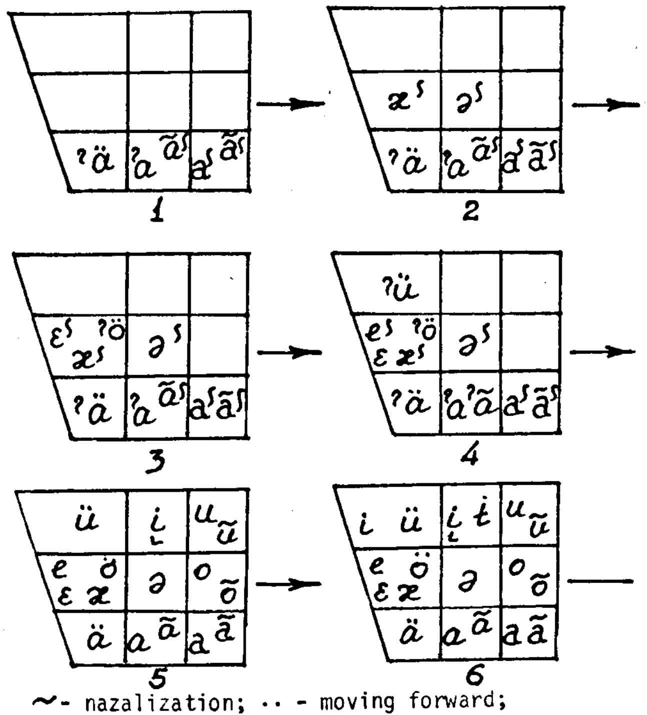
In his case anarthria appeared as a result of the birth trauma. The patient's central nervous system abnormality implicates the basic mechanism of speech synergism. The neurological and psychological examinations showed that the child had an inborn disability of coordinating the muscles of the vocal tract and of producing intelligible speech |3|. His pronunciation was similar vocalizations of children at the cooing and pre-babbling stages. But at the same time he had normal hearing and could understand spoken language but there could be no question of his understanding Russian completely. He could cry and laugh, and it sounded normal. He was able to make short coughlike grunts to accompany his pantomimed communications. His cognitive and communicative activity was very high, but his general knowledge is below the norm for his age.

THE RESULTS AND THEIR DISCUSSION

Vowel-likes inside syllable-likes |V|. Our material allows us to describe the acquisition of vowel-likes at the pre-speech stage in normal and pathological development (see Table 1). Table 1 shows that the first to appear are vocalizations producing the impression of mid vowel-likes of non-high. From the point of view of articulation these vocalizations are the simplest: the mouth opens widely, the position of the tongue is neutral, the muscular strain is minimal, the vocal bands are weakening.

Table 1

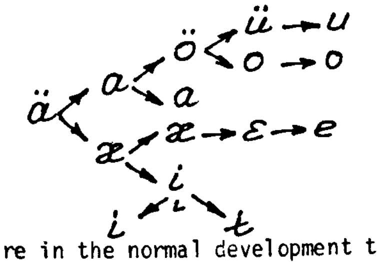
The Normal Development of Vowel-Likes at the Pre-Speech Stage (from cooing to babbling).



~- nazalization; ·· - moving forward;
ʟ - moving backward.

In the course of child's development the speech organs are perfected and this gives him an opportunity to produce not only mid vowel-likes but sounds which are more narrow and front, like |ε|. The intensity of such vocalizations is smaller, but the muscular strain is greater. We also observed the tendency to smooth the marginal positions of the tongue. For instance, in the development of |u| and |o| we observed the movement of the tongue forward to |ü| and |ö|; in the development of |i| and |e| the tendency of drawing the tonque off and down and the appearance the vowel-like |ä|.

In the development of phonetic field of vowels one can observe the gradual differentiation of vowel-likes and the appearance of the connection between the production of sounds and its acoustic form.

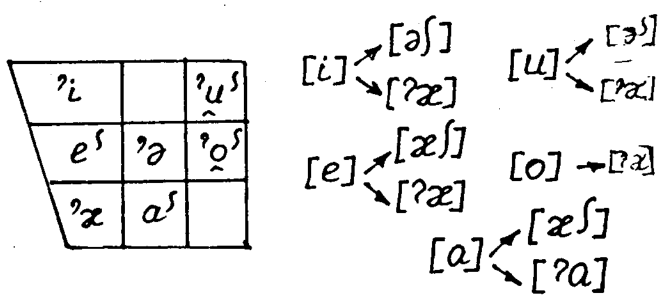


Therefore in the normal development the acquiring of vowels is the process of detalization and differentiation of the initial mid vocalizations.

In the patient's vocalizations some other types of vowel-likes may be found (see Table 2).

Table 2

The System of the Patient's Vocalizations



a) Patient's babbling; b) The repetition of the syllables |V| given by the examiner.

Table 2 shows that babbling vocalizations have a greater variety than vocalizations in his repetition. It means that the patient has some difficulties in bringing his voicing mechanism under voluntary control.

In the repetition there are substitutions of vowels, and the general tendency is to produce mid vowel-likes without any differentiations. Labialized vowels like |o|, |u| were substituted by mid partly labialized |æ̹|, |ə̹|. Such acoustic features as loudness, timbre and duration were not stable. The most difficult task for the patient was to produce vowel-likes with high F_1, which need for their articulation efficient differentiations.

Analysing Tables 1 and 2 we found out that in vocalizations of normal children at the cooing and pre-babbling stages and in the vocalizations of the patient vowel-like sounds are accompanied by noisy on and off glides, glottal stop |ʔ| or voiceless indistinct mid sound |ʃ|.

Consonant-likes inside syllable-likes |CV|. The first consonant-likes appear in normal development at the cooing stage. We can point out in children's sounds as well as in the patient's vocalizations the presence of partly voiced and moderatly palatalized consonant-likes. In addition in the patient repetition of syllables |ta|, |t'a|, |b'i|, |n'u|, |po| we see the regular substitutions of the first consonant component with |w| or |j|. Most of consonant-likes as well as vowel-likes receive additional nasal articulation, and are accompanied by the vocal on and off glides |ʃ̃ə|.

All these facts can be explained from the physiological point of view: the epiglottis is high, pharyngal modulations are minimal |4|. For children it is impossible to mantain a fixed position of their speech organs, and as a result their articulation is gliding. In normal children's vocalizations in contrast to our patient, there are many consonant-likes - they greatly exceed those in the speech of adults, surrounding the baby. At this period in vocalizations of Russian children, for instance, it is possible to find sounds like clicks. Normal children vocalizations have no connection with adult speech. This is the so-called pre-phonemic level.

In normal and pathological vocalizations max- and min-contrast syllable-likes are absent because consonantal elements are accompanied by the vocal on and off glides and vocal elements - by the noisy glottal stop or the voiceless indistinct sound. This results in the increased sonority in the first case and reduced sonority in the second [5].

$$C^{v} \longleftrightarrow {}^{c}V$$

It determines the absence of coarticulation between consonantal and vocal elements inside such syllable-likes. The appearance of max-contrast syllable-likes is impossible.

There is a similarity in vocal-consonantal vocalizations between normal babies at the cooing and pre-babbling stages and the patient with anarthria.

The divergence in the acquisition of vowels and consonants in normal and pathological development begins at the babbling stage.

At this stage in normal acquisition the epiglottis is descending. This is the physiological requisition for the articulatory oppositions of sounds.

Changes of the speech mechanism and its connection with perception of adult's speech (echolalia) are the basis for the formation of phonological oppositions as such. In normal development in contrast to anarthria we can observe the tendency in vowel and cosonant-likes of loosing their noisy and vocal on and off glides, glottal stop. The articulation becomes even more differentiated. As a result in normal development max-contrast syllables like |*pa*|, |*ta*| appear. Therefore the presence of max-contrast syllable-likes in babies vocalizations is the major requisition for the opposition of sounds according to degrees of sonority, when on the one hand there are wide non-high vowels like |*a*|, but on the other one there are voiceless stop consonants like |*p*|, |*t*|. This is a manifestation of the first general phonological opposition: vowel/consonant.

This opposition is the earliest in child development and is a universal one since according to R. Jakobson, it is observed in all the languages of the world [6].

This opposition is absent in the patient's speech production but is present in his speech perception. At the end of the pre-babbling stage in normal development it is possible to distinguish sounds according to the types of resonators (mouth resonator - nose cavity). As a result, we can find oral and nasal vowel- and consonant-likes. This distinction in the resonator's types is the physiological requisition for the forming at the babbling stage of the phonological opposition: oral/nasal.

Then babies begin to split both consonantal and vocal components and other differentiations oppositions also appear.

CONCLUSION

At the cooing and pre-babbling stages in normal development and with our patient we find vocalizations in which features of articulation contrasts are mixed up. As a result, the appearance of max-contrast syllables is impossible.

The same was established by N.I. Žinkin in the sound system of hamadryads. He pointed out that in their vocalizations combinations of a vocal element with a noisy consonant-likes do not occur; only combinations of a vocal element with a sonant-like are possible [4].

The perfection of the speech mechanism and its connection with the children's perception of adult speech brings forth the appearance of max-contrast syllables, which in its turn stimulates the formation of the first general opposition: vowel/consonant. Various other oppositions modifying and attenuating the primary contrast of consonant and vowel follow.

The dominating influence of adults' speech on the acquisition of the phonological oppositions is proved by the presence of such oppositions in the patients' speech perception, but their absence in his speech production. This fact shows that until a certain moment the absence of speech production skills doesn't interfere with a more or less adequate understanding of speech.

These results may be used for patients' rehabilitation and in language teaching.

REFERENCES

[1] Тонкова-Ямпольская Р.В. Развитие речевой интонации у детей первых двух лет жизни. - Вопросы психологии, 1968, № 3.

[2] Жинкин Н.И. Новые данные о работе речедвигательного анализатора в его связи со слуховым. "Известия АПН РСФСР", вып. 81, М., 1956.

[3] Панченко И.И. Дизартрические и анартрические расстройства речи у детей с церебральными параличами и особенности логопедической работы с ними. АКД, М., 1974.

[4] Жинкин Н.И. Звуковая коммуникативная система обезьян. "Известия АПН РСФСР", вып. 113, М., 1960.

[5] Носиков С.М. Опыт фонетического описания лепета (организация слога и ритмической структуры). - В кн.: Становление речи и усвоение языка ребёнком. М., 1985.

[6] Jakobson R. and Halle M. Phonology in Relation to Phonetics. In: Manual of Phonetics, Ed. by L. Kaiser, Amsterdam, 1957.

SPEECH DISTURBANCES CAUSED BY CLEFT PALATES
(Phoniatric aspects)

IRAIDA KRUSHEVSKAYA

Dept. of Oto-Rhino-Laryngology
Research Institute of Capacity to Work
Minsk, Byelorussia,USSR,22008I

Palate developmental defects result in voice and speech disturbances due to: a) incomplete closure of the throat ring; b) disturbances of the resonator function of the mouth cavity. In spite of an obvious theoretical value this problem has an urgent practical aim of restoring speech communication of the cleft palate carriers, and their social and labour rehabilitation.

With the modern rates of development of social,political and scientific life, the actuality and social significance of the problem of restoring the lost communicative functions have grown greatly. Hearing , speech and voice disturbances should be looked at from the standpoint of pathophysiology of these organs, which makes it possible to develop a more rational system of measures to restore these functions.

Clinical and social observations indicate that the restoration of speaking and vocal functions of the cleft palate carriers is a compley process of rehabilitation, and it is insufficient to make one operation to create a plastic resonator. The analysis of the results of the investigations, which we have carried out, shows that the presence of pathophysiological and conditioned-reflex relations in the central nervous system of the cleft palate carriers before the operation has caused the absence of acquired reflexes of correct phonational respiration, of the voice formation process, and have resulted in disturbances of the neurophysiological speech mechanisms.

According to the statistic data, cleft palates are a frequent occurence: I.5 - 2 cases per IOOO new-born children. Face and jaw developmental defects may be caused by various exogenous and endogenous factors affecting the fetus at the early stage of its development before 7-9 weeks. Cleft lips and palates are one of the most serious psychotraumatizing defects since the early childhood, as they create a feeling of inferiority of their carriers.

The representatives of phonetic sciences will certainly get interested in and find useful the submitted results of investigations of a live model of an anatomic defect of the mouth cavity resonator with all the disturbances, which follow, including the muscular system function of the loud speech motor apparatus:breathing, phonation and articulation muscles. In this pathology a hearing disturbance aggravates the influence upon the phonetic system of speech. The most characteristic feature of a speech disturbance at cleft palates is rhinolalia sperta: nasalization, which has appeared due to the absence of a demarcation between the nasal and mouth cavities,changes greatly the acoustic characteristics of phonemes. A voice disturbance is versatile. The most prominent features are timber alterations, the presence of an unpleasant nasal resonance, a clear nasal shade of oral sounds. The nasal sounds (M,H) are pronounced quite normally. The sounding of vowels changes insignificantly. Rhinophonia may be accompanied ny rhinolalia, i.e. incorrect pronunciation and distortion of sounds in the following cases: I) if the acquired factors, developed due to a cleft palate, begin to make its influence during the first years of a child's life when the articulation mechanisms have not yet been formed; 2) if an articulation disturbance of the central origin joins; if a hearing disturbance (even of short duration), causing the formation of wrong articulation reflexes, joins during the articulation formation period. Palate developmental defects result in voice and speech disturbances due to: a) incomplete closure of the throat ring; b) disturbances of the resonator function of the mouth cavity; c) accompanied hearing disturbances.

At the absence of voice caused by deformations in the mouth cavity and incomplete closure of the throat ring,functional derangements are observed in all resonator cavities.

Pathological changes of the sort palate muscles usually develop at the age of 4 - 5. Due to a lower functional load in the muscles and mucous pharynx, a dystrophic process grows progressively worse. The mucous membrane of the back wall of the pharynx becomes gradually pale, atrophic. The absence of a pharyngeal reflex is indicative of the atrophy of muscular fibers of the pharynx constrictor, and of degenerative changes of the sensitive and trophic nerve fibers of this region.

The chronaximetry data (time necessary for the muscles to react to an electric stimulus) testify to a significant disturbance of the muscle function of the closing throat ring expressed by increased chronaxy of these muscles from 0.32 to 0.40 mm/sec. Eventually chronaximetric asymmetries appear between the right- and leftside muscles, if the clefts have not been operated on. The upper pharynx constrictor, whose chronaxy becomes longer and longer, is subject to much deeper dystrophic and functional changes, and then the muscle ceases reacting to an electric stimulus. In cases of disturbances of the closing throat ring function, the speech becomes monotonous without any melody or accent.

Investigations of the external respiration function have revealed a versatile respiratory impairment. At congenital clefts the phonational respiration suffers most of all: at phonation children and teen-agers continue to breathe simultaneously through the nose and mouth at the exclusively clavicular breathing.During the process of expiration a large amount of air (from 20 to 32 per cent) escapes through the nose, thus shortening the time of expiration, lowering the air pressure in the suprafold space.

Hence, the phonational respiration becomes shallow and hurried. From the age of 7 - 8 a functional derangement of the motor muscles and of the diaphragm in particular is revealed: the function of these muscles becomes weaker, their contractions are flabby, slow. Very often they are asymmetric and not co-ordinated with the phonation and articulation. The time and degree of expressiveness of the above-mentioned pathology depend upon the cleft morphology with regard for the defect width. Such patients have a low, constrained, weak and thin voice with a vivid nasal shade. Acoustic changes of the voice spectrum deprive it of clarity and make the speech less legible.

A change of the voice timbre of the cleft palate carriers is connected with an anatomic defect of the supratracheal pipe, which results in construction asymmetries of the resonator cavities of the larynx, pharynx, nose as well as discoordinates the function of the palate-larynx complex, in which the palate plays the role of a starting mechanism. At palate clefts the phonation mechanism is so specific that at rhinolalia the voice is singled out as a separate disturbance and is called "palateny dysphonia" or "palatophonia".

The combination of an anatomic defect of the palate, laryngeal sound formation,motor disfunction with an incorrect voice behaviour provokes the development of organic changes in the larynx of the type of nodulations and chronic inflammatory processes, motor - as a cut of internal muscles of the larynx, functional - as phonasthenia.

Violations of the integrity, anatomical and functional asymmetries of the soft palate and pharynx muscles bring with age to a functional asymmetry of the vocal folds, which is well determined with characteristic asynchronism of the vocal folds vibration at the electronic laryngostroboscopy. This pathology of the functional state of the internal muscles of the larynx as well as the asymmetry of forms of the larynx resonator cavities are clearly expressed since the age of 9 - II.

There are three main reasons for the voice pathology:

I. Additional articulation function of the larynx. The laryngeal way of forming a number of voiced consonants, their sounding by the friction of air along the edges of the vocal folds result in a functional overload of the vocal apparatus and a growth of organic motor or functional diseases of the vocal folds.

2. The cleft palate carriers have a low voice because since their childhood they consider themselves to be inferior members of our society, are ashamed of their face malformation and speech defects, and don't want to attract the attention of those who surround them.

3. The muscles, lifting and stretching the palate, work as antagonists instead of being synergisms, their functional load becomes lower and the dystrophic process worsens.

At cleft palates the speech develops under pathological conditions, and so it suffers more than other functions. The absence of a palatopharyngeal closure makes the nasal cavity a double resonator of the mouth cavity giving a nasal timbre to all phonemes. The degree of the sound nasalization expressiveness depends on the inadequacy of closure, the mobility of the palate curtain and the co-ordination of the tongue and soft palate motions. Due to the escape of air into the nose, the pressure falls sharply and it becomes impossible to sound the apertures (closure breakage) during the articulation of consonant phonemes. Besides,the escape of air into the nose makes it more difficult to form a directed air flow in the mouth, and as a result almost all the plosive and fricative voiceless consonants are pronounced in a pharyngeal way. The mediolingual palatal and backlingual palatal sounds cannot be articulated because of the absence of one of the closure components - palate. The forelingual T,T^I,D,D^I become weaker or are replaced with a laryngeal or pharyngeal closure on H, H^I.

All the latest results of the pathophysiological investigations, which have revealed detailed peculiarities of the phonational respiration, voice and speech formation at rhinophonia and rhinolalia, have been assumed as a basis of methodical recommendations developed in our country by I.I.Yermakova to correct the speech of children and teen-agers at rhinolalia. The author has taken into account that no spontaneous speech occurs after uranoplasty, but the pathological sound formation at rhinolalia has anthropophonic (sound distortion) and phonologic (replacement of one phoneme with another) signs. The correction of each sound provides the following: I) an ability to single it out from others; 2) to correlate it with some definite articulation; 3) to correctly pronounce the articuleme; 4) to use this ability in a flow of connected speech.

In spite of an obvious theoretical value this problem has an urgent practical aim of restoring speech communication of the cleft palate carriers, and their social and labour rehabilitation.

Verbal development dysontogenesis in children with velopharyngeal incompetency

G.V.Chirkina

Defectology Research Institute, Moscow, USSR

Children with velopharyngial incompetency make up one of the most severe forms of speech pathology. Linguistic and psychological-pedagogical study of the defect suggests development of advanced correction methods. Interactions of articulation and receptive mechanisms in verbal activity are described.

Children with velopharyngeal incompetency resulting from cleft lip and palate (speech therapy diagnosis: rhinolalia) belong to one of the most severe forms of speech pathology. We undertook investigation of the defect structure in linguistic and psychological-pedagogical aspects in the frames of the system approach. This study showed that the children with two given disorders do not make up a homogenious group. Most common characteristics of speech deficiency in this case are found in acquisition of phonetics which is developing in abnormal anatomic physiological conditions.

Children with rhinolalia are characterized by changes in oral sensitivity in mouth cavity as well as impairments in stereognosis caused by sensory-motor conduction tract dysfunction (what results from deficiency of feeding in infant age). These children were found to have certain specific characteristic in prelingual developmant, insufficient activity of babbling, late appearance of speech, long laps between appearance of the first words and phrase speech.

Peripheral defects of articulation organs result in development of compensatory changes of articulation organs positions when sounds are pronounced: high position of the tongue root and its backward shift in the oral cavity, lips insufficiency in labial vowels, bilabial and labio-dental consonants, excessive activity of the tongue root and larinx, tension in mimic muscles. The most essential defect of oral speech phonetics is that of impairment of all this oral sounds, resulting from changes in aerodynamic conditions of phonation and involvement of nasal resonator.

Besides regular nasalization children with rhinolalia are characterized by some specifically coloured consonants (often velar ones): what is the effect of participation of pharyngeal resonator. Pharyngealization, i.e. excessive articulation resulting from tension in the walls of pharynx, appears as a compensatory means. There are also additional articulations in larinx what furnishes the speech of rhinolalics with a specific "clicking" on-glide.

Besides these mentioned tendencies in adaptive changes of speech, there are found many more particular articulatory defects. The làtter depend greatly on positional changes in a word, phrase, text. The most typical are:

1. Omission of initial consonants
2. Neutralization in the manner of production
3. Multiple various substitutions of sounds
4. Abrupt discontinuance of sounding (fricatives in the final position)
5. Pronounciation of hushing sounds is accomponied by hissing noise and v.v.
6. Sonorous sounds in the final position are strongly devecalized
7. Manner of sound production is changed: explosives are substituted by fricatives
8. Vibrant /r/ is either missing or substituted by the second /i/ in strong breath
9. Additional noise in nasalized sounds (hushing, hissing, aspiration, hoarseness, laringeal on-glide etc.)
10. Backward shift of articulation focus (as a result of high position of the tongue root and insufficient participation of lips in articulation)
11. Children having regular lessons with speech therapy teacher are sometimes characterized by hyper correction phenomena, i.e. forward shift of articulation. E.g. /s/ (frontal dorsal) is substituted by /f/ (labio-dental) without changing the manner of articulation.

Interconnections between nasalization and distortions in separate sounds articulation are rather multifold.

It's impossible to establish an immediate correlation between the degree and form of palate defect and extent of phonetic impairment. Compensatory modes children use for speech production are too variouse, very much also depends on relations among resonators and on diversity of individual differences in the configuration of mouth and nasal cavities. Besides that there are other less specific factors also influencing the degree of distinctness (developmental, individual-psychological characteristics, social-psychological factors and many others).

The described characteristics of phonetics in children with rhinolalia suggest the conclusion about "phonetic uncertainly" of speech sounds and developmental backwardness of prosodic elements.

Speech legibility varies from 28,4% to 55,6%. It brings around serious bounds over speech as a means of communication. Disorders of acoustic aspects may be grouped in the following way:

1. Disorders directly relating to anatomy defects
   a) articulatory disorders
   b) aerodynamic disorders
   c) phonatory disorders
2. Disorders related to motor control defects
   a) eurhithmic-sylabic disorders
   b) disorders in consonant confluence

The described characteristics of pronounciation in children with rhinolalia result in disappearance of distinctive features and delayed or distorted phonological development. The functions of distinction and identification of language sounds are disturbed, what impaires phonological aspect of acoustic functional system. Disorder of interaction between auditory and speech motor analysator affects acquisition of written speech. In writing substitutions are found: /m/ for /b,p/, /n/ for /t,d/ and v.v. What is due to absence of the oppositions in oral speech. There are also other types of substitutes of vowels, hushing and hissing, voiced and surd sounds, what proves disorder of the whole phonematic system.

The degree of writing disorder is defined by combination of factors: defect of articulatory system, character and terms of speech therapy, compensatory capablities of a child, influence of verbal environment. The children need specially organized correction in disgraphia performed simulteneously with modification of child's phonological system. These data were taken into consideration of the reform in principles of organization of verbal material, used for correction goals.

Study in other aspects of verbel activity of children with rhinolalia of different age groups revealed a certain dynamics in interaction of pathogenic factors, differing in its nature, degree and turn of influence. In preverbal and early verbal period the greatest negative effect is produced by anatomic-physiological defects influencing development of phonetics (I stage). In the period of active development of verbal activity deficient conditions of speech generation, deprivation of motor component of speech trigger psycholinguistic factors. They cause diversity of deficiencies in speech generation and perception (II stage). On the III stage when the language system has to be acquired social-psychological factors are added which hinder communication and information exchange (education). Use of such a model supplies a speech therapist which a means for defining correction strategy and ways for prophilactics of secondary aftereffects of the defect.

A DEVICE FOR CORRECTION OF RHYTHMICAL DISORDERS OF SPEECH FUNCTIONS

Y.N.GNATIV, Y.M.RASHKEVICH, Z.Y.SHPAK

Department of Automatics
Lvov Polytechnic Institute
Lvov, Ukraine, USSR, 290013

ABSTRACT

The absence of effective approaches to removing human stammering makes it necessary to develop some devices and techniques of raising the efficiency of medical treatment of speech disorders. A device is presented for the implementation of basic approaches used in logopaedics to produce stable rhythmical patterns and rates of speech processes of logoneurosis patients. The performance of the device is based on the introduction of changes into the time and frequency parameters of the speech signal.

INTRODUCTION

The disorders of the speech functions of human beings are considered serious and common diseases.

The accepted view of stammering is as of a stable pathological state of speech [1,2]. Medical treatment of stammering involves a number of procedures, which are designed to destroy stable pathological states of stammering and to create new functional relations, corresponding to healthy speech processes [3]. Here we may include light and sound effects, selection of word pronunciation speed, speech delay, change of qualitative characteristics of speech, etc. According to [2,3,4] an integrated application of these techniques depending on the individual peculiarities of stammering is rather effective.

The above speech treatment procedures require some dedicated technical system.

At present the logopaedic treatment makes use of separate devices for time delay of speech, producing periodical sound signals, etc. So far we haven't come across any mention in logopaedic literature of a technical device for increasing or reducing the length of speech pronunciations, in spite of the fact that "slow" speech is considered a classical method of speech treatment [4,5].

The creation of an integrated logopaedic device capable of producing a desired speed of speech reproduction is considered to be the basic requirement in speech therapy.

MAIN FEATURES OF THE DEVICE

Our device is designed to implement following tasks:

- gradual slowing down or speeding up the reproduced speech while retaining its prosodic characteristics;
- introduction of controlled time delays into the output sound signals;
- introduction of additional rhythmical sound or light stimulation of the preset frequency and signal amplitude;
- muffling the speech with "white" noise;
- radical change of the voice quality;
- control of the volume of speech signals.

The device provides two modes of reproducing the speech: variation of the speech rate and time delay. Other corrective modes (rhythm, speech muffling, sound amplification) can be applied independently or in combination with first two ones.

When varying the tempo, the device reproduces a previously recorded text at a higher or lower rate, i.e. increases or reduces the length of sounding without losing the natural quality of the voice, legibility, the key and other speech values. The variation coefficient can be set within the limits of 1-2.5 with the discreteness of 0.1.

If a speech signal is fed into the device in the real time scale, i.e. directly from the speaker, the tempo control produces the displacement of spectral composition of the speech proportional to the variation coefficient. The output speech signal is reproduced at the original speed, but the quality of the voice changes, i.e. an "alien" voice is heard.

The mode of delayed speech enables us to obtain the input speech information at the output of the device with the delay of 100-250 ms (the step being 10 ms). This covers the most favorable range of delays [2] used for the corrective treatment of stammering in patients of varying degree of the disease. An additional amplification of the output sound signals increases the effectiveness of the correcting procedures.

When producing the rhythmical effect the device creates periodical sound and light signals of excitation. The sound rhythms are in the frequency range 0.5-2.5 Hz. During photostimulation light flashes have the length of 70 mc with the rhythm of 2-5 Hz, which is quite suitable for developing necessary correct speech habits [3].

When producing background a casual noise signal is fed into the device output. It prevents the reception of patient's own voice, thus removing pathologically stable reactions to the incorrect speech [2]. To increase the achieved results the patients speech can be recorded and played back at a higher or lower speed for a better evaluation of the deviations.

THE DESIGN OF THE DEVICE

Fig.1 shows the functional layout of the device with the external sound recording and sound reproducing units. When the device is in the process of regulating the speech rate, the input signal is fed into it from the recorder, the latter being controlled by a special circuit according to the preset variation coefficient. Having been converted to a digital code, this signal is recorded by the two memory blocks connected in parallel, each having the capacity of 2Kx10 bits. Every new counting is stored in the place of the oldest one. The value of the speech segment in the memory depends on the tempo vaviation coefficient k. The recording frequency is $f_1 = kf_2$, where f_2 - reproduction frequency with a constant value of 16 kHz. The readout from the two memory units is carried out simultaneously, but to different adresses. It

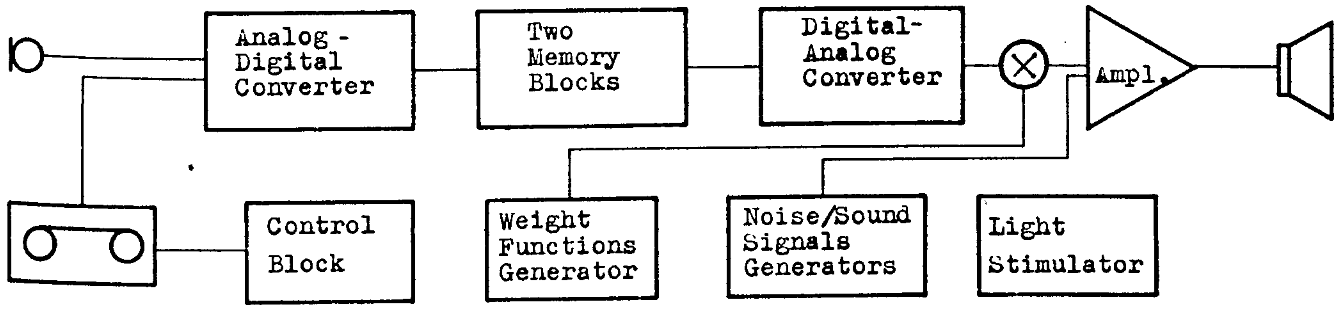

Analog - Digital Converter
Two Memory Blocks
Digital-Analog Converter
Ampl.
Control Block
Weight Functions Generator
Noise/Sound Signals Generators
Light Stimulator

Figure 1. Hardware Description of the Device

enables us to store all the input information in the output signal. The reproduced readings from the memory units are passed to the digital-analogue converter, where they are multiplied by the basic voltage, produced by the weight functions generators. It removes disconnections on the junctions of the speech segments in the output signals [6].

When the device works in the time delay mode, both of the memory units are connected in series. Digital readings of the speech signal received at the input of the device are stored in the memory with the frequency $f_1 = f_2 = 16$ kHz. The difference of the adresses of recording and reproducing is determined by a present time delay. The data obtained from the memory device are fed into the output channels of the system. The generators of sound and noise may be connected to these channels.

THE CONTROL OF THE SPEED OF SPEECH UTTERANCES

According to the phonetic theory of speech formation [7], most of the meaningful information of a speech signal is contained in the transient sections of sounds, the stationary segments being informationally poor. An experiment was conducted to determine the content percentage of various segments in speech utterances. The original speech signal was divided into sections of 20 ms, which equals the length of explosive sounds. On computing the degree of difference between the segments by the DELCO algorhythm [8], these sections were transformed into the output fields, the spectrum composition of which was in fact stationary. The results obtained show that with the normal speech rate the transient sections and short sounds take up about 25 percent of the whole time of the speech signal, the rest being filled with stationary pauses and long sounds. Speeding up or slowing down the tempo in oral speech it take place through the changes of the duration of the stationary sections. The short speech elements change slightly [9]. The idea of regulating the speech rate reproduction makes use of the abundance of speech signals. By excluding short sections of speech signals or by introducing additional short signals it is possible to reduce or increase the time of the phonation of speech utterances, at the same time saving the information content and individual peculiarities as well.

The regulatory process of the reproduction of speech information rate is shown in Fig.2. The original speech signal $x(t)$ is previously stored in some storage device (magnetic tape, digital memory,etc.) during the time of its pronunciation T_p at the rate V_p. The reproduction of the recorded information is carried out during the preset time interval T_r at a suitable rate of $V_r = k V_p$, where $k = T_p/T_r$ – the coefficient of the variation rate. Since it is accompanied by the change of the frequency constituents of the signal, the function of the speech rate regulator is to recover the original spectrum in the output signal $u(t)$ at the time interval T_r.

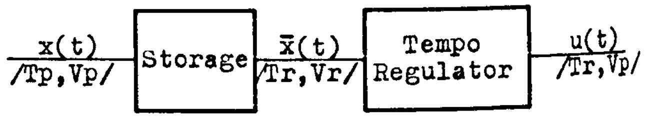


Figure 2. The Speech Rate Regulation Model.

We can single out two groups of methods in the regulation of speech tempo:

- the division of the signal into short uniform sections and changing their length proportionally to the preset rate regulation coefficient;
- the division of the original signal into quasistationary sections with the uniform spectral composition, followed by the variation of phonation time of each of them depending on its original length.

Among the methods of the first group is a selective segmentation of signals [10]. It is simple, easily operated, but has some significant shortcomings. It may remove short sounds and even whole syllables from the speech utterances, which causes distortion of speech and considerably restricts the regulating potential. In combining the segments after the length transformation, phase and amplitude drops occur at the junctions, reducing the quality of the obtained speech.

To remove the possibility of the loss of some useful sections of signals from the speech utterances we suggest using two channel regulation with partial overlapping of the neighbouring segments. To remove the amplitude drops, each of the output sections is multiplied by the weight "window". Partial imposition of segments compensates energy losses in weighing.

The analysis of speech types used for stammering correction shows that all kinds of stammering are characterized by the following regularities [6]:

- the length of every syllable is increased;
- the number of long syllables in a phrase is increased;
- the length of all syllables tends to be equal.

The first two types are positively solved in the presented device. The use of the device permits multiple reproduction of the recorded text with varying speed. On the other hand it allows to listen to an accelerated recording of the patients speech and to determine the intensity of the disorders. Other approaches to speech therapy are possible by varying the time of the sounding of speech information. To make the length of speech syllables equal requires more discriminating approach of the speech signals, which is characteristic of the second group of tempo regulation methods.

CONCLUSION

The performance of the device which permits to carry out a set of correcting logopaedic procedures is described. The most significant of them is the changing of the rate of speech reproduction. The best results in treating speech disoders are to be obtained by using regulator, which makes the length of stationary sections of sounds of equal duration. Modelling confirmed the effectiveness of these approaches allowing to create a wide range of tempo variations while retaining high quality characteristics of the reproduced speech.

REFERENCES

I. Заикание / Под ред. Н.А. Власовой, К.-П. Беккер // М.: Медицина. - I978. - 200 с.
2. Данилов И .В., Черепанов И.М. Патофизиология логоневрозов // Л.: Медицина. Ленингр. отд. - I970. - I59 с.
3. Андронова Л.З., Лохов М.И. Использование методов дестабилизации устойчивого патологического состояния в клинике и лечении заикания // Физиология человека. - I983. - Т. 9. - № 5. - С. 854-859.
4. Куршев В.А. Заикание // М.: Медицина. - I973. - I59 с.
5. Андронова Л.З., Арутюнян М.А. Анализ временных характеристик видов речи, применяемых при коррекции заикания // Дефектология. - I984. - № 4. - С. 34-37.
6. А.С. I173438 СССР. МКИ^3 G10L3/02. Устройство для изменения темпа речевой информации / С.В. Балицкий, Т.Н. Гнатив, В.В. Грицык, А.Ю. Луцык, Ю.М. Рашкевич. - Опубл. I5.08.85. Бюл. № 30.
7. Пирогов А.А. Вокодерная телефония. Методы и проблемы // М.: Связь. - I984. - 384 с.
8. Маркел Дж., Грей А. Линейное предсказание речи // М.: Связь. - I980. - 308 с.
9. Агафонова Л.С. и др. О некоторых характеристиках русской речи в зависимости от темпа произношения // Слух и речь в норме и патологии. Л.: Наука. - I977. - С. 25-39.
I0. Шиффман М. Регулятор темпа речи воспроизводящего магнитофона // Электроника. - I974. - № I7. - С. 24-35.

AN ARTICULATORY SPEECH SYNTHESIZER TALKING GERMAN

Gernot KUBIN, Vytautas PIKTURNA +)

Institut für Nachrichten- und Hochfrequenztechnik, TU Wien
Gusshausstrasse 25/389, A-1040 Vienna, Austria

ABSTRACT

Modern digital signal processing technology opens the way to real-time implementation of articulatory speech synthesizers as the phonetic-acoustic conversion module in text-to-speech systems. An outline of a workstation for the development of such a prototype synthesizer for the German language is given. This workstation is equipped with fast interactive graphics and acoustics processing capabilities and is used as a tool for both the study of articulatory phenomena as such and the development of simplified algorithms needed for the prospective target realization of the articulatory synthesizer.

1. INTRODUCTION

Until now most of the development in articulatory speech synthesis [1-6] has originated within a phonetic research environment. Only recently [7] attempts have been made to tailor this methodology into a form susceptible for real-time implementation with modern digital signal processing technology.

As articulatory synthesis is expected to yield high synthetic speech quality we have chosen this line for the development of the phonetic-acoustic conversion component within a text-to-speech system for German [8]. Our goal is not to refine the knowledge of human articulation in numerous experiments but rather to use the available knowledge for an operational speech prooduction model.

Therefore we confined the development environment to be small from the beginning: a specifically designed workstation that provides close-to-real-time operation of both the computer animated articulatory model and the acoustic signal synthesizer. The workstation facilitates changes in the detailed model and synthesizer structures while preserving the hard- and software characteristics of the envisaged target system.

2. SYSTEM OVERVIEW

In the text-to-speech system GRAPHON the articulatory synthesizer bridges the gap between the string of phonetic symbols derived from morphological word parsing [9] of the input text on one side and the synthesized acoustic speech signal on the other side. To this end, the articulatory synthesizer must provide the following four steps:

(1) Interpretation of phonetic symbols in the articatory domain by means of look-up tables containing geometry and timing parameters. Only ***essential*** or non redundant parameters are used for the definition of a phone, leaving the final determination of the time-varying vocal-tract contours to the following step.

(2) Synthesis of articulatory kinematics by interpolation in the articulatory domain. Thereby ***non-essential*** or redundant parameters (e.g. lip rounding in the articulation of a German [t]) are generated. Secondly, intermediate positions of articulary movements can be generated at an arbitrary rate.

(3) Graphical display and evaluation of sequences of mid-sagittal views. Speech organ contours are generated mathematically from complete sets of geometric parameters defined for a certain time step. Vocal-tract area functions are estimated from linear distances measured between speech organ contours.

(4) Acoustic synthesis with a wave digital filter implementation of a vocal tract model controlled by the time-varying area functions, cf. [7].

As the basic principle of operation has already been discussed in [8,10] only a few points of special interest will be discussed in the sequel.

3. ARTICULATORY PHONETICS AND COMPUTER ANIMATION

It appears obvious 'hat a full account of human articulation is impossible: neither the neuromuscular control of the speech organs nor the dynamics of their movements is fully understood. Even a phenomenological description of their kinematics seems quite untractable as the motion of three-dimensional non-rigid bodies is involved. Especially the continuous change of the tongue shape and position is hard to measure and to model adequately. What is left are a few basic facts describing certain stable articulatory mechanisms either in the steady state [11] or in transitions [12]. The rest is hypothesis.

How can this incomplete knowledge be exploited for speech synthesis? The answer is that even a rudimentary articulatory model introduces an additional level for the representation of spech phenomena such as coarticulation, reduction, assimilation, homorganic articulation, and other contextdependent allophonic variation. This additional level appears more suited to human intuition in the manipulation of hypotheses than the lower levels such as an exclusively acoustic signal description. Furthermore, it opens certain degrees of freedom hidden to the human experimenter at the acoustic level: some simple articulatory movements may induce very complex acoustic mechanisms that would not be recognized as basic to the speech production process at the acoustic level as they appear buried in the mess of signal variability.

Summarizing, the actual structure of an articulatory model is only partly determined by human articulation itself whereas an even larger part is due to the ***means of representation*** used in the desired application. As our application requires interaction with a human experimenter, a graphical display of speech organ movements is indispensible. Thus the principles of ***computer animation*** govern largely the design of our articulatory model.

(1) Animation of axionometric displays of three-dimensional shapes would be too clumsy on the envisaged "small" hardware environment.

(2) Two-dimensional shapes can be adequately displayed by their ***contours.*** Representing speech organs by their contours only, deliberately dismisses all knowledge concerning their morphology and internal dynamics. The prevailing information about articulator contours is conveyed by mid-sagittal (cine-)radiographies [11,13,14]. These are taken as the starting point for modeling the vocal-tract geometry.

(3) The methods for the synthesis of articulatory movements may be classified according to [15] as follows:

□ ***Image-based key-frame animation*** generates intermediate frames from fully specified ***key-frames*** by interpolation of shapes without taking into account any structural information about this shape. This principle is similar to the diphone synthesis concept in exclusively acoustic speech synthesizers. As the velar movement shows a single degree of freedom it can be adequately modeled by this technique.

□ ***Parametric key-frame animation*** has previously been used in articulatory synthesis [3] for a *synthesis-by-script* mode of operation. Still the human experimenter provides fully specified key-frames but these are interpreted in a parameter domain so that interpolated frames preserve certain structural characterictics of the parameterized shape. This principle is similar to the (allophone) synthesis by rule concept in exclusively acoustic speech synthesizers.

□ ***Kinematic algorithmic animation*** is our approach for the modelization of highly mobile variable-shape articulators, in particular the tongue, lips and epiglottis. There exists no similar concept in exclusively acoustic speech synthesizers. The synthesized frame sequence is no more specified from key-frames but from algorithmic parameter control laws. Because there is a direct *open-loop relationship* between the control laws and the controlled geometry and timing parameters this technique is a *kinematic* one. In our system, typical laws specify the durations of on-glide, stationary, and off-glide phases in the movement of a particular articulator within a given phone [16]. These durations may assume negative values e.g. to emphasize anticipatory coarticulation or reduction.

□ ***Dynamic algorithmic animation*** requires the replacement of the above kinematic laws by models of the internal speech organ ***dynamics.*** This approach [6, p. 279] goes beyond our previous option for simple contour line geometry. It introduces an additional level of representation, i.e. complexity, which we consider only worthwile to be studied in the context of text-to-speech synthesis after completing the study of pure kinematic models.

(4) Sampling of articulatory movements is sufficient at a rate of approx. 20 frames/sec for the human eye. However, this rate does not fully captûre the true motion of speech organs. For this purpose, a rate of at least 50 frames/sec should be used. Yet, it is important to separate the two rate requirements when implementing the computational model: every second, 50 frames must be calculated and evaluated by the graphics processing system while only 20 midsagittal contour plots must be output via the video display.

4. ACOUSTIC PHONETICS AND SIGNAL PROCESSING

Acoustic phonetics is seemingly more tractable than articulatory phonetics as there exist highly refined models of the vocal-tract acoustics such as [17]. More often than not, these models are delineated as an electrical circuit analog which can in turn be transformed into a digital circuit. The most elegant strategy consists in the ***wave digital filter*** (WDF) concept [18] which provides a direct translation of the analog voltage and current relations into the digital domain. A reasonably simplified WDF version of [17] has been implemented for the development of a quasi-articulatory speech synthesizer in [7].

We adopt this procedure while modifying its implementation according to our hardware system that comprises a vector-oriented bit-slice signal processor for the acoustic signal synthesis . This processor is controlled by a MC68000 microprocessor system developed at our department with special attention to the fast high-resolution graphics as needed for the animated articulatory model. The two processor systems are coupled via a parallel interface with a transfer rate of up to 3 Mbyte/sec. This interface transmits the area function values

+) On leave from Kaunas Polytechnical Institute, Jurostr. 65-302, 233028 Kaunas, USSR estimated from linear distances between speech organ contours on the basis of piece-wise approximation formulae given in [4]. The vocal-tract synthesis filter is tuned according to the area function in a time-varying manner. The operation of the signal synthesis can be supervised with a waveform editor and linear predictive analysis module integrated in the workstation utilities.

5. PERCEPTUAL PHONETICS AND SYSTEM EVALUATION

There are a lot of open issues that can only be studied in perceptual experiments implementing a feedback loop for system opimization through a *human experimenter*:

(1) How accurate must an acoustic vocal-tract model be, given its control by a fairly coarse articulatory model?

(2) What is the adequate level of representation for various speech phenomena? Adequacy should be defined by the human listener's judgement while the choice among several adequate representations should be made such that implementation complexity is minimized. For instance, it is not at all clear which articulatory transitions really *need* to be represented in the articulatory domain and which could be established by simplified rules operating directly on area functions or acoustic parameters.

(3) Feedback control should be made possible at all system levels. This calls for comparison mechanisms for mid-sagittal views and area functions as well as for spectrographic measurements. To fulfill this requirement, an interactive phonetic editor is built with thumb-wheel control of articulatory geometry and real-time output of the speech organ contours, the area function, and the synthetic speech signal.

(4) Special attention is devoted to rapidly time-varying speech events such as the explosion in stop consonants. For their detailed study both adaptive methods as well as new time-frequency analysis methods [19] are under investigation.

6. CONCLUSION

Several concepts fundamental to the design of a workstation for the development of a real-time articulatory speech synthesizer have been discussed. At the present state of the system, articulatory kinematics can be computed and displayed by our graphics system at a rate of 10 frames/sec approx. Speech signals can be produced with a sampling rate of 10 kHz. For a target system with 50 frames/sec and 20 kHz sampling an increase in computational capacity by a factor of 5 is needed. This is well within reach of off-the-shelf components (e.g. MC68020 with floating-point co-processor and 4 DSP chips such as TMS 32010). These data show an impressive technology step when they are compared to run-time data of articulatory models published a few years ago, e.g. 360 times real time in [2] or 20 to 60 times real time in [3]. Taking up this step is essential for applied articulatory synthesis.

7. REFERENCES

[1] J.L. Kelly, C.C. Lochbaum: Speech Synthesis. 4th Int. Congr. on Acoustics, Copenhagen, August 1962, paper G42.

[2] C.H. Coker: A Model of Articulatory Dynamics and Control. Proc. IEEE 64 (1976), pp. 452-460.

[3] Ph. Rubin, Th. Baer, P. Mermelstein: An articulatory synthesizer for perceptual research. JASA 70 (1981), pp. 321-328.

[4] G. Heike et al.: Berichte des Instituts für Phonetik der Universität zu Köln, IPKöln-Berichte 10(1980), 12(1982), 13(1986).

[5] T. Thomas, F. Fallside: A new articulatory model for speech production. Proc. IEEE Int. Conf. on Acoustics, Speech, and Sig. Proc. ICASSP'85, Tampa (Fla.) March 1985, pp. 1105-1108.

[6] R. Carré, R. Descout, M. Wajskop [eds.]: Articulatory Modeling and Phonetics. G.A.L.F. Groupe de la Communication Parlée - Proc. of the Symposium at Grenoble, July 1977.

[7] P. Meyer, R. Wilhelms, H.W. Strube: An efficient vocal tract model running in real time.
In: I.T. Young et al. [eds.]: Signal Processing III: Theories and Applications (Proceedings of EUSIPCO'86), North-Holland 1986, pp. 377-380.

[8] G. Dorffner, M. Kommenda, G. Kubin: GRAPHON-The Vienna Speech Synthesis System for Arbitrary German Text. Proc. IEEE Int. Conf. on Acoustics, Speech, and Sig. Proc. ICASSP'85, Tampa (Fla.) March 1985, pp. 744 - 747.

[9] A. Pounder, M. Kommenda: Morphological Analysis for a German Text-to-Speech System. Proc. 11th Int. Conf. on Computational Linguistics COLING'86, Bonn(FRG) August 1986, pp.263-268.

[10] G. Dorffner, G. Kubin: Artikulatorische Sprachsynthese. Mikroelektronik in Österrreich.
Wien: Springer 1985, pp. 456-461.

[11] G. Fant: Acoustic Theory of Speech Production. s'Gravenhage: Mouton 1970(2).

[12] O. Fujimura: Temporal Organization of Articulatory Movements as a Multidimensional Phrasal Structure.
Phonetica 38 (1981), pp. 66-83.

[13] Wängler: Atlas deutscher Sprachlaute.
Berlin: Akademie-Verlag 1981(7).

[14] G. Lindner: Optische Analysen der Koartikulation durch Röntgenkinematographie. Hochschulfilm T-HF 719. Berlin: Institut für Film, Bild und Ton, 1972.

[15] L. Forest, N. Magnenat-Thalmann, D. Thalmann: Integrating Key-Frame Animation and Algorithmic Animation of Articulated Bodies.
In: T.L. Kunii [ed.]: Advanced Computer Graphics. Tokyo: Springer-Verlag 1986, pp. 263 - 274.

[16] G. Dorffner, M. Kommenda: Ein Artikualtionsmodell zur Sprachsynthese. Fortschritte der Akustik - DAGA'85.
Bad Honnef: DPG-GmbH 1985, pp. 615-618.

[17] J.L. Flanagan, K. Ishizaka, K.L. Shipley: Synthesis of Speech From a Dynamic Model of the Vocal Cords and Vocal Tract.
BSTJ 54 (1975), pp. 485-506.

[18] A. Fettweis: Wave Digital Filters: Theory and Practice. Proc. IEEE 74 (1986), pp. 270-327.

[19] W. Wokurek, G. Kubin, F. Hlawatsch: Wigner Distribution - A New Method for High-Resolution Time-Frequency Analysis of Speech Signals. Proc. 11th Int. Congr. of Phon. Sciences, Tallinn August 1987, this volume.

APPENDIX

As a reference to our articulatory model two figures are presented:

Fig. 1 shows the parameterization of the articulatory geometry by approximation of the speech organ contours with simple mathematical functions (circle, tangent).

Fig. 2 shows a (subsampled) synthetic frame sequence for the German word [matəma:tik].

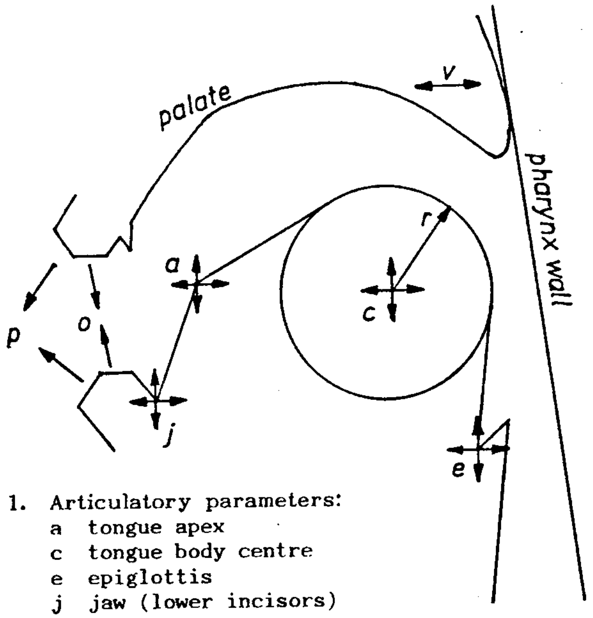


Fig. 1. Articulatory parameters:
a tongue apex
c tongue body centre
e epiglottis
j jaw (lower incisors)
o lip opening
p lip protrusion
r tongue body radius
v velum
palate and *pharynx wall* are fixed reference positions

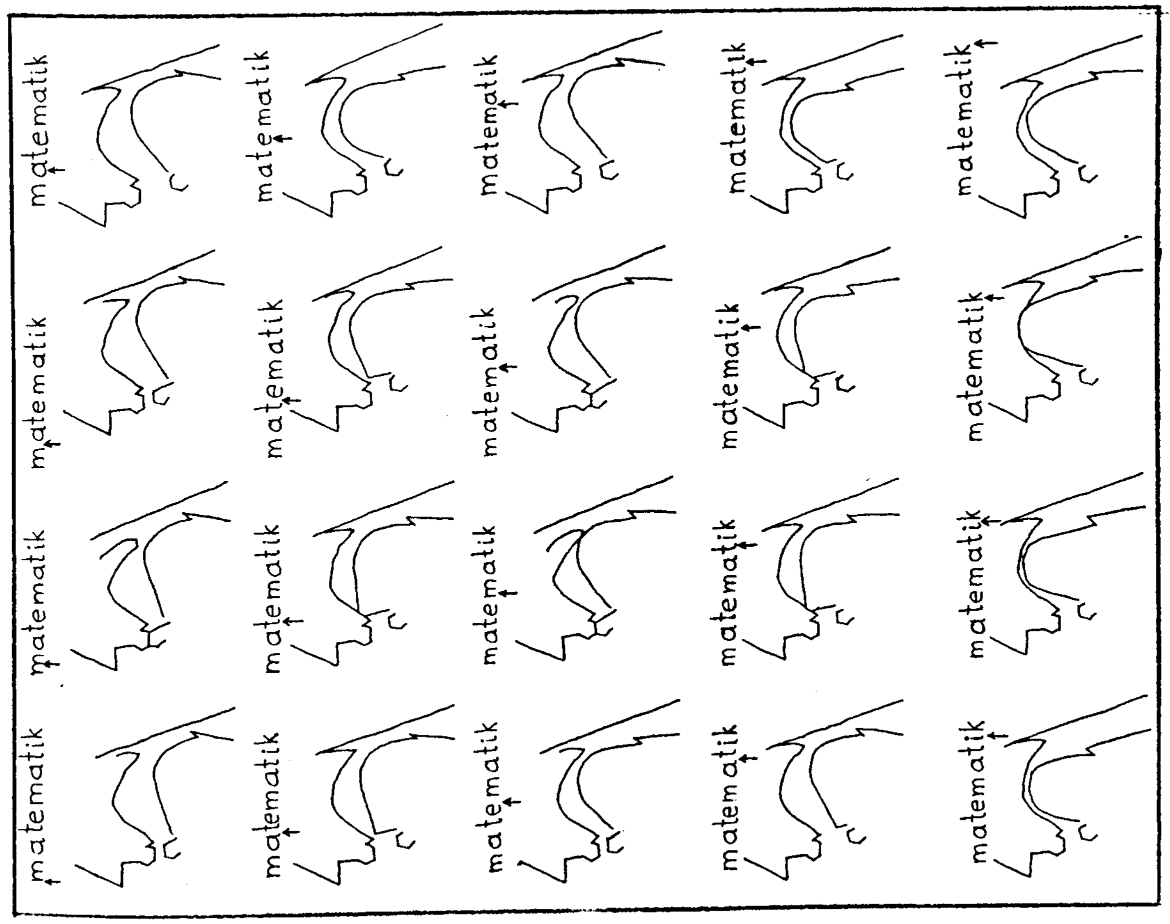


Fig. 2. (Subsampled) synthetic frame sequence for [matəma:tik], the arrow points at the current time position of the frame within the whole word.

'COARTICULATION' IN AN ARTICULATORY SYNTHESIS MODEL OF GERMAN

GEORG HEIKE

Institut für Phonetik
Universität zu Köln
Greinstraße 2, D-5000 Köln 41

ABSTRACT

'Coarticulation' is the main problem in speech synthesis. In the case of German we show that the control of articulatory parameters is dynamic in nature, i. e. depends on effort and time of articulatory gestures.

The purpose of this contribution is the development of a framework in which 'coarticulation' rules for the articulatory synthesis of German can be established. Starting point is the conclusion that the traditional concept of 'coarticulation' must be rejected as inadequate because it presumes discrete phonetic segments as input units into a coarticulatory module. Therefore a target oriented model of articulatory control is proposed. Input units to the control module are labelled by phonetic symbols. They are defined by at least one target value of one parameter (e. g. for bilabials) or more parameters (in the case of most other sounds).

With the exception of the bilabial (rounded) [ʃ] German consonants are defined by one target value only, namely the constrictional position of the lips, the glottis, the anterior part of the tongue or the dorsum. The remaining articulatory configurations, for example the shape of the lips and of the dorsum in the case of an apico-dental consonant, have to be specified according to the syllabic context. This specification, usually termed as 'coarticulation', is language dependent and has to be formalized in an articulatory synthesis model. Since there is a lack of sufficient experimental investigations (especially x-ray studies) in the case of German, our method is restricted to the articulatory interpretation of sonagrams, self-observation, and auditory control of synthesis output.

Preliminary results suggest the hypothesis that the complete articulatory specification of German consonants depends on the factors: vowel context, type of consonant, position within the syllable, speed of articulation. These dependencies will be exemplified in the case of the apicodental [l] and the dorso-velar (or palatal) consonant [k]. In an initial prevocalic position (e. g. [li:] as in 'Liebe') the most economic (and hence 'coarticulated') position of the tongue would be the same as for [i:] except for the elevation of the tongue tip. This would result in a 'l' with palatalized dorsum. Although German is not said to be characterized by such consonants, the above-mentioned case can be observed in fluent speech, especially in intervocalic position, as e. g. in 'die Liebe'. There is, however, a remarkable difference in the 'coarticulative' effect of the vowel context between slow and fast articulations. In the case of relatively slow articulation - which means slower movement of the tongue tip and greater duration of dental contact - there is enough time for the tongue body to move towards the neutral position. The same principle holds for stop consonants, but since closure and release gestures are clearly separated, relatively fixed in time, and hence independent of speech tempo, a very distinct difference between syllable initial and final position can be observed. The articulatory position of release, e. g. in 'lieg' [li:k^h], may result from a backwards movement of the dorsum during the closure time interval, whereas in 'Kiel' [k^hi:l] the dorsum must in any case take the appropriate palatal position for the following [i].

Fig. 1a shows the sonagram of the VCV-portion of a relatively careful pronunciation of 'die Liebe'. An appropriate articulatory synthesis of the VCV-gesture can be achieved with a tongue profile of l with central position of the dorsum, as midsagittal tracings (fig. 1c) of i and l and the sonagram of the synthesis output (fig. 1b) show.

Fig. 2 presents, in a case of more rapid pronunciation, both reduction of the unstressed i in 'die' and the palatal position of the tongue dorsum for l, largely identical with that of i. Thus

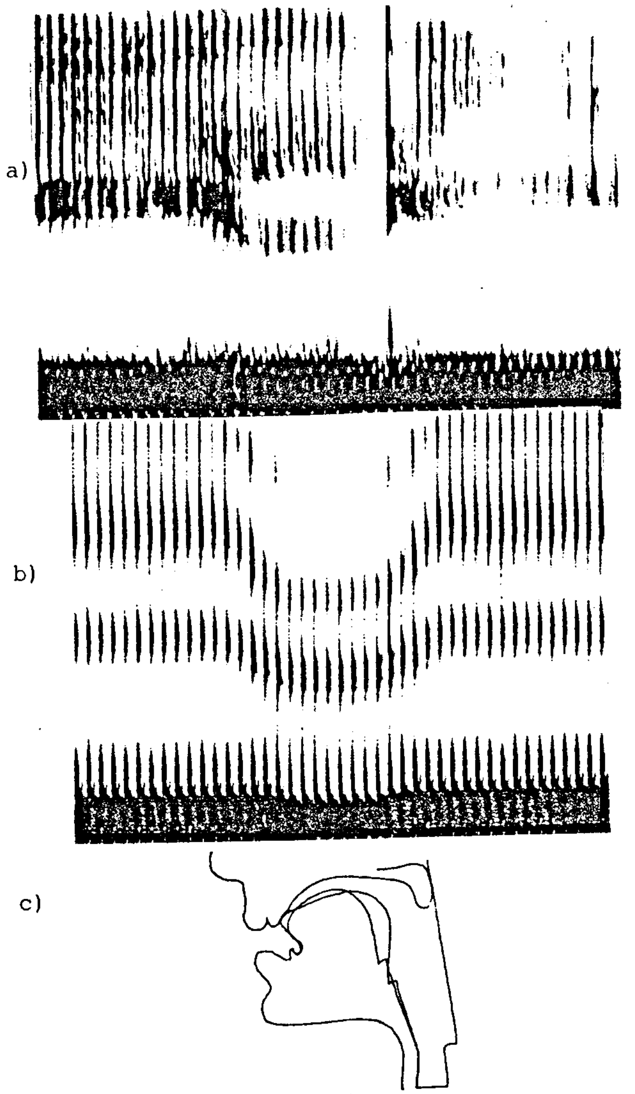


Fig. 1: Spoken (a) and synthesized (b) VCV-portion [ili] of German 'die Liebe', relatively slow speech. Midsagittal computer tracings (c) show corresponding target positions

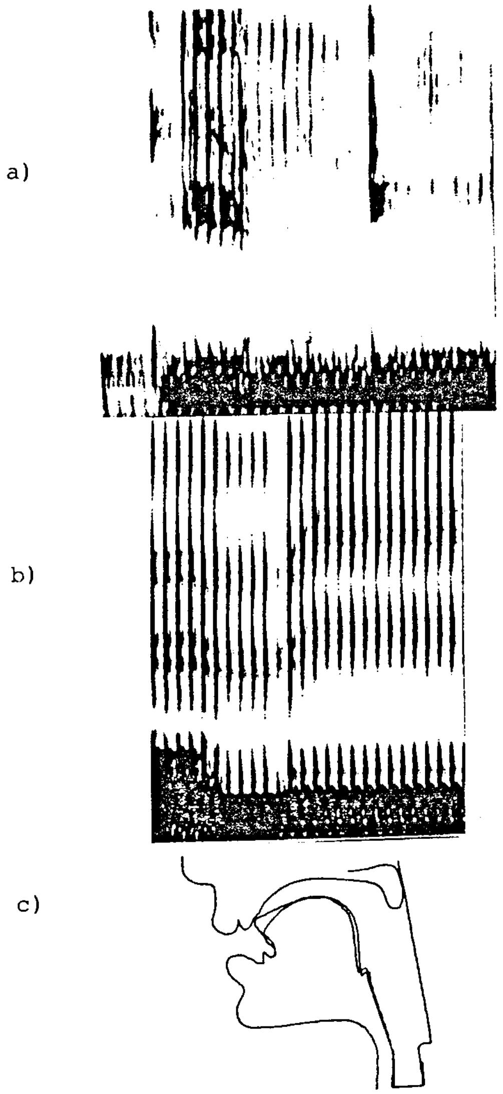


Fig. 2: The same VCV-transitions as in fig. 1, spoken with relatively rapid speech tempo

we may hypothesize that the control of the tongue dorsum as a function of speech tempo works differently for vowels and consonants. We observe reduction (towards neutral) of vowels with high tempo, whereas with consonants a similar effect appears with slow tempo.

But both phenomena can be explained by one principle: the economy of effort as a function of time. Effort may be defined in two respects: as the effort of reaching a target, and as the effort of maintaining a position different from neutral. With high speech tempo reduced effort results in an 'undershot' of the movement towards a target (vowels), with slow tempo the effort of maintaining an extreme position of the dorsum (unnecessary for consonants) is reduced and compensated by a movement to a central position. The same principle holds for the parameter of lip rounding. In the sequence [uli], for instance, we notice (with normal-to-slow speech tempo) a lip spreading gesture from u to l and vice versa. In synthesizing ['uli] ('Uli') we would therefore expect the spreading gesture of the l to be continued in the transitional movement towards i (see fig. 3). Comparison with the rather rapidly spoken utterance shows not only a remarkably reduced u (more centralized and less rounded), but also an anticipation of the spreading gesture within the duration of the vowel. This demonstrates

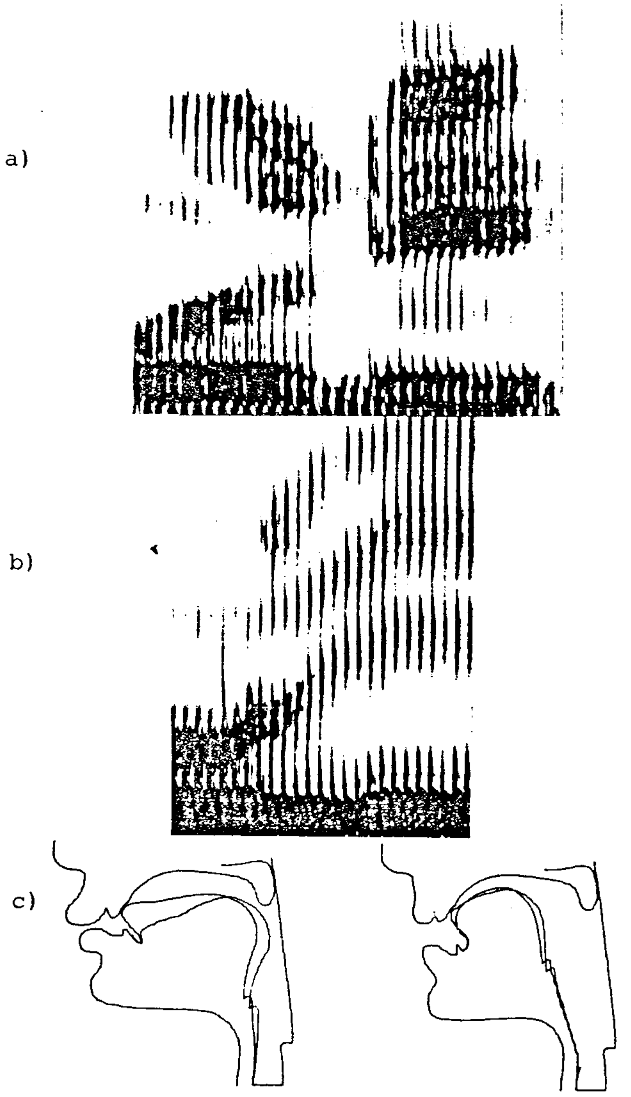


Fig. 3: VCV-transitions of 'Uli';
a) spoken,
b) synthesized without anticipation of lip spreading.

the usefulness of articulatory synthesis for the study of 'coarticulation' phenomena (i. e. articulatory control as a function of time).

Quasi-Articulatory Real-Time Speech Synthesis

Peter Meyer, Reiner Wilhelms and Hans Werner Strube

Drittes Physikalisches Institut, Universität Göttingen,
Bürgerstrasse 42-44, D-3400 Göttingen, Fed. Rep. of Germany

ABSTRACT

To produce natural sounding transitions with a speech synthesizer by simple interpolation of its control parameters, these parameters should have articulatory meanings. In this case the synthesizer must have the form of a vocal tract. We embedded such a model into a simple dynamic articulatory system and applied Kalman filtering methods to estimate the articulatory parameters. From these parameters we extract simple rules for speech synthesis. The synthesizer is based on a signal processor system and runs in real time.

THE ARTICULATORY MODEL

The articulatory model is controlled by seven parameters ($a_1,\dots,a_7$) which determine a discretized 10 tube model of a vocal tract and a 7 tube model of a nasal tract. Parameters a_1 and a_2 describe the tongue body and the shape of the pharynx in a simpified manner by linearly superposing two basic vocal tract shapes and a constant neutral shape. The different places of articulation in the palatal and alveolar region can roughly be described by them. The front palatal and dental articulation is described by parameters a_3 and a_4. a_4 represents the place of the tip of the tongue and a_3 is treated as a parameter of the strength of articulation. a_5 and a_6 determine the radiation from the vocal tract, which is simulated by discretized horns terminating the vocal and nasal tract. a_7 determines the coupling of the nasal tract to the vocal tract.

MODEL FITTING

In order to get transitions of parameters suited for speech synthesis, the model must be fitted to natural speech, that is, we have to find a mapping from an acoustic parameter space to the space of articulatory parameters. It is known from theoretical and practical considerations that this mapping cannot be unique. Thus, we have to restrict ourselves to searching for trajectories that do not contain jumps and that give a representation of measured short-time spectra in a least squares sense. Our method to find this mapping is based on Kalman filtering and Kalman smoothing. We extended the 7-vector of articulatory parameters to the 21-state vector of a dynamic model which is a critically damped 2nd order system with unknown white noise input and unknown control input. Formally:

$$x = (x_1;\ x_2;\ u)'\ ,\ x_1 = (a_1,\dots,a_7)'$$

x_1: vector of articulatory parameters,
x_2: delayed articulatory parameters,
u : unknown control input.

$x_{n+1} = \Phi x_n + w_n$,
w_n : vector of white noise with
$\langle w_n \rangle = 0,\ \langle w_n w_n' \rangle = Q$
The transition matrix is

$$\Phi = \begin{pmatrix} 2A & -A & (I-A)^2 \\ A & 0 & 0 \\ 0 & 0 & I \end{pmatrix}$$

A is a diagonal 7x7 matrix of $\exp(-T/\tau)$;
T: frame length, τ: time constant.

The trajectories of the dynamic system are to be estimated in accordance to natural short time spectra. Thus, based on utterances of one speaker that are digitized with 10 kHz after preemphasis, we estimate ARMA coefficients every 2.5 ms using a Hamming window of 25 ms, and we take the smoothed logarithmic ARMA spectra as a reference. The resulting sequence of short time spectra is called the real measurement process z(t).
The analysis procedure computes the acoustic velocity transfer function of the model's vocal and nasal tract, given the actual estimate of the state $\hat{x}$. The logarithmic spectrum of the transfer function, which is called $h(\hat{x})$, is the model measurement. Then we formally assume that the measurement process z(t) is produced by the model and disturbed with noise:

$$z(t)=h(x(t))+r(t),$$

and thus is related to the 'true' state x. r is a random vector with zero mean and covariance R, it is assumed to contain measurement noise and all model inadequacies as well.
The computation of h(x) requires some computational expense. For this purpose the vocal tract is described by four-terminal networks

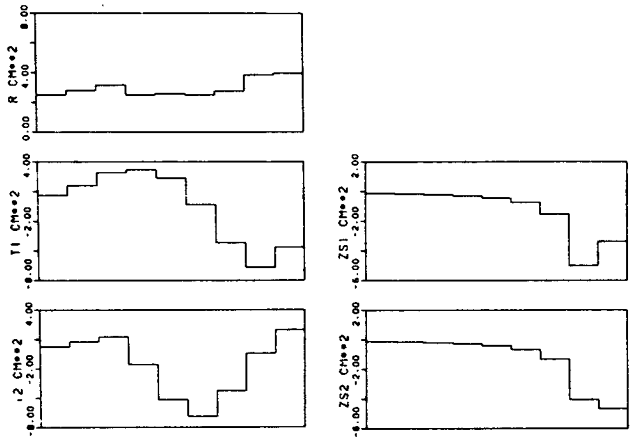


Fig. 1: The articulatory model.
Left: Constant neutral and two basic shapes.
Right: Tongue-tip component in its two extremal points of articulation.

in the view of electrical circuit analogue. The transfer functions from the vocal source and from the fricative source (only one assumed) to the mouth and to the nostrils can be computed, and h(x) is obtained by adding the partial transfer functions and computation of the power spectrum.

As h(x) is a nonlinear function, we can not apply the linear Kalman filter algorithms, an iterated Kalman filter is more convenient. It requires that at each step of iteration the matrix H(x) containing the partial derivatives of h(x) must be evaluated numerically at the actual estimate of x. As h(x) is of high dimension (we take a 64-point FFT), we make use of an inverse covariance Kalman filter. The algorithm works as follows:

Let $\hat{x}^-$ be the estimate of the state at time t with covariance P^-, before the actual measurement z is incorporated. Starting with $x_0=\hat{x}^-$ it iterates:

$$x_{k+1} = x_k + (H'R^{-1}H + (P^-)^{-1})^{-1}$$

$$\cdot(H'R^{-1}(z-h(x_k)) - (P^-)^{-1}(x_k-\hat{x}^-))$$

$$\text{for } k=0,\dots, K. \quad \hat{x}^+ := x_K.$$

$$P^+ = (H(\hat{x}^+)'R^{-1}H(\hat{x}^+) + (P^-)^{-1})^{-1} .$$

The inverse covariance R^{-1} of the measurement noise r is simply defined as a time varying diagonal matrix. It plays the role of a weighting function for the particular measurements. The algorithm presented above can be extended to a (computationally rather tedious) smoothing algorithm by requiring that the estimate of state $\hat{x}$ at time n is not only determined by the measurement history up to time n but also by future measurements up to time n+m. For one update of the smoothing algorithm a Kalman filter, starting with the present state of the smoother, first runs forward up to time n+m, then, using the adjoined backward dynamic model, backwards to time n. At each measurement it makes an update of its state. The state of the smoother is then updated by assembling its actual state and the state of the backwards filter.

INCORPORATION OF ARTICULATORY CONSTRAINTS

As could be expected, the described procedure works sufficiently for most of the pure vocalic transitions. For voiced-unvoiced transitions and for nasalized transitions some constraints have to be incorporated into the model. We do this in a straightforward way: If we know, e.g. that the velum must be open at an interval and closed at another, we 'tell' the Kalman filter that we measure the behavior of the velum parameter, that is, we include the pseudo-measurements into the general measurement history z(t) and give them more or less influence by defining the corresponding entries in the diagonal of R^{-1}. In a similar manner all parameters that are functionally related to the state of the model can be prescribed, such as place of articulation and strength of fricative excitation.

ANALYSIS OF FITTED ARTICULATORY TRAJECTORIES

For analysing and testing the fitted articulatory trajectories, the vocal tract model was implemented on a signal processor system. This system consists mainly of a fast signal processor TMS-32010 from Texas Instruments, a fast parallel interface and a 16-bit D/A-converter. The signal processor is fast enough to calculate the vocal tract model in real time. The articulatory model as well as the "articulator-to-filter" transformations are done on a laboratory computer (Gould 32/9705). The filter parameters are transferred to the signal processor system at a frame rate of 200/s.

It was possible to resynthesize intervals of an adapted trajectory as well as fixed parameter sets. On this way it was possible to extract subjectively constant parts of vowels or consonants.

Fig. 4 shows extracted vowels in the plane of the first two articulatory parameters (a_1, a_2). If we interprete the first parameter as a front versus back, the second as a high versus low parameter, the positions of the vowels are close tongue hump positions. In this plane some vowels like /e:/ and /i:/ or /a/ and /ə/ are very close. They differ from each other in the third (tongue tip) parameter. For example this parameter is higher for /i:/ than for /e:/ which can be interpreted as an articulation more in the front of the tract.

For the most cases the transitions between phonemes are regular. Attempts to find a fixed dynamics for the articulation failed. Every transition seemed to have a different dynamics. The best and most easy description of the transitions was linear interpolation or a critically damped second order system with varying parameters.

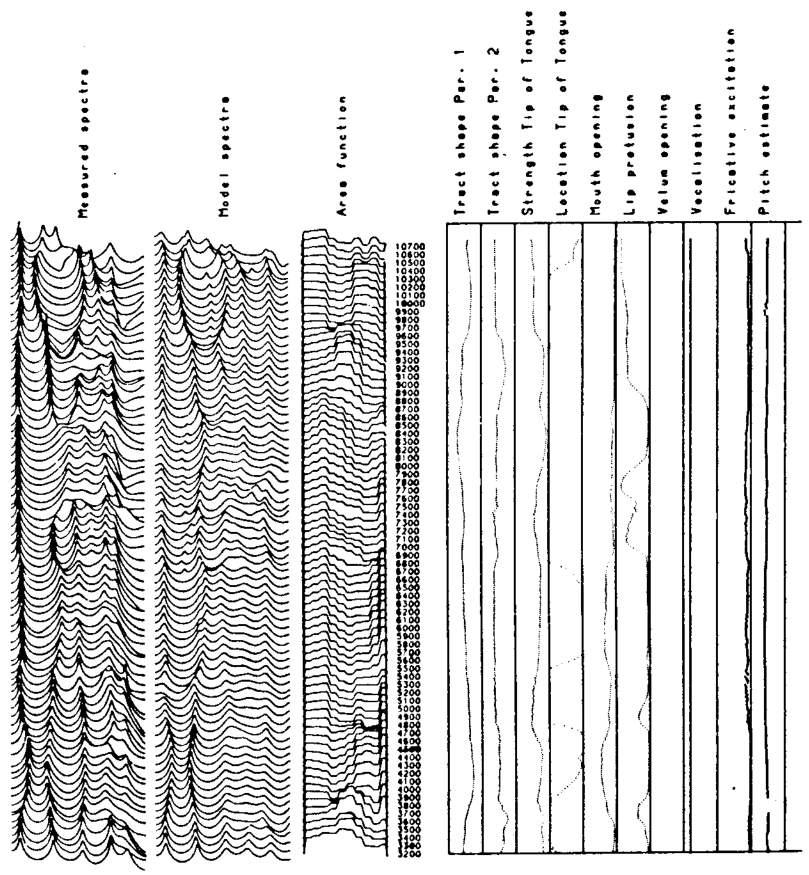


Fig. 2.
Fitted utterance /la:le:li:lo:/.
Time runs from bottom to top.
Velum opening was fixed to zero.

For both figures on this page:
The vocalization parameter is chosen 'by hand'. It determines the energy of the input signal of the four terminal network which represents the transfer from the glottis to the radiation. This parameter also determines the glottal shunt, which increases for low vocalization. The fricative excitation parameter is computed by the fitting procedure based on simple assumptions about the relation between closest constriction and turbulent noise strength.

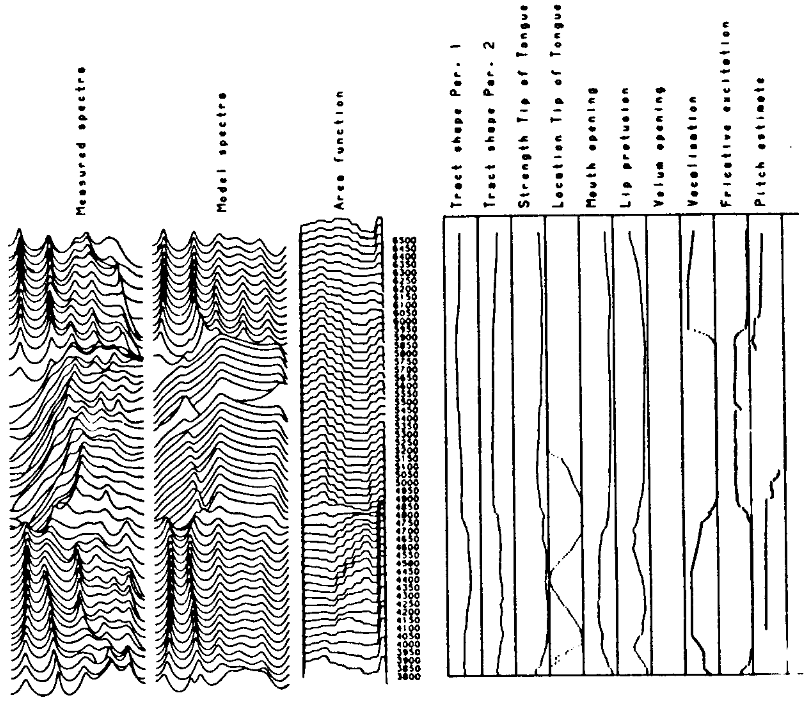


Fig. 3.
Fitted utterance /taʃə/.
Starting at the voice onset of /ta/

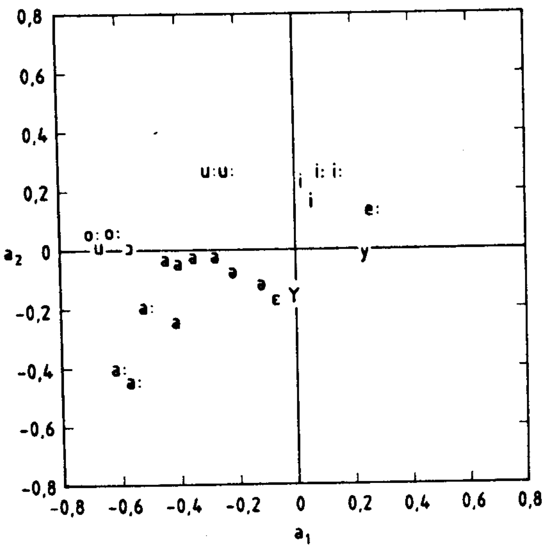


Fig. 4. Points of articulation of some German vowels in the plane of a_1,a_2.

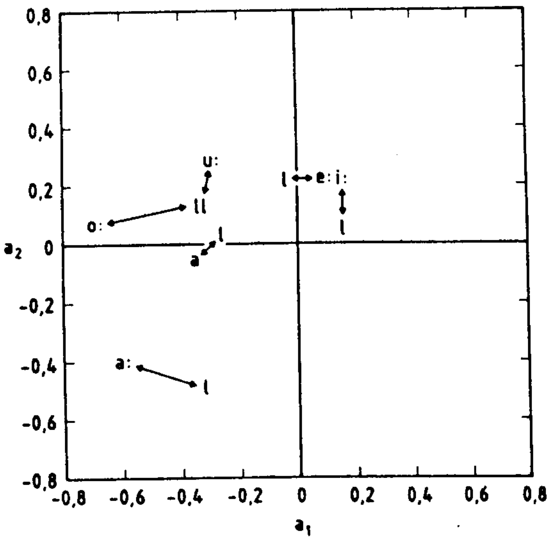


Fig. 5. Points of articulation for liquid /l/ embedded in different vowel context like /a:la:/.

The adapted articulatory parameters show strong effects of coarticulation. We examined different VCV-transitions with the liquid /l/ like /a:la:/,/e:le:/... Fig. x shows the points of articulation of the l's in the plane of the first two parameters. The articulation is close to the position of the surrounding vowel. The main articulation is done by the third, the tongue tip parameter. For all shown articulations this parameter produces a constriction of the 8'th or the 9'th tube segment of about 0.5 cm^2. This effect can be seen if we look at the articulation of the l's in the plane of the first and third parameter where they lay on a straight line. Similar results can be found for other consonants, for example for the nasal /n/.

Strong articulatory effects can be found for the articulation of plosives which is close to the following vowel. This effect is extreme for the plosives /p/,/b/ which are articulated nearly in the same way as the following vowel with a mouth opening of zero.

SYNTHESIS

For speech synthesis purposes we stored 36 parameter vectors of different German phonemes in a table. This table contains also information about voiced or unvoiced exitation, the strength of the fricative exitation, duration, voice onset times etc. Only for few transitions it is possible to make a synthesis by interpolating between these parameter vectors. So we add further vectors which all represent the same consonant to describe coarticulatory effects, such as the articulations of different /l/ in Fig. x. If we synthesize a transition from a vowel to a consonant we interpolate to the representant which is nearest to the vector of the vowel. If we want to synthesise a VCV-transition we interpolate during the 'constant' consonant part to the representative which is nearest to the following vowel vector. These rules are described by addresses to the consonant vectors written in a 36x36 matrix. This matrix also contains information about the duration of the transition.

Parameters like voice onset times for plosives and strength of fricative exitation are chosen subjectively.

SPEECH MOVEMENT RESEARCH USING THE NEW X-RAY MICROBEAM SYSTEM

ROBERT D. NADLER

Waisman Center
University of Wisconsin
Madison, WI, USA 53705

JAMES H. ABBS

Depts. of Neurology and
Neurophysiology
Waisman Center
University of Wisconsin
Madison, WI, USA 53705

OSAMU FUJIMURA

AT&T Bell Laboratories
Murray Hill, NJ, USA 07974

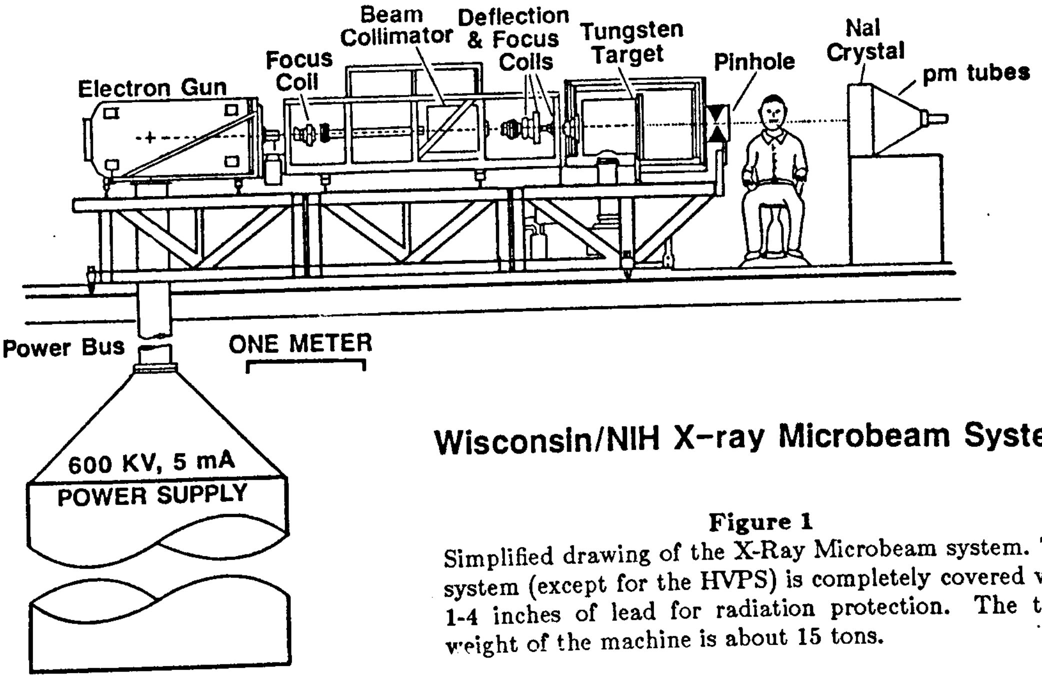
ABSTRACT

A new X-Ray Microbeam system for studying tongue movements and other articulatory gestures has been constructed to serve as the core instrument of a shared speech production research facility. Preliminary speech movement data has been obtained and this system is currently capable of tracking multiple articulatory pellets (up to 12) at aggregate sampling rates of about 1000 per second. Radiation exposures are very low due to the narrow x-ray beam and localized computer-controller scans used for tracking. The facility includes parallel capability for data display and analysis for multiple experimenters.

[Support for this facility is provided by NIH-NINCDS grant NS-16373.]

NEED FOR FACILITY

The progress of research into human speech production is severely limited by a number of factors inherent in the speech production process itself. In-depth physiological investigation of critical speech processes (e.g., neural activation, muscle activity, movement, *in vitro* biomechanics, etc.) cannot be conducted optimally because there appears to be no truly suitable animal model. Because the human speech apparatus has multiple overlapping functions (breathing, chewing, swallowing, *and* speech), functional inferences made from anatomical investigations provide only fragmentary and potentially confusing results. For example, while the masseter and temporalis muscles are capable anatomically of substantial jaw closing forces, they are generally inactive during movements of the jaw for speech production. These limitations require that many of the significant physiological issues surrounding human speech production *be addressed with human subjects under normal conditions of speech.*



Wisconsin/NIH X-ray Microbeam System

Figure 1
Simplified drawing of the X-Ray Microbeam system. The system (except for the HVPS) is completely covered with 1-4 inches of lead for radiation protection. The total weight of the machine is about 15 tons.

While speech production processes are inherently difficult to investigate, recent advances have demonstrated that the careful application of analysis and interpretation techniques adapted from systems physiology, biomedical engineering, and signal processing provide a means to exploit the limited available data. That is, in the last ten years, the methods for interpretation of multivariable speech events have become increasingly utilized by the speech physiologist.

To advance this important work, we have developed an x-ray system with increased capabilities to obtain large samples of the relevant movement, EMG, aerodynamic, and acoustic data *simultaneously*, data with which to elaborate and refine preliminary models and further test their viability. The X-Ray Microbeam system that previously existed at the University of Tokyo [1] provided preliminary studies with suggestive information demonstrating the strength of this method [2,3,4,5].

X-RAY SYSTEM

Figure 1 is a simplified drawing of the x-ray generator. Its major components include a 600 kV, 5 mA power supply, a source electron gun and accelerating column, beam-line components for electron beam focusing and deflection, a thin (900 micron) water-cooled Tungstun target for photon generation, x-ray pinhole, and NaI detector.

Sampling of the x-ray detector output, control of the beam deflection and x-ray scanning, and implementation of the pellet spherical pattern recognition/background subtraction is done with specialized digital hardware controlled by two fast microsequencers which in turn are remotely controlled by the main computer processor (VAX 11/750). All of these processors communicate through a shared four-port memory. This implementation has been designed (in particular, the 600 kV acceleration voltage) to allow us to track pellets in the presence of common tooth filling amalgams. Development of the specialized algorithms and support computational hardware to provide this capability is currently under way.

As shown in Figure 2, small (2-3 mm) gold pellets are placed, for example, on the tongue, lips, maxillary and mandibular teeth, and velum. The x-ray generator emits a narrow x-ray beam whose two-dimensional position on the object field of the subject's head is computer controlled. The x-ray beam passes through the object field and is detected by a scintillation counter. The output of the scintillation counter reflects the relative radiopacity of the image field through which it passed.

Based upon the previous pellet positions or the pellet position determined during an initial scan, the main computer transmits a set of initial X and Y coordinates to the digital scan controller. The computer system locates the position of a pellet by first obtaining the x-ray image generated using a locally restricted raster scan (as shown in Figure 2). This image, after background subtraction, is subjected to a global template recognition algorithm to determine the pellet location within the scan area. The predicted pellet position is determined using an algorithm in the main computer that utilizes previous pellet displacement, velocity, and acceleration. Each of the other pellets are scanned in turn with associated image processing to provide time-motion tracking of all the pellets.

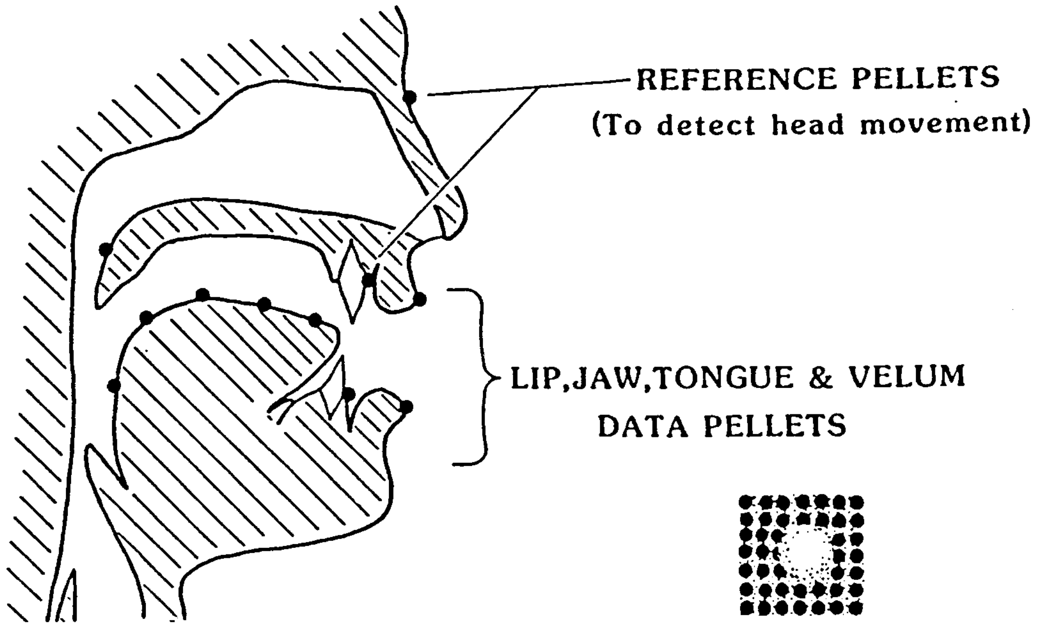


Figure 2
Mid-sagittal view of speech articulators with attached gold pellets. Small computer generated scan showing image of a gold pellet.

Computer Image
of
Microbeam Scan Pattern

Radiation dosages to the subject are limited to very low levels due to a combination of many factors. They include the relatively small size of the x-ray beam (approximately 1 mm at the subject mid-sagittal plane), the limited time (pixel exposure is 10 microseconds maximum), the beam is allowed to expose any particular area of the oral cavity, tissue not in the immediate vicinity of a pellet is not exposed, and secondary photon scatter (Compton effect) is reduced by the high energy of the primary photons (photon energies below 100 KeV are filtered out). Multiple measures and estimates indicate that the system at an acceleration voltage of 400 kV and an electron beam current of 5 mA will yield average entrance radiation exposures of 0.56 mR/minute, or a total of 8.4 mR for 15 minutes of data acquisition. Other measures, such as the peak radiation exposure for a given small volume of tissue (1.0 cm^3) in the worst case, are also very low (11 mR/minute or 165 mR for 15 minutes of data). These exposure levels compare very favorably with exposure levels for other clinical and experimental procedures. For example, a single dental bitewing yields an entrance exposure of 650 mR while the same 15 minutes of speech data with cineradiography would require a prohibitive entrance exposure of 7.5 R. Given these measures and comparisons, we are confident that the radiation exposure as a result of this procedure will offer negligible risk if appropriate precautions are observed.

ACQUISITION OF OTHER SIGNALS

Instrumentation associated with this facility also allows for simultaneous detection/transduction and conditioning for the following speech production parameters: (1) the airborne speech acoustic signal, (2) an accelerometer/throat microphone signal, (3) up to four channels of aerodynamic signals or signals from other strain gage transducers and (4) up to ten channels of electromyography (EMG). A custom designed A/D subsystem which acquires this analog data at aggregate rates up to 125K samples/second (15 bits, isolated) has been built to provide this capability. This implementation provides for differential sampling across up to 64 channels that is time-synchronized with the x-ray system pellet movement data. A multi-channel D/A subsystem is also provided for audio playback of speech acoustic data as well as analog output of physiological data.

NETWORKING AND ANALYSIS CAPABILITIES

A general purpose networking system (consisting of Ethernet and Pronet local area networks) provides high bandwidth inter-computer communication capabilities for both data acquisition and analysis functions. Each processor on the network (data analysis graphics workstation or data acquisition CPU) has shared access to the central file servers. SUN graphics workstations provide a bit-mapped 1024 X 1024 monochrome display and are well suited for manipulation of multi-channel physiological and acoustic data. A custom designed data base has

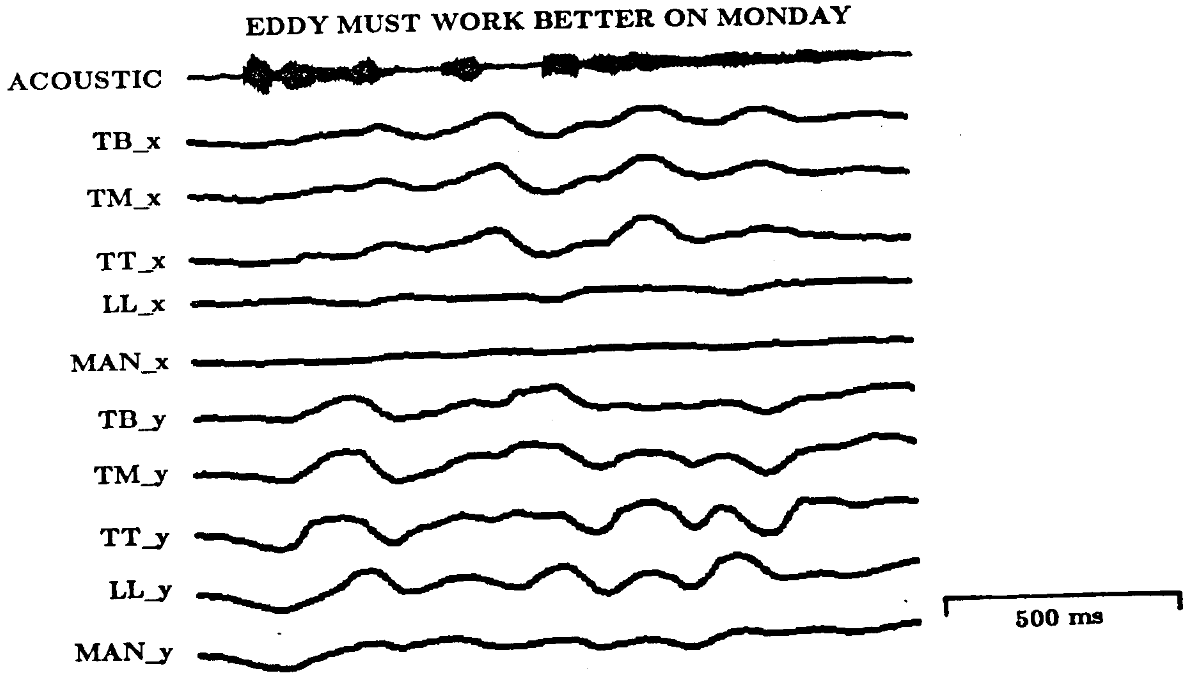


Figure 3
Simultaneously acquired acoustic speech signal and X- and Y-coordinate data from 5 pellets tracked with the x-ray microbeam system. See text for details.

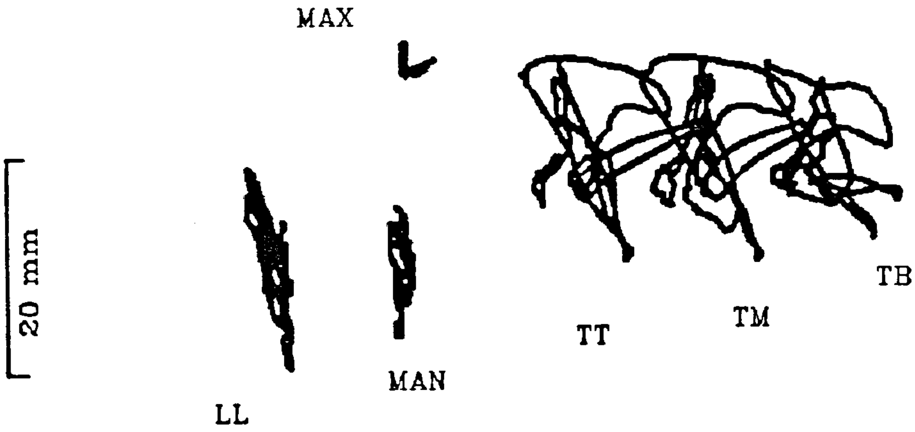


Figure 4
Cartesian coordinate plot of the same data from Figure 3. A head reference pellet is also included.

been constructed with binary descriptive and data files designed for optimum storage efficiency and access. In addition to the general windowed/mouse environment provided by the SUN workstation, graphics applications have been developed specifically for multiple data signal display, manipulation, and analysis. In addition, the 'S' statistical package (licensed from AT&T Bell Laboratories) provides an interactive computing environment and statistical data analysis language as well as a wide variety of specialized graphics capabilities.

EXPERIMENTAL RESULTS

Figures 3 and 4 present an example of typical data acquired using the x-ray microbeam system. This experiment used three tongue (TB, TM, and TT), a mandible (MAN), a lower lip (LL), and two maxillary reference pellets (MAX). Each of the data pellets was acquired at 100 samples/second, the reference pellets at 50 samples/second, along with a single channel of speech acoustic data at 10,000 samples/second. These data have not been digitally filtered or corrected for head movement. Normally, at least two head reference pellets are sampled so that articulatory pellet movement data can be corrected for head movement (translation and rotation in the mid-sagittal plane) prior to data analysis.

REFERENCES

[1] Fujimura, O., Kiritani, S., and Ishida, H. Computer controlled radiography for observation of movements of articulatory and other human organs. *Comput. Biol. Med.* · 1973, *3* , 371-384.

[2] Kiritani, S. Articulatory studies by the X-ray microbeam system. In M. Sawashima and F. S. Cooper (Eds.), *Dynamic aspects of speech production*. Tokyo: University of Tokyo Press, 1977, 171-194.

[3] Harris, K. S., Tuller, B., Bell-Berti, F., and Kelso, J. A. S. *Some temporal rules in speech production.* Paper presented to the 105th meeting of the Acoustical Society of America, Cincinnati, Ohio, May 1983.

[4] Hirose, H., Kiritani, S., and Sawashima, M. Patterns of dysarthric movements in patients with amyotrophic lateral sclerosis and pseudobulbar palsy. *University of Tokyo Annual Bulletin,* 1980, *14* , 263-272.

[5] Fujimura, O. Relative Invariance of Articulatory Movements: An Iceberg Model. In J. S. Perkell and D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes,* London, Lawrence Erlbaum Ass., 1986, 226-234.

TENSOPALATOGRAPHY DYNAMIC TECHNIQUE AND ARTICULATORY TENSION STUDY COMPARED WITH SPEECH ACOUSTIC PARAMETERS (SYLLABLE PRODUCTION IN SPOKEN SPEECH AS AN ARTICULATORY STRUCTURE)

SKALOZUB L.G., PAVLICHENKO A.N.,
TERYAYEV D.A.

Dept. of Filology, Laboratory of Experimental Phonetics Kiev State University
Kiev, Ukraine, USSR, 252017

The paper presents phonetico-experimental data concerning dynamic tenso palatography technique which makes it possible to correlate the dynamic articulatory processes as the last link of the syllables, words, syntagmas in speech production with their physical characteristics. Analysed were tensooscillograph records of Russian syllables with the vowel "a" and with initial stop front lingual consonants differing in their hard, palatalized, voiced, voiceless and sonant features.

The study and a statistical analysis of the data received yielded in: 1. The syllable is produced as an articulatory integrity. 2. The syllable includes an aggregate factor manifesting itself in a lesser/greater homogeneity of the components due to muscular tension. 3. Syllables differ according to the mode their articulatory tension develops.

The Kiev State University Experimental Phonetics Laboratory (KSUEPHL) experimental phonetics study of over some recent decades has aimed at a detailed describing the speech production articulatory aspect which was due to both practical tasks (to understand articulatory standardization, to correct pronunciation) and an important objective dealing with syllable- and word-structure.

A syllable, a minimal structural unit of a spoken speech stretch, has an aggregate factor enabling to find its wholeness and continuity as an original chain structuring lexical and syntactical language events.

The study of syllables and words dynamic manifestations makes us possible to treat them resulting from the final chain speech production - one of the speech activity processes. An articulatory process thus turns out to be immediately related to the language (phonological) speech study: the syllable/word articulation presupposes the producing of language units functioning as spoken text components elsewhere.

The academecian L.V.Shcherba theory suggests that the syllable production and division is dependent on the muscle tension impulses which are responsible for a consonant power changeability within a syllable. /10, 4/

This hypothesis experimental test made us to design a technique enabling us to record articulatory tension and its progress within a syllable, a word and word sequence.

The papers and books on phonetics have not yet had experimental evidence concerning the tension growth within a syllable, though the syllable peak (a vowel in most languages) is believed to be made with the most tension possible. There were some ideas presented as to the consonants articulation heterogeneity which is due to the tension. /10/

Yet there are no data available on vowels heterogeneity based on the feature in point. The consonants tension is assumed to result from their being voiceless/voiced; from the syllable being stressed/unstressed; from their position in a syllable or a word (opening/closing). /3, 1, 2, 4, 9, 10, 8/

Strong-ended consonants, voiceless and consonants in stressed syllables opening a word are believed to be more tense.

The KSUEPHL has designed a technique for the tongue pressure on the palate (palatum durum) to be investigated.

The power exercised by the tongue muscle is known as a mechanical one. Thus with consonants this power can be defined as the pressure upon some rigid surface. A technique combining tensometric processing with palatography and oscillography has been created to answer a number of points: how articulatory tension changes in lingual consonants within articulated syllables and words; the way the muscle tension manifests itself within a syllable; what the syllable peak is (whether it is a definite and the most tense point, the attainment of which is immediately followed by a relaxation, or whether it is a segment more or less elongated); the way the motor impulse of muscle tension is being produced, whether there is the incessability of the impulse and what it is manifested in.

The technique is called tensopalatography.

The technique in point (see first described - 7) makes it possible to correlate the physical characteristics of speech (of syllables, words and sense groups) to simultaneously registered articulatory organs movements. The oscillogram is recording not only acoustic signals but the tongue pressure impulses upon the roof of the mouth as well, which when analysed can appreciate the articulatory tension, the articulatory duration and compare these features with signals acoustic duration.

The sensory (sensing) elements in the pressure electric measuring elements were minute (2 mm base) wire tensometres. /7/

The measuring elements taring enabled us to value the tongue pressure impulses. The recording apparatus used was a rotating mirror oscillograph. The oscillograph record simultaneously showed pressure signals from two measuring elements and the speech acoustic picture. (Comp.12 and 5).

To choose indicator position on the palate plate there has been made a standardization of consonants and vowels articulatory contacts according to the Russian speech sounds palatography evidence taken from the KSUEPHL phonetical archives. The contacts in question were grouped into three types. Each of the three speakers, the participants of the experiment, had three separate palates specially made. Each of the palates with two (front and side) detectors attached to it served as an integrate detector (see fig.1) through the two channels of which the oscillograph record had signals registered. The registered tongue pressure signals looked like impulses having the front, the peak and the cut-off of their own. The impulses did not superimpose with acoustic signals boundaries.

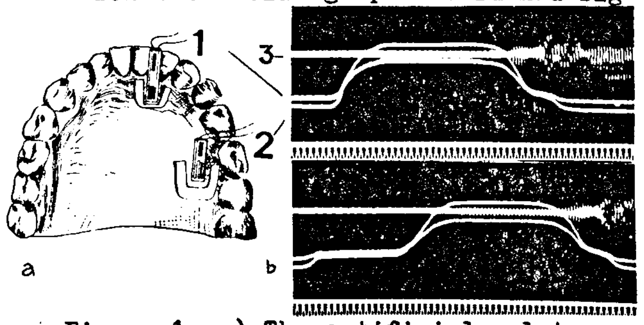


Figure 1. a) The artificial palate with the front and side pressure detectors; 1 - front detector, 2 - side detector. b) Tensooschillograph record of the syllables /t'a/, /d'a/; 1, 2 - pressure signal from the pressure detectors, 3 - acoustic signal from the microphone.

The impulse form and such of its parameters as amplitude and peak duration burdened with one or several pedestals, constantly varied due to the syllable components structure.

Different impulse forms were analysed; this resulted in descerning six impulse types: 1) a rectangular one having the peak length with relatively low amplitude; 2) a triangular one having a minimal peak duration, an increasing front and a descending cut off; 3) a bell-formed one also having a minimal peak yet having soft prominent front lines and cut off; 4) impulses with a complicated compound form having but one peak; 5) complex impulses with two peaks; 6) blending impulses having adjoining peaks. (See fig.2).

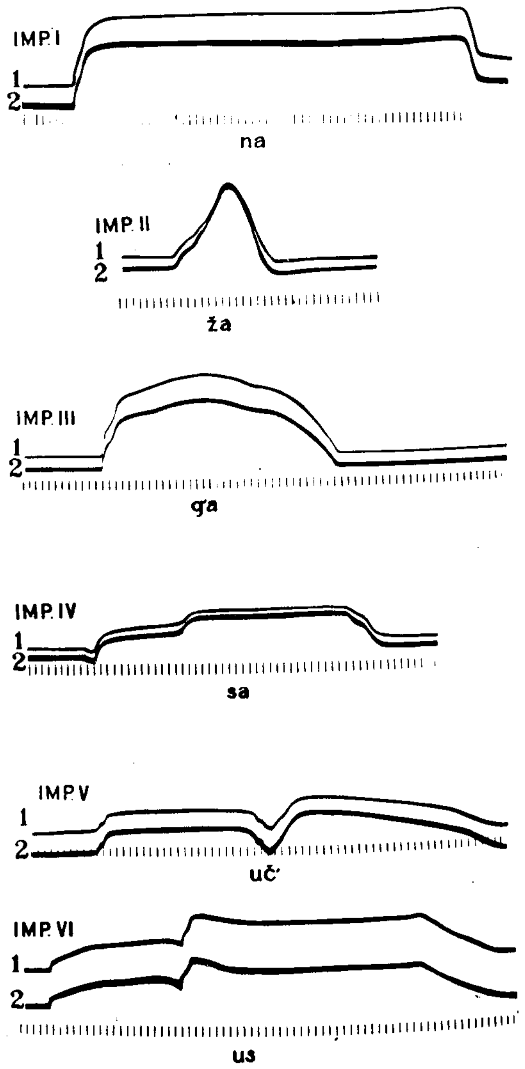


Figure 2. Types of the impulses.

Further tongue pressure impulses analysis and measuring made us have the parametres as follows: a front duration, a peak duration, an articulatory integrated duration, an impulse amplitude peak. The front duration and the pressure increase velocity were defined as correlated values: the less the duration the greater the velocity and consequently, the higher pressure at the syllable beginning.

The cut off duration determined the pressure descending time, tension heterogeneity/homogeneity at the juncture of a consonant and a vowel. The angle between the zero-th and the beginning front line as well as the front duration indicated the pressure progress speed; the angle growth corresponds to the pressure increase speed growth.

The impulse peak duration was interpreted in terms of: a) a peak line uniformity/non-uniformity; b) the duration of the level part of the peak; c) the existence of the rise peak followed by a lesser part of a peak duration.

The uneven peak line growth indicates the uneven pressure manifestation with its maximum observed within a relatively stable segment.

The technique serves for describing and measuring the boundaries correlation between the syllables and word articulatory and acoustic duration. The tensooscillograph record of syllables with initial stop consonants displays the development of their first components - consonants; their acoustic signal being registered in zero-shape. It makes it interesting to describe coincidence/noncoincidence of the final segment in the pressure impulse and the start of the acoustic signal for the syllables differing in modal indications: the voiced beginning syllables, voiceless beginning syllables and sonant beginning syllables.

The basic indicator employed was an acoustic signal. All the impulse pressure observed after the acoustic signal switched on were assumed to be retarding or slow and had a minus index during the evaluating procedure. Plus index was attached to the parametres being ahead of the signal attack. Thus, the technique makes a foundation to investigate the problem of interdependent articulatory and acoustic features. / 6/

Investigated were Russian vowel syllables and those with initial predorsal hard and dorsal palatalised stop consonants (the first type of impulses) (see fig.3). The analysis aims at having tension articulatory characteristics and its distribution within syllables, at the correlation of the articulatory and acoustic duration of syllables.

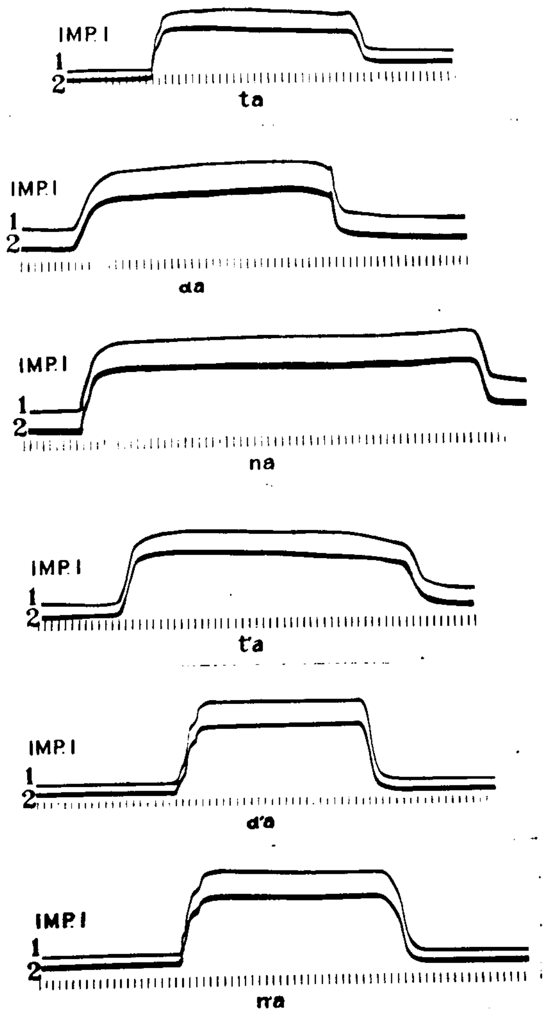


Figure 3. Impulses of the investigated syllables.

The syllable analysis based on impulses shape yielded the description to follow. The initial components of all syllables displayed on the oscillograph record took the shape of the first class impulses. Hard consonants syllables preferred the 1, 4 shape (rectangular impulse with a graduated front and a vertical cut off), syllables with initial palatalized stop consonants preferred the 1, 2 shape (rectangular with a stepped front and cut-off).

Initial voiceless hard syllables differed from those with initial voiced and nasal consonants by belonging to the 1, 4 class, while voiced and nasal consonants - to the 1, 2 class which indicates the lack of identity in the growth and decline of tension in the syllables.

The palatalized stops syllables and those with voiced and voiceless differed from hard consonants syllables by their belonging exclusively to the 1, 2 class; however the voiced and voiceless cut impulse was usually longer. The homogeniety in the difference of voiceless and voiced syllables for both tested groups (hard and palatalized) shows that voiced and nasal syllables differ from voiceless as having a special, smoother tension growth at the consonant-vowel transition. The tension growth in voiceless syllables with initial hard and palatalized consonants is of overfall nature - an abrupt transition at the consonant-vowel boundary.

The analysis of all syllables with hard and palatalized consonants with respect to their impulses shape enabled us to distinguish long syllables impulse manifestations with retaining one and the same amplitude and a relatively unchanging manifestation of peak duration. Here belong the initial voiceless syllables and those with hard and palatalized consonants. The second group contains the syllables with the initial voiced and nasal consonants having the impulse shape whose tension is found less stable concerning their peak duration manifestation and respectively, a short period of homogeneous amplitude.

The first syllables have more consonantal features, the second group is of mixed character, i.e. of a consonant-vowel nature. We can now state that the tension change (the overfall during consonant - vowel juncture) is more pronounced with the voiceless syllables and less pronounced with the voiced syllables and those with the sonants. The tension change feature manifests reciprocal derivational relations between the syllable components.

The analysis of voiceless syllables on the one hand and of voiced and sonant ones on the other hand presents them as having different growing modal features.

This brought to the measuring and analysis of the parameters describing the modal features growth: the front duration, the cut off duration, the amplitude value, the peak duration, the correlation between the peak duration and that of the amplitude; the correlation of the front duration and that of the amplitude; the correlation of the cut-off duration and that of the amplitude. Taken into account was the correlation of integral impulse length and the syllable acoustic length.

Initial voiceless hard syllables and those with initial palatalized consonants are different classes in terms of their absolute mean front length: the front length in hard syllables (HS) is constantly lesser than that in palatalized syllabless (PS) which shows a greater pressure growth in the initial segment of HS. The absolute value of HS mean cut-off duration is less than that of PS mean cut-off duration, the length of HS and PS cut-off being less than that of the front.

With PS the front is longer than the cut-off, while both segments impulses are longer than their HS counterparts. This evidence distinguishes syllables according to the degree of tension at the consonant-vowel juncture. With PS this is of less contrast nature yet having a relatively greater growth velocity, which can be read in the front-cutoff ratio (FC ratio). /F > C/

Relative values resulted from the comparison of the peak duration and the maximum amplitude of voiceless HS impulses and those of PS can be written as follows: $\frac{\tau_\ell}{J}HS > \frac{\tau_\ell}{J}PS$

The comparison of the front and the amplitude F/A and the cut-off and the amplitude C/A with HS and PS positively distinguishes HS into a type with a more tense first component growth, when combined with the vowels it becomes more prominent (contrastive). The latter characterizes occlusion as a consonant marker.

The analysis of the parameters realizing the tension dynamics describes the initial syllable components as a process capable to affect the acoustic signal duration.

There is a regular greater articulatory impulse and acoustic signal duration observed in PS, while HS show lesser articulatory and acoustic signals duration respectively. The second group of syllables with initial voiced consonants, hard and palatalized, and sonants, hard and palatalized, recorded with identic amplification and from the same artificial palate, was described similarly.

Tendencies discovered while analysihg voiceless syllables proved to be regular for syllables with initial voiced and sonorous consonants. First of all it has to do with the following parameters correlated:the front duration /the amplitude value; the cut-off duration / the amplitude value; the peak length / the amplitude value; a total articulatory duration / a total acoustic duration.

As it was already mentioned, the voicelees syllables are contrasted to those of the voiced and the nasal for having different modal dynamic features.

A common feature both for voiced and for sonant syllables impulses is the peak duration and the amplitude ratio. The ratio can be written as follows:

$$\frac{\tau_\ell}{J_{max}}(da) > \frac{\tau_\ell}{J_{max}}(na)$$

Similar exponents ratio of the parameters in point were found when D'A and N'A were compared. The parameters comparison show greater ties of the hard voiced and the palatalized voiced consonants with a vowel to follow.

The tension of the voiceless stops, of the voiced consonants and the nasal sonants is growing in different ways, which displays both intrinsic features of each consonant and the syllable features as the articulatory entireties. The greatest liasion of the components is observed in the voiced syllables (which is indicated by the cut-off value and the amplitude cut-off ratio).

The voiced and the nasal syllables are contrasted to the voiceless ones as a special modal class of syllables, the chief feature of which being realized in greater derivational ties between the components.

On the oscillograph records of the syllables with initial voiceless, voiced and nasal consonants different types of relations between the articulatory peak duration and the amplitude peak (see A); between an integrate duration of the tension impulse and the acoustic duration signals (see B) is regularly indicated as follows:

A $\frac{\tau_\ell}{J_{max}}(ta) > \frac{\tau_\ell}{J_{max}}(da) > \frac{\tau_\ell}{J_{max}}(na)$

B $\frac{\tau_{art.}}{\tau_{ac.}}(ta) > \frac{\tau_{art.}}{\tau_{ac}}(da) > \frac{\tau_{art}}{\tau_{ac.}}(na)$

The most autonomous (see A and B) are the initial voiceless components; the voiced and the sonant syllable enable us to assume a noncontrastive, relative homogeniety of the components (see above C/A; F/C) which brings about a greater interdependence between them and is manifested in the vowel duration growth. The similar relations are likely to have resulted from its greater articulatory tension. The intersyllabic relations between the syllable components is based on the feature of a higher or lower homogeniety of the articulatory tension development. Therefore the syllable is articulated as a naturally organized integrity of interdependent components. The syllable has an agregate factor manifesting itself in a greater or lesser similarity of the syllable components based on the muscular tension, which fins its expression in a specially structured impulse of the tongue pressure.

There is born a possibility to classify syllables in terms of relations of their components. In the class of syllables with hard consonants it looks as follows: DA > NA > TA.

The analysis evidence suggests that the tensopalatography is suitable for studying the tension feature in its dynamic manifestation within the articulated syllable, word and sense group. The factor working within the syllable and uniting its components undoubtedly proves not only the articulatory entirety of the syllable, but also its predetermination in speech production.

The table of syllable tension impulses. Parameters and ratios

Syllables Parameters	ta	da	na	t'a	d'a	n'a
Front (F) msec	36	59	61	64	91	118
Cut-off (C) msec	34	50	52	58	65	65
Amplitude (A) mm	43	42	41	37	36	32
Articulatory duration of signal (D art.) msec	348	325	325	352	326	339
Acoustic duration of signal (D ac.) msec	334	438	420	310	452	448
Parametres ratios						
Front:Cut-off	1,1	1,2	1,3	1,1	1,4	1,9
Front:Amplitude	0,8	1,4	1,5	1,7	2,5	3,7
Cut-off:Amplitude	0,8	1,2	1,3	1,6	1,8	2,1
D art.: D ac.	1,0	0,7	0,8	1,1	0,7	0,8

Note: the table gives statistic data obtained from tensooscillograms with constant amplification on the same artificial palate of one speaker.

REFERENCE

I. Абеле А. К вопросу о слоге. "Slavia", Ш, I924, с.I-34.

2. Богородицкий В.А. Фонетика русского языка в свете экспериментальных данных. Казань, I930. - 356 с.

3. Бодуэн де Куртенэ И.А. Введение в языковедение. В его кн.: Избранные труды по общему языкознанию. М., I963, т.2, с.246-293.

4. Зиндер Л.Р. Общая фонетика. М., I979. - 3II с.

5. Кузьмин Ю.И. Динамическое палатографирование. - Вопросы психологии, I963, № I, с.I37-I4I.

6. Лийв Г., Эзк А. О проблемах экспериментального изучения динамики речеобразования: комплексная методика синхронизированного кинофлуографирования и спектографирования речи. - Изв. АН Эст.ССР, I968, т.I7. Биология, с.78-I02.

7. Рузга З. Электрические тензометры сопротивления. М., I964, с.I5-44.

8. Рушковская Л.М., Скалозуб Л.Г. Артикуляция звонких и глухих согласных в слогах русской речи:(По данным кинорент-

генографирования). - Русское языкознание, Киев, I98I, вып.2, с.94-I00.

9. Скалозуб Л.Г., Лебедев В.К. Тензометрирование как прием исследования давления языка на нёбо при речи. В кн.: Механизмы речеобразования и восприятия сложных звуков. М.-Л., I960, с.56-62.

I0. Щерба Л.В. Фонетика французского языка. М., I953. - 3II с.

11. Sawashima M.Temporal Patterns of Articulatory and Phonatory Controls. - Ann.Bull.RILP, 1979, 3, p.1-13.

12. Stetson R.H., Hudgins C.V. and Mosos E.R. Palatograms change with rates of articulation, 1940, Arch.neerl.phonet. exper., 16: 52-62.

COMPLEX SIGNAL REFLECTION IN THE PERIPHERAL PART OF THE HEARING SYSTEM AND DESCRIPTION OF PHONETIC ELEMENTS

LYUDMILA BABKINA
Dept. of Physics
Leningrad State
University
Leningrad, USSR

BORIS DENISON
Dept. of Physics
Leningrad State
University
Leningrad, USSR

SVETLANA GOROKHOVA
Leningrad
Polytechnic
Institute
Leningrad, USSR

ALEXANDR MOLCHANOV
Dept. of Physics
Leningrad State
University
Leningrad, USSR

ABSTRACT

Frequency structure of signal reflection in the peripheral part of the human hearing system is evaluated in terms of the combined cochlear potential observed at the ear-drum level. The reflection appears to include components missing in the signal spectrum. The explanation proposed implies the possible effect of a hearing feedback which, unlike the hearing reflex, provides for the appearance of signal envelopes propagating along the cochlear partition as separate waves.

The study of signal processing in the peripheral part of the hearing system (PPHS) is essential for getting an insight into the mechanism of human sound-information perception. Complex signal reflection in PPHS is of particular importance. Here signal reflection will be defined as a spatial distribution of exciting effects along auditory-nerve-fiber endings, formed as a result of the signal transformation by hearing mechanisms, allowing for feedback effects.

Until recently feedback mechanisms had been overlooked in simulating signal transformation processes in PPHS. The implications were that a result of signal processing in PPHS is a frequency-coordinate transformation similar to spectral analysis which is correlated with the excitations of auditory-nerve-fiber endings. A reflection of this type is also extensively used in phonetic studies in the form of dynamic spectrograms.

Recent electrophysiological experiments, however, have provided evidence for the propagation of vibrations, corresponding to complex-signal combination tones even at low stimulation levels, in the cochlear hydrodynamic system /7/. The fact that frequency components missing in the signal spectrum may appear in the signal reflection is incompatible with the idea of PPHS as a linear system which deals only with separating the signal into frequency components.

In the literature available combination frequency vibrations are often viewed as a product of signal distortion in its non-linear transformation in the cochlear vibration system. However, experiments with narcotized animals involve certain difficulties in determining the informational significance of the combination vibrations observed. To investigate the role of combination vibrations in signal reflection in PPHS it is necessary that the fact of their existence should be established and their level estimated. When using phonetically meaningful sounds as stimuli, the existence of a certain component in the reflection can be correlated with a certain characteristic of its perception. Of particular importance is to establish that vibrations with frequencies missing in the signal spectrum do exist in human PPHS, and to lay down a model of the mechanism causing their occurrence.

In this study the method of electrocochleography involving analog and digital accumulation was used in combination with fast Fourier transform /4, 8, 3/ to obtain combined cochlear potentials (CCP) and to analyze the frequency structure of vibrations in the human cochlea.

Assuming that receptor structures of organ of Corti interact in an electromechanical way with the cochlear hydrodynamic system, a variable component of combined cochlear potentials is considered to reflect the motion of cochlear mechanical structures under the effect of the stimulus or vice versa /6/.

The experiment was intended to identify, in the signal reflection in PPHS, the components missing in the sound stimulus spectrum by means of analyzing the CCP appearing at the human ear-drum under the effect of a complex sound stimulus.

Fig.1 shows two-tone stimulus spectrum (I) and typical CCP spectra successively for one subject, given two values of volume of sound.

Fig.2 shows the spectrum of vowel "a" (I) and the CCP spectrum (II) for the same subject. The comparison of the stimulus spectra with the CCP spectra reveals that the latter include components missing in the former. With a two-tone stimulus, a component of this kind is primarily the f_1-f_2 frequency component. The level of the newly appearing components has a value close to that of the level of response to spectral components present in the spectrum.

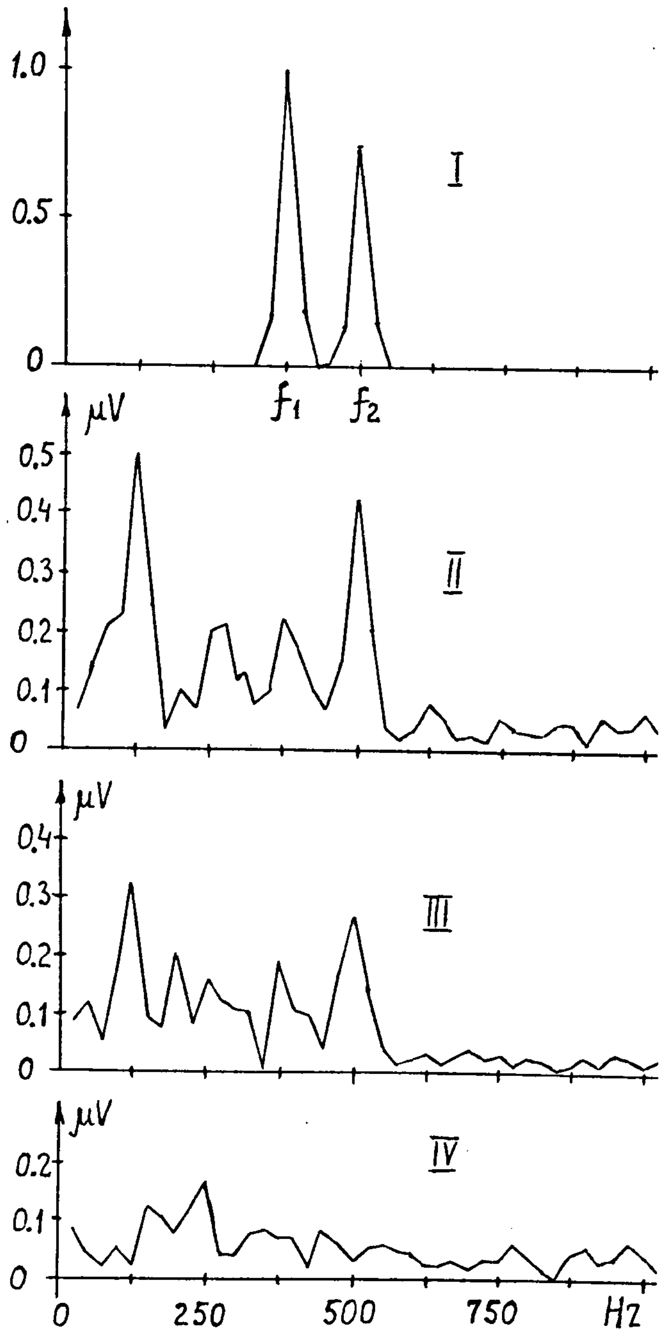


Fig.1. Two-tone stimulus reflection in the spectrum of CCP measured at the human ear-drum.
I - two-tone stimulus spectrum; II - CCP spectrum at the 100 db SPL stimulus level; III - CCP spectrum at the 85 db SPL stimulus level; IV - spectrum of noises measured at the human ear-drum level in an analogous accumulation mode. The pattern of the stimulus spectrum is shown in relative normalized counts in Y-axis. Quantization range - 156.4 mcsec.; number of counts in a sampling - 256; number of accumulated samplings - 1024.

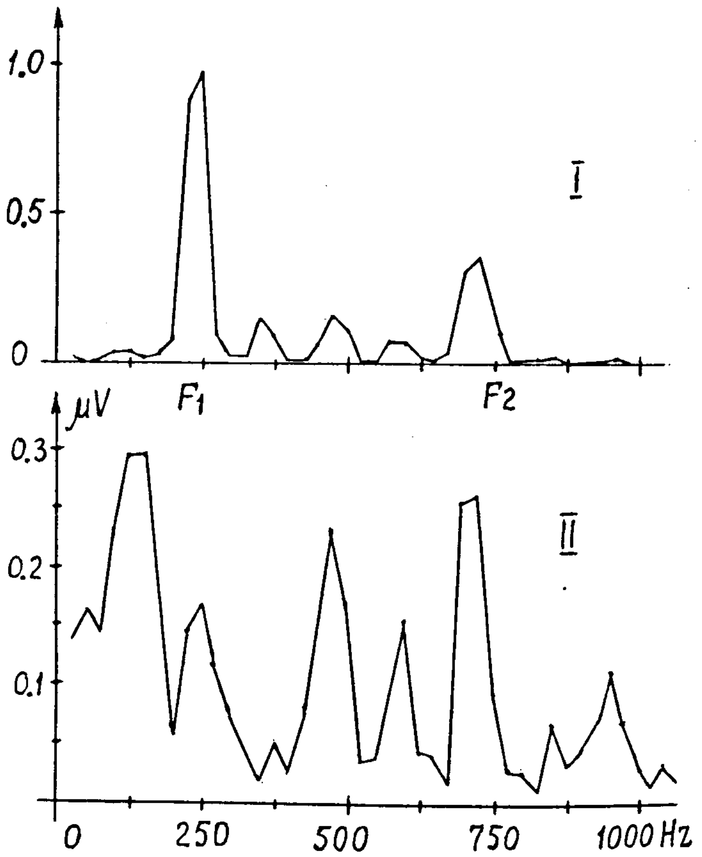


Fig.2. Vowel spectrum reflection in the spectrum of CCP measured at the human ear-drum.
I - spectrum of vowel "a"; II - CCP spectrum at the 95 db SPL level of volume of sound. The measuring conditions are identical to those listed in the caption of fig.1.

Fig.2 demonstrates that the general pattern of the spectrum of response to a vowel is significantly different from that of the spectrum of the vowel presented at the input of the human hearing system.

The most convenient way of discussing the results obtained is to make use of the model of signal transformation in PPHS. The functional structure of such a model was described in /10, 11/. Compared to the earlier models of signal transformation in PPHS /3/, the model under discussion includes a mechanism realizing the feedback which significantly affects signal reflection in PPHS.

Consider the possible properties of the mechanism in question.

The possibility that in addition to the feedback circuit ensuring hearing reflex there exists in PPHS a feedback effected along the signal envelope was first suggested in /10/. The mechanism realizing the latter feedback was termed "hearing feedback". It was also shown that the action of this mechanism may account for the effects such as residual tones and inhibition of the first harmonic of microphonic potential /12/.

It is obvious that the inhibition of the first harmonic of microphonic potential /12/ can be accounted for by the existence of the hearing feedback, provided the value of a difference-frequency component stipulated by its effect is comparable to that of the response to the first harmonic of the stimulus. The experimental results shown in fig.1 indicate that the amplitude of the f_2-f_1 frequency component of CCP spectrum and that of the f_1 frequency component of CCP spectrum are values of the same order of magnitude.

Thus, experimental evidence has been obtained for the assumption that the inhibiting effect may be accounted for by the effect of the difference-frequency component of CCP spectrum. Again, the value of this component being great, it is possible to assume that its informational significance is by no means less than that of the CCP spectrum components caused by the effect of those components which are present in the stimulus spectrum. Accordingly, a similar explanation of residual tone perception is available.

The fact that the relative amplitude of the CCP spectrum component resulting from the effect of the stimulus whose spectrum does not include such component is not dependent on the stimulus level indicates that the component in question is caused by the action of a specialized parametric mechanism dealing with separation of the signal informational characteristics rather than by non-linear distortions in transforming the signal in PPHS.

A problem to be solved concerned experimental identification of the paths taken by the signal envelope to get back to the analyzer part of PPHS, i.e. to the cochlea, upon being formed. One possibility suggested in /10, 11/ was the "cochlea - receptor cells - auditory nerve - facial nerve - stapes - cochlea" circuit. The newly obtained experimental data make it possible to consider the "cochlea - receptor cells (acting as envelope extractors) - cochlea" circuit as well.

Upon getting to the inner ear by either way, the envelopes are propagated along the basilar membrane and form the maximum deflection at a corresponding point, thus producing a new channel where a new envelope can be extracted whose variable component will again pass along the feedback circuit and will be summed up with other envelopes etc. until a dynamic equilibrium reflection of the stimulus is obtained. Thus, the hearing feedback model appears to be an integral part of the model of PPHS analyzer part and the whole system should be viewed as a parametric non-linear signal analyzer, with its characteristics depending, alongside with other factors, on the type of signals being analyzed. Realization of the hearing feedback model requires concrete definition of the envelope, formulation of the rules of its formation and introduction of PPHS in the analyzer part of the model.

A possible technical realization of the hearing feedback model is described in /1/. As follows from the fundamental scheme of the model/1/, the output signal reflection will include frequency components missing in the analyzed signal, their frequency values characterizing the mutual disposition of the signal spectral components. Occurrence of reflection components resulting from secondary interactions is also possible.

Correlating vowel spectra to the frequency structures of their reflections in PPHS shown in fig.2, it can be seen that the latter include spectral components missing in the stimulus when the frequencies of components in the stimulus spectrum are close enough. Thus, the CCP spectrum of vowel "a" includes F2-F1, F2+F1 frequency components.

The foregoing implies that envelope extraction in non-linear analyzer channels is significantly affected by a frequency-selectivity formation mechanism referred to in the literature as that of sharpening of cochlear gain-frequency characteristics (GFC). As stated above, the earlier studies /3/ make it possible, by using non-linear transformations, to lay down a model ensuring a sufficient degree of cochlear GFC sharpening to account for the difference between the shape of auditory-nerve frequency-threshold curves and GFC of cochlear hydrodynamic system.

Recent experiments /13/ have demonstrated, however, that at low signal levels the cochlear GFC themselves appear to have a shape close to that of auditory-nerve frequency-threshold curves.

The only seemingly possible way of accounting for the above effects is to assume the existence of an electromechanical interaction of receptor cells with cochlear vibration systems, presuming the interaction to form local feedbacks of quick-response leading to regeneration processes.

Alongside with the new experimental evidence, a model of the sharpening mechanism must also make allowance for the whole complex of properties recognised in the earlier studies of signal processing in PPHS disregarding feedback effects.

The basic principles of a model of cochlear GFC sharpening mechanism amount to the following.

1. At low vibration levels cochlear GFC are to be close to auditory-nerve frequency-threshold curves.
2. At high vibration levels cochlear GFC are to be close to those measured by von Békésy.
3. The structure of spatial-frequency signal reflection in PPHS is characterized by the location of auditory-nerve fibers with given characteristic frequencies in the low-frequency slope area of the amplitude-coordinate characteristic of the basilar membrane /6/.
4. The effect of one harmonic signal involves an increase of neuron pulsation frequency above the threshold value only in a relatively narrow range near the values of the signal frequency close to the characteristic frequency.
5. With the effect of two signals, one being tuned to the characteristic frequency of the neuron observed and the other being a test signal, neuron pulsation frequency at low test-signal intensities is considerably higher than the spontaneous one throughout the range of test signal retuning.
6. The increase in test signal intensity with certain kinds of detuning is accompanied by the formation of inhibition areas. The width and depth of the areas increase with an increase of test signal intensity.
7. The inhibition areas are asymmetrical in relation to the characteristic frequency, being deeper towards the high-frequency region.

The above requirements are met by the model of PPHS GFC formation which includes a frequency-coordinate transformer /5/ with a frequency-dependent voltage transformation device /2/. The degree of feedback can be controlled as described in /9 /. A calculation has revealed that the scheme allows sharpening of PPHS GFC by a factor of 20 to 24, while preserving a phase characteristic close to the linear one.

From the above considerations the following conclusions can be drawn. Signal reflection in PPHS appears to be a result of both a complex interaction of non-linear mechanisms of vibration processing in the inner ear and the effect of feedback circuits due to electromechanical interaction of receptor systems of organ of Corti with cochlear partition vibration system, as well as of the circuits realizing hearing feedback. Since the formation of signal reflection in PPHS involves the appearance of components missing in the spectrum of the stimulus signal and may be accompanied by secondary interaction of these components, one should expect the reflection to differ considerably from the stimulus spectrum, particularly with speech signals whose form is fairly complex.

REFERENCES

/1/ L.N.Babkina, A.P.Molchanov. Spectrum analyzer. Patent 792172. Bulletin No.48. 1980 /in Russian/.

/2/ L.N.Babkina, A.P.Molchanov. Frequency -dependent voltage transformation device. Patent 1275316. Bulletin No.45. 1986 /in Russian/.

/3/ L.N.Babkina, A.P.Molchanov, T.I.Tereshchuk. Mathematical models of signal transformation in the peripheral part of the hearing system. In Sensornye sistemy, Leningrad, 1982 /in Russian/.

/4/ L.N.Babkina, A.P.Molchanov, T.I.Tereshchuk. A comparative analysis of the gain-frequency characteristic of human microphonic potentials obtained experimentally and by mathematical modelling. - Fiziologiya cheloveka, 1983, No.2, 223-231 /in Russian/.

/5/ O.Ye.Buslovsky, V.K.Labutin, A.P.Molchanov, Ya.I.Panova. Frequency-coordinate transformer. Patent 223164. Bulletin No. 2. 1968 /in Russian/.

/6/ H.Davis. An active process in cochlear mechanics. - Hearing Research 9 (1983), 79-90.

/7/ G.L.Gibian, D.O.Kim. Cochlear microphonic evidence for mechanical propagation of distortion products (f_2-f_1) and ($2f_1$-f_2). - Hearing Research 6 (1982), 35-39.

/8/ K.W.Humphris, P.B.Ashcroft. Extratympanic electrocochleography. - Acta Oto-Laryngologica, 1977, v.83, No.3-4, 303-309.

/9/ A.P.Molchanov. Adaptable supergenerator. Patent 355730. Bulletin No.31. 1972 /in Russian/.

/10/ A.P.Molchanov, L.N.Babkina. On the possibility of existence of a hearing feedback in the peripheral part of the hearing system. - Doklady AN SSSR, 1977, No.4, 958-961 /in Russian/.

/11/ A.P.Molchanov, L.N.Babkina. Electric models of cochlear mechanisms, Leningrad, 1978 /in Russian/.

/12/ N.N.Sanotskaya. Microphonic potentials of cat's cochlea in two-tone harmonic signals of different phase spectra. - Fiziologicheskij zhurnal im. Sechenova, 1977, No.7, 976-983 /in Russian/.

/13/ P.M.Sellick, R.Patuzzi, B.M.Johnstone. Measurements of basilar membrane motion in the guinea pig using the Mössbauer technique. - JASA 1982, v.72, No.1, 131-141.

STATIC AND DYNAMIC STRUCTURE OF VOWEL SYSTEMS

L.F.M. ten Bosch L.J. Bonder L.C.W. Pols

Institute of Phonetic Sciences, University of Amsterdam

ABSTRACT

A phonetic/phonological model has been developed for describing the structure of natural vowel systems in terms of configurations consisting of N points in the formant space. These configurations (abstract vowel systems) are defined as solutions of an optimalisation algorithm. This search algorithm uses an optimality strategy that is based upon two extra-linguistic principles, one dealing with the articulatory effort, the other with perceptual ease. The model is evaluated by comparing the model results with available phonological data.

INTRODUCTION

The model that we present is developed in order to find basic structure principles underlying the architecture of vowel systems. It uses as a starting-point the dispersion model of Liljencrants and Lindblom (1972). They tried to describe natural vowel systems by maximizing an acoustic distance measure between N points, all of them positioned within a predefined fixed region in the formant space. The novelty of the present model is the extension of the acoustic principle (with respect to vowel dispersion only) with an articulatory minimal effort principle.

In the following three sections, we will gradually unfold the model. Section 1 poses the two basic structure principles we are using. Section 2 describes the model itself: 2.1 deals with the technical translation of the basic principles into an appropriate mathematical formulation and a search algorithm for the abstract vowel systems; 2.2 describes the comparison of these abstract systems with the vowel systems from natural languages; and 2.3 will briefly deal with the implementation of *dynamic* aspects of vowel systems: the long/short-opposition and the diphthongs. In section 3 we will give a summary of the present results. In section 4 we conclude with a discussion.

1. THE PRINCIPLES

We use two principles dealing with the structure of vowel systems which are supposed to be of primary importance:

(a): *minimality of effort* of (static) vowel pronunciation;

(b): *minimality of inter-vowel confusion*.

Vowel systems are said to be 'optimal' if they optimally satisfy both principles simultaneously.

Evidently, the consequences of these principles separately are conflicting: (a) yields minimal overall articulatory vowel distances, whereas (b) leads to maximal inter-vowel distances. In order to be able to handle both principles in an appropriate way, they have been translated into specific mathematical formulae. Some of these formulae directly deal with both the formant position of vowels and the vocal tract area function, other ones are based upon arguments concerning probability and optimalisation techniques (see section 2.1, the search algorithm).

2. THE MODEL

2.1. The Search Algorithm

Each vowel system is represented as a point in a so-called 'state space', in which principles (a) and (b) define an optimality strategy. The search for optimal vowel systems can be considered as looking for stable solutions in this state space. In order to specify the search algorithm, we introduce the following formulae (classified into basic, derived and evaluational ones):

2.1.1. *basic formulae*

These formulae play the most elementary role in the model.

The *inter-vowel acoustic distance* dF between v1 and v2 is defined as follows:

$$(dF)^2 = (\log(F_1(v1)) - \log(F_1(v2)))^2 + (\log(F_2(v1)) - \log(F_2(v2)))^2 \qquad (1)$$

Only the relative positions of vowels in a vowel system are relevant. The logarithms of the frequencies are used to meet with the perceptual behaviour of the basilar membrane. This closely relates d_F to empirically determined acoustic distance measures involving mel or bark scales.

The expression for the *inter-vowel confusion probability* p(v1, v2) reads:

$$p(v1,\ v2) = \exp(-\alpha * d_F(v1,\ v2)) \qquad (2)$$

α being a positive scaling parameter.
Before actually evaluating vowel systems we first introduce the following probabilistic concept. We hypothesize an exponential relation between the inter-vowel confusion probability p and the inter-vowel acoustic distance d_F. This relation can be globally verified by inspecting the perceptual vowel confusion matrices in several languages.

We define the *articulatory effort* d_A:

$$d_A = \sum (Si - 1)^2 \quad (i = 1,\ \ldots,\ 4) \qquad (3)$$

This expression relates the shape of the vocal tract (which is approximated by the straight 4-tube, consisting of 4 segments of equal length with areas Si (cf. [1], [2])) to an articulatory effort value (see figure 1).

2.1.2. *derived system formulae*
In order to be able to define the structure principle for vowel systems as a whole, we introduce the system counterparts of d_A and d_F.

The expression for the *total articulatory system effort* D_A reads:

$$D_A = \max(d_A) \qquad (4)$$

The articulatory effort value of a vowel system is defined as the maximal value of the articulatory effort values of its members.

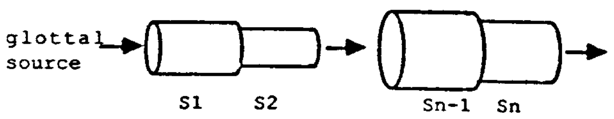


Fig 1. An example of a general n-tube with segment areas Si.

The total *perceptual system discriminality* D_F will be

$$D_F = \prod (1-\ p(v_i,\ v_j)) \qquad (1 \leqslant i < j \leqslant N) \qquad (5)$$

1- p(v1, v2) denotes the probability of vowel v1 and vowel v2 not being mutually confused. Therefore D_F is a measure for the total discriminality of an N-vowel system. Consequently we have $D_F = 1$ in case of perfect discriminality and $D_F = 0$ in the worst case.

2.1.3. *evaluation formulae*
We have to minimize the articulatory effort D_A and to optimize the discriminality measure D_F simultaneously. Therefore we introduce the *penalty parameter* Q relating both aspects:

$$Q = (D_A)^2 + S * (D_F - 1)^2 \qquad (6)$$

This type of expressions is well-known from optimality theory and is in fact a natural choice here. Indeed, minimization of Q logically implies minimization of D_A towards zero and optimization of D_F towards unity simultaneously. The rate of convergence of this process is controlled by the slack variable S (S being a large positive number). Optimal vowel systems are locally found by iteratively improving the position of all vowels in the system while decreasing the value of Q.

2.2. Evaluation Part
The evaluation part of the algorithm described above in fact consists of a measurement of the goodness of fit of the acoustic model output in relation to the more phonologically specified data from language databases ([3], [4]). For the time being we confine the evaluation to vowel systems without dynamic structure (without short/long opposition, without diphthongs). Presently, these latter effects contribute less to a general insight as they are second-order consequences.
In the model a method is implemented for actually effectuating the phonetic/phonological comparison. It is based upon essentially the same probabilistic motivations as already used in formula (5). The result of the comparison is expressed in terms of the *similarity probability* (denoted SP) of the respective abstract phonetic vowel system and a phonological system after having optimally paired each unlabelled v_i in the model system with a vowel v_j in the phonological reference system.

$$SP = \prod \exp(-\alpha * d(v_i,\ v_j)) \qquad (7)$$

If SP = 1, the similarity is perfect. The model evaluation now consists of the evaluation of all SP values between a model solution containing N vowels and all known phonological N-vowel systems. The present result of this evaluation is plotted in figure 2.

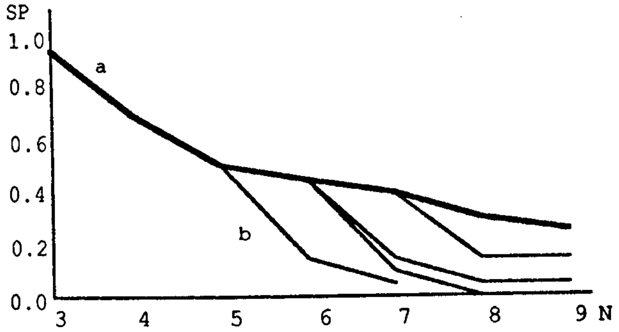


Fig.2. Goodness of fit of the present model in terms of the SP value.
The heavy line a connects all the found maxima, b shows some possible ramifications. N denotes the number of vowels in the model system and the phonological reference system.
One observes the decreasing SP value for increasing values of N. Probably this phenomenon can be traced to
- the declining fit of the model itself
- the increasing number of linguistic possibilities for large N.

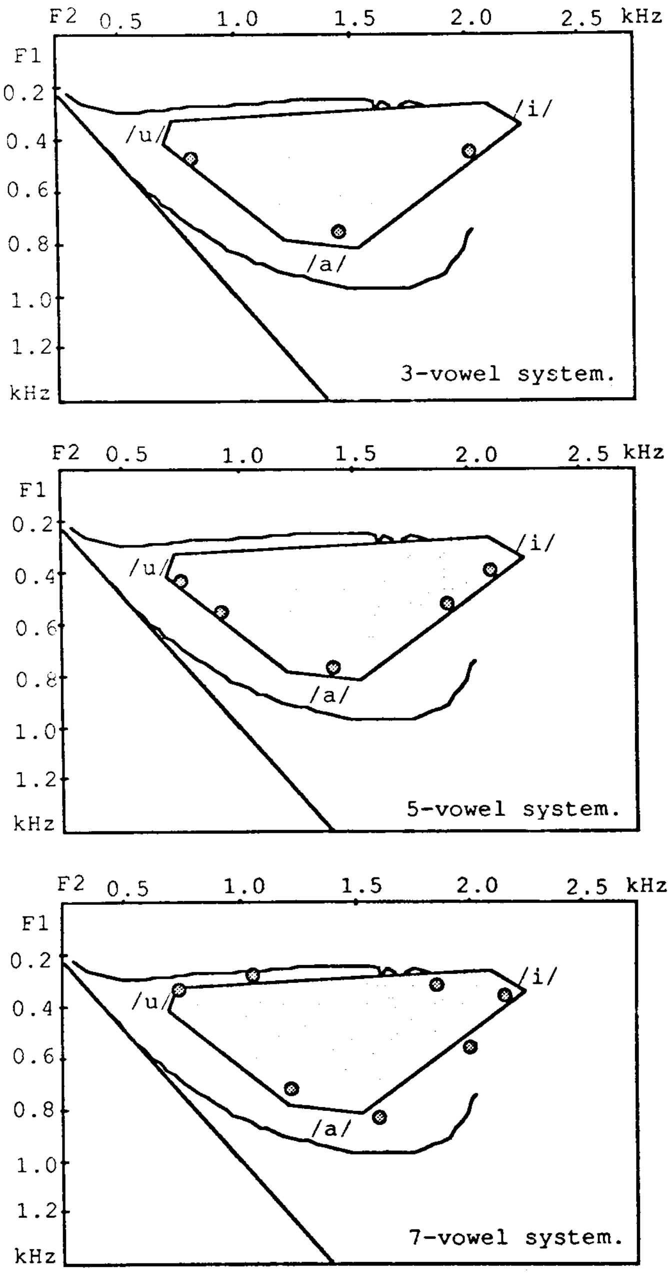
2.3. Dynamics
The description of the dynamic part of vowel systems appears to involve more linguistic details then are contained in the model described above. The model has proved to be inadequate for predicting actual diphtongs and long vowels in a specific language, but it merely defines and bounds the set of physical possibilities out of which a language may select.
In order to study these possibilities in more detail we use a *vowel structure matrix* of which the entries represent the long vowels and diphthongs. The short vowels constitute the elements along the two axes. Evidently, long vowels emerge as geminates along the main diagonal and diphthongs off the diagonal. In order to evaluate the entries we considered the acoustic gain relative to the articulatory effort. We give the results of such a calculation in figure 3. One may observe a preference for diphthongs to start in the /a/-region (i.c. to show decreasing first formant frequency).

ɔ	0.5	0.3	0.3	0.6
I	0.6	0.5	0.7	0.4
ε	0.4	0.5	0.4	0.3
α	0.7	0.3	0.5	0.4
	α	ε	I	ɔ

Fig 3. Gain of acoustic contrast in relation to articulatory transitional effort. The transitions are now described as concatenations of two short vowels out of the indicated set of four short vowels. Horizontally, we denote the vowels in initial position and vertically the short vowels in final position are shown. All entries (quotients of acoustic contrast and articulatory effort) have been rescaled to values between 0 and 1. They give an indication of the preference of the corresponding combination of short vowels.
In this four-vowel system the preference for transitions to start in the a-like region of the formant space is demonstrated by the values figuring in the first column relative to those in the other columns. The overall-preference for gemination can be deduced from the values along the diagonal. In general it does not have to be the case that these geminates correspond to actual long vowels such like /a/, /e/ etc. This identification is in fact a phonological item. The quotients have been specified up to only one decimal place in order to express their tentative character. They only have relative significance.

3. RESULTS OF THE MODEL

In the figures 4, 5 and 6 we give the present model solution in case of N = 3, 5, and 7 respectively. The closed contours represent contour lines of the articulatory effort function d_A. One observes:
- the preference for the vowel /a/, followed by /i/ and /u/;
- the preference for vowels along the lines /a/-/i/ and /a/-/u/;
- the limitation of the available vowel space without predefining a fixed boundary in the formant space.



Figures 4, 5 and 6 show the model solution in the formant space. For reference the grey area indicate the region which is used by most languages. The staight line denotes the line F1 = F2. The other two lines are contour lines of the articulatory effort function dA, which gives an idea of the theoretically shaped vowel space by using an effort principle (see the text). In case of the 7-vowel system, some of the vowels are positioned outside the grey area, as a consequence of the subtile imperfection of the balance between the two principles (a) and (b) (see the text).

4. DISCUSSION

In our project, we explicitly deal with the model in relation to other recent vowel dispersion theories as well as with recent improvements. The present results have led to the following two suppositions:

a) natural vowel systems may adequately be considered as derivations of specific 'abstract' vowel systems, while
b) the structure of these abstract vowel systems is defined by two extra-linguistic principles:
   - reduction of perceptual vowel confusion probability and
   - reduction of articulatory effort.

The present model certainly does not pretend to be the final answer to the question of the structure of vowel systems in general but it may stimulate a further fundamental approach to the subject. In our presentation we will briefly mention some of the parallels with recent phonological theories, e.g. [5]. Our model does not predict all linguistic details of vowel systems as it is not based upon such linguistic or other language-sensitive principles. However, some important tendencies are clearly demonstrable: tendencies in the appearance and behaviour of vowel systems are described by combining a few, indeed simple arguments concerning articulation and perception. The main question will be the search for a convincing theory relating vowel systems as they are actually observed on the one hand to the results of a stipulative or normative model at the other hand.

ACKNOWLEDGEMENTS

This project is sponsored by the Netherlands Organization for the Advancement of Pure Research ZWO (project no. 300-161-030).

REFERENCES

[1] Dunn, H.K. (1950). The calculation of vowel resonances, and an electrical vocal tract. J. Acoust. Soc. Amer., vol. 22., pp. 740-753.

[2] Bonder, L.J. (1983). The n-tube formula and some of its consequences. Acustica vol. 52, pp. 216-226.

[3] Crothers, J. (1978). Typology and universals of vowel systems. In: Universals of human language. Ed. J.H. Greenberg. Vol. 2. Stanford University Press, Calif., pp. 93-152.

[4] Maddieson, I. (1984). Patterns of sounds. Cambridge studies in speech science and communication. Cambridge University Press.

[5] Kaye, J., Lowenstamm, J. and Vergnaud, J.-R. (1985). Vowel systems. Paper presented at the 1985 GLOW colloquium, Brussels.

AN ARTICULATORY DYNAMIC MODEL FOR DIPHTHONGS AND TRIPHTHONGS IN CHINESE

Yang Shun-an

Institute of Linguistics, Chinese Academy of Social
Sciences, No.5 Jianguomennei Dajie, Beijing, China

ABSTRACT

The present paper describes the Exponential Dynamic Model for compound vowels such as diphthongs and triphthongs. With this Model, actual formant frequencies of all the allophones occurred in different phonological and phonetic contexts can be generated. The 9 diphthongs and 4 triphthongs in Standard Chinese constituted by 30 allophones can thus be generated with the target values of 6 phonemes.This Model is applicable to speech synthesis, so that data memory size can be decreased, and both intelligibility and naturalness of the synthesized diphthongs or triphthongs can be improved.

INTRODUCTION

The changing of sound color in compound vowels like diphthongs and triphthongs is mainly produced by the continuous movement of the speech articulators, i.e. by the continuous movement of the vocal tract. According to the acoustic theory of speech production, a given set of formant frequencies correspond to a given shape of the vocal tract. Therefore,the time-varing characteristics of formants can reflect the dynamic features of the compound vowels. Because of the practical need in speech synthesis and automatic speech recognition, it is necessary to formulate a functional model for describing the time-variation of the formant frequencies in dynamic vowels. And only after the formulation of such a model can we discuss the process of transformation between the discrete speech code and the continuous speech sound waves.

This paper proposès an Exponential Dynamic Model based on the analysis of the formant frequency data of the 9 diphthongs and 4 triphthongs in Standard Chinese. Parameters for the Model were obtained through analysis-by-synthesis, and the dynamic trajectories of formant frequences are in close approximation with the observed data. The utilization of this Model in the Synthetic System for Standard Chinese has both improved the quality of the synthetic sound and reduced the memory size for the synthetic parameters.

FORMULAS OF THE EXPONENTIAL DYNAMIC MODEL

The observed time-varing trajectories of the formant frequencies indicated that the formant frequencies of a diphthong are constantly changing from one set of target values to another set, and the overall tendency of such dynamic trajectories is to have relatively stable parts at the beginning and the end of the vowel and to change rather abruptly at the transitional part. And, compared with the typical formant values of the phonemes composing a given diphthong, the starting and ending frequencies of the formants are only approaching the target values rather than actually reaching them. This condition is very like a curve obtained by joining two reverse e exponential functions. We thus hypothesise that a formant trajectory of a given diphthong can be approximated by the following formulas (Fig.1).

$$\left.\begin{array}{l} F(t)=Fc+0.5S*Fd\{1-EXP[-\alpha(t-t0)S]\} \\ \\ Fc=0.5(Fb+Fe) \\ Fd=Fe-Fb \\ S=1 \quad (t-t0\geqslant 0) \\ S=-1 \quad (t-t0<0) \end{array}\right\} \quad (1)$$

Here,
Fb is the beginning target value;
Fe is the ending target value;
t is normalized time;
t0 is the time of division; and,
α is the factor of transitional rate.

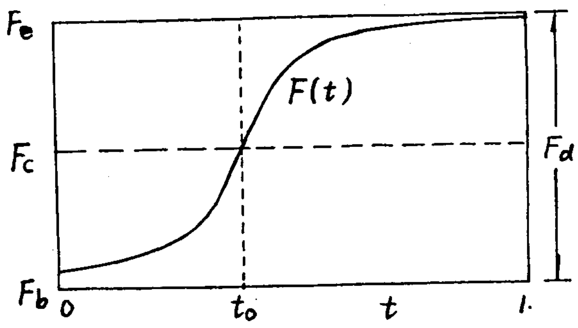


Fig.1 Schematic Dynamic Model for Diphthongs

Fig.2a and 2b show respectively the dynamic trajectories of formant frequences when t0 and α are altered. It is quite clear that in this model Fb and Fe can only approach the two target values rather than actually reaching them. The closer the division point is to the beginning point, the it is harder for the formant frequencies to reach the target value of Fb, and the easier it is for the ending formant frequencies to reach the target value of Fe; and, the greater the α, i.e. the factor of transitional rate, the easier it is for the formant frequencies of both of the extremities to reach their target values.

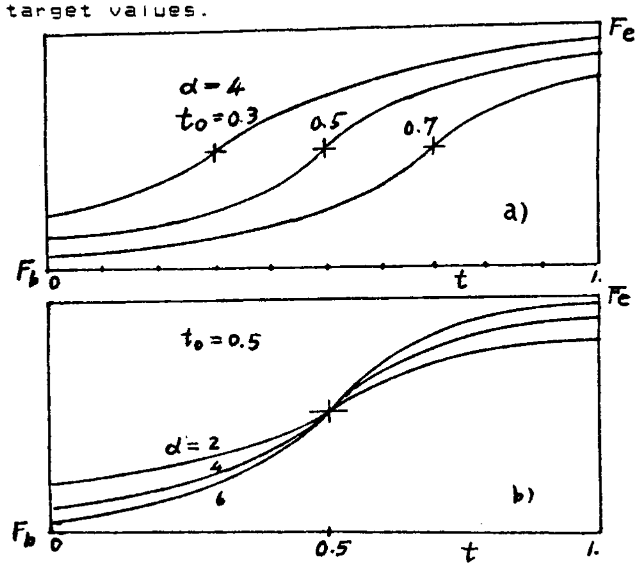


Fig.2 Variation of the formant trajectories with a) t0 and b) α

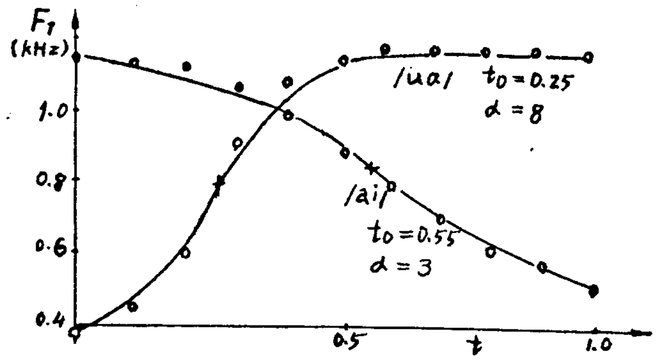


Fig.3 The measured formant frequency values and formant trajectories estimated with the formulas

The parameters Fb, Fe, t0 and α in the Model can be determined through analysis-by-synthesis. In Fig.3, the small circles represent the observed values of the first formant in /ai/ and /ua/, while the thin solid line is the trajectory calculated with formula (1) after the parameters for the Model had been determined. It can be seen that the two are in close approximation. As examples, the fitting values of t0 and α for F1 and F2 of the nine diphthongs are listed in Table 1.

Table 1 The fitting values of t0 and α for the 9 diphthongs in Standard Chinese

	/ai/		/ei/		/ao/	
	F1	F2	F1	F2	F1	F2
t0	0.55	0.43	0.27	0.19	0.52	0.49
α	3.0	4.0	3.1	4.2	1.9	2.1

	/ou/		/ia/		/ie/	
	F1	F2	F1	F2	F1	F2
t0	0.50	0.44	0.25	0.25	0.45	0.45
α	1.9	2.3	8.0	7.6	3.4	3.4

	/ua/		/uo/		/ye/	
	F1	F2	F1	F2	F1	F2
t0	0.20	0.25	0.35	0.42	0.35	0.23
α	7.1	7.8	3.8	3.8	3.5	3.5

Now, we can easily extend the Exponential Dynamic Model to include triphthongs. For triphthongs, considering the coarticulation effect between the three component phonemes, the dynamic trajectory of a given formant can be approximated by the following formula (Fig.4).

$$F(t)=Fb,m(t)+Fm,e(t)-Fm \qquad (2)$$

(for the meaning of the symbols here please refer to Fig.4)

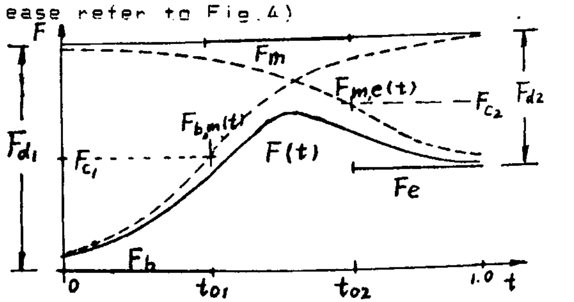


Fig.4 Schematic Dynamic Model for triphthongs

In this way, the dynamic aspect of a given formant in a triphthong can be defined by the tree target values Fb, Fm and Fe and the two division times t01 and t02 and the two factors of transitional rate α1 and α2, 7 parameters in all.

In overall generalization, for a given dynamic vowel that has n target values Fn, the dynamic trajectory of the frequency of a given formant can be approximated with the following formula:

$$\left.\begin{aligned}
&F(t)=\sum_{i=1}^{n-1}Fi,i+1(t)-\sum_{i=2}^{n-2}Fi \qquad (n\geqslant 2)\\
&Fi,i+1(t)=Fci+\\
&\quad 0.5S*Fdi\{1-EXP[-di(t-t0i)S]\}\\
&\qquad Fci=0.5(Fi+F_{i+1})\\
&\qquad Fdi=F_{i+1}-Fi\\
&\qquad S=1 \qquad (t-t0i\geqslant 0)\\
&\qquad S=-1 \qquad (t-t0i<0)
\end{aligned}\right\}\ (3)$$

SYNTHETIC PROOF AND APPLICATIONS

To verify the validity of this Exponential Dynamic Model, we had a synthetic experiment with the Software System for Chinese Syllables [1, 2]. This system uses a cascade formant synthesizer; with 10 KHz of sampling frequency, and 12 bit of precision for D/A converter. The synthesis was operated on a BCM-3 microcomputer. the frequencies of the first three formants for the 6 target phonemes used for synthesizing the 9 diphthongs and 4 triphthongs in Standard Chinese are listed in Table 2. The F4 and F5 were fixed at 3500 Hz and 4500 Hz respectively.

Table 2 Frequency values of the first three formants for the 6 target phonemes used for synthesizing the 9 diphthongs and 4 triphthongs in Standard Chinese

	/i/	/e/	/A/	/o/	/u/	/y/
F1(Hz)	270	520	1070	600	360	300
F2(Hz)	2350	2030	1200	1000	600	1890
F3(Hz)	3050	2720	2600	2500	2200	2250

Chinese is a tone language, and the F0-contour of each of the compound vowels were generated by a Tone Model [2].

All the syllables containing compound vowels in Standard Chinese were successfully synthesized. Fig.5 shows the spectragrams of four syllables, both natural

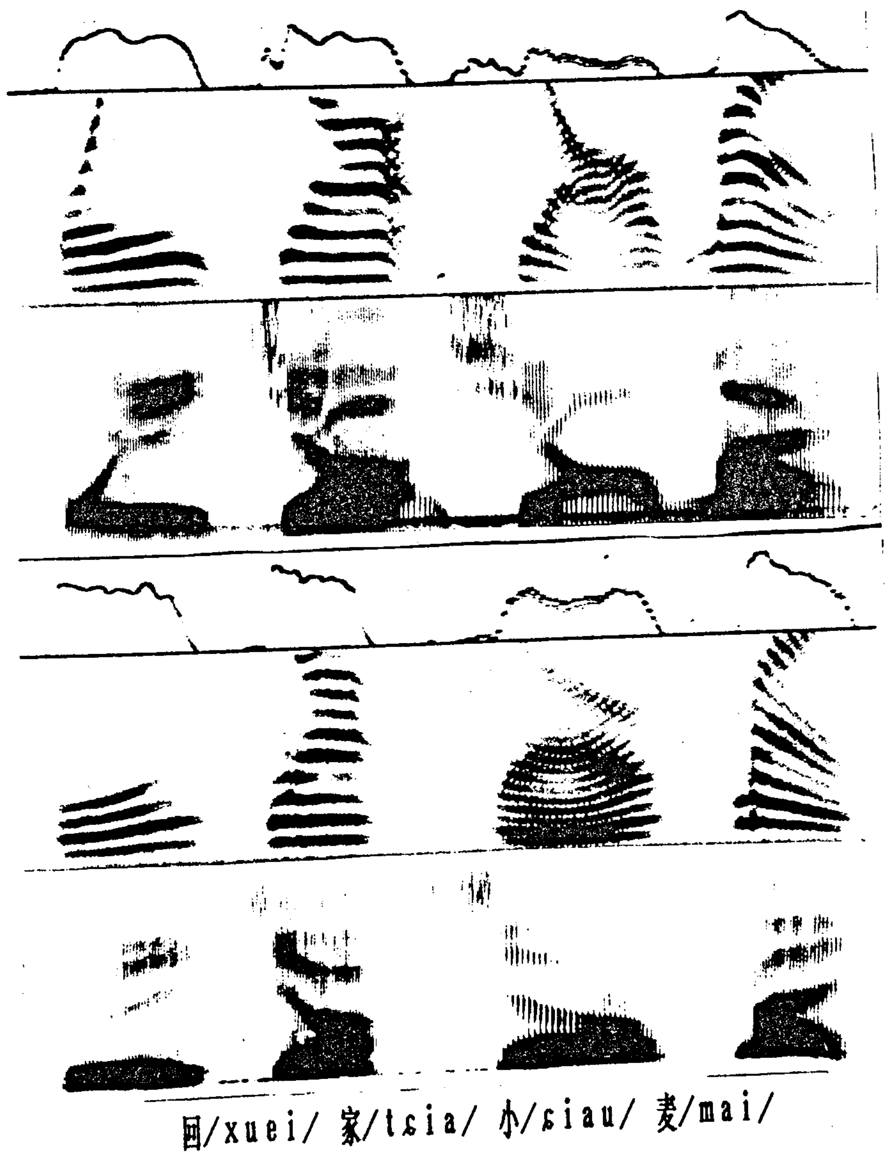


Fig.5 Spectragrams of four syllables containing diphthongs and triphthongs. The upper part for the natural ones and the lower part for the synthesized ones.

and synthesized, each containing a diphthong or a triphthong. It can be seen that the formant transition of those compound vowels are very smooth. Listening tests also indicated that both intelligibility and naturalness of the synthetic syllables were very close to those of the natural ones.

DISCUSSIONS

For synthetic application, there are two related features in this Exponential Dynamic Model: first, reltively few target values needed in input and storage, and second, better representation of the coarticulation effect. Speech analysis shows that one and the same phoneme in different compound vowels has different sound values. For example, the actual value of /ai/ and /ia/ are [ai] and [iA] respectively. Even two given vowels narrowly transcribed as the same sound in two different dynamic vowels, e.g. the [i] in [iA] and [iao] can have differences that should not be ignored. It means then, for synthesizing the 9 diphthongs and 4 triphthongs that are close to the natural ones, we will need 9*2+4*3=30 sets of target values. However, thanks to the ability of "approaching rather than actually reaching" the target values in the Exponential Dynamic Model, as few as 6 sets of target values listed in Table 2 are almost enough for this purpose. For instance, in synthesizing /ai/, [A] and [i] are used as target values; t0 is right in the middle and is relatively small. As a result, the beginning point is close to a open front vowel [a] rather than [A], and the ending point is a lower front vowel [I] rather than [i]. In synthesizing /ia/, [i] and [A] are also used as target values with t0 close to the beginning part and a relatively great , and the result is be that the two extremities are close to [i] and [A] respectively, and the /a/ part is relatively long and stable. In the acoustic vowel diagram in Fig. 6, the dynamic tracings are drawn for the synthetic /ai/, /ia/, /ao/, /ua/, /iao/ and /uai/ which use [i], [A] and [u] as the target values. The diagram shows that the beginning, middle and ending point of each of the compound vowels are just in their right places. In this sense, the synthesis of dynamic vowels with this Model is a synthesis with phonemic targets.

As a comparison, the trajectories generated by the exponential dynamic model reported in reference [3] and [4] always starts from the same first target value, disregarding the difference in factors like second target values and so on. The coarticulation effect is thus inadequately represented.

REFERENCE

[1] Yang Shun-an and Xu Yi (1987): A software system for synthesizing Chinese speech, Proc. 1987 Inter. Conf. on Chinese Information Processing, Aug. 4-6, Beijing, China.

[2] Yang Shun-an (1986): The effect of the dynamic characteristics of voice source upon the quality of synthesized speech, Zhongguo Yuwen, 1986 No.3, pp. 173-181 (in Chinese).

[3] Rabiner, L.R. (1968): Speech synthesis by rule: An acoustic domain approach, Bell System Tech. J., Vol.47,pp.17-37.

[4] Fujisaki, H. et al. (1973): Automatic recognition of connected vowels using a functional model of the coarticulatory process, J. Acoust. Soc. Japan, Vol. 29, pp. 636-637.

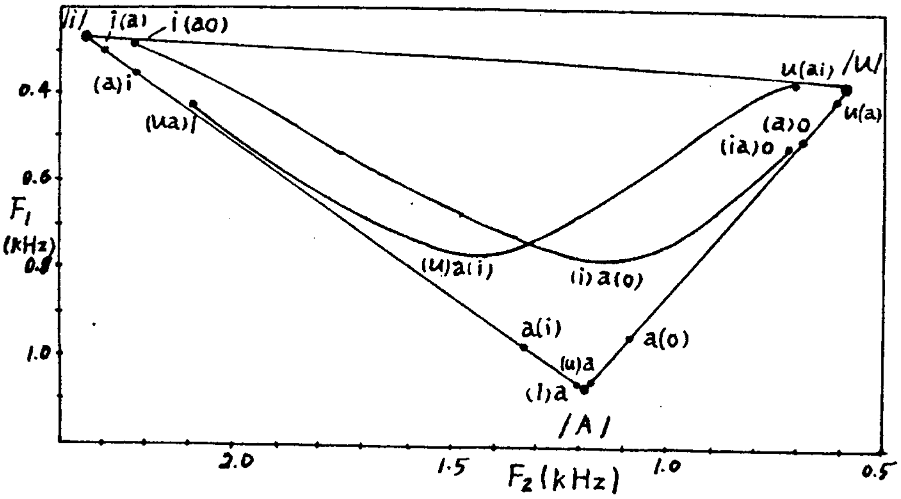


Fig.6 Acoustic vowel plot for the four diphthongs (/ai/, /ia/, /ao/, and /ua/) and the two triphthongs (/iao/ and /uai/).

MATHEMATICAL MODELLING OF THE FORMANT STRUCTURES OF VOCALIC SOUNDTYPE SYSTEMS

ALEXEY TYAPKIN

Institute "Sapsibnipiagroprom"
Novosibirsk 64, P.O.B. 95, USSR
Post Index 630064

ABSTRACT

Some problems of the formant structure modelling of vocalic soundtype systems (VSTS) by methods of multidimensional analytic and descriptive geometries as well as theories of convexes and inequalities are treated.

GENERAL CONCEPT

A concept of the acoustic structure modelling of VSTS was presented in a most general form formerly. By modelling we understand in this case a multistage process involving diverse aspects of the formant structure transfer of VSTS through mathematical structures and their graphic representation. Philosophy of modelling has exhaustively been considered elsewhere. Here, some basic problems, pertaining to principal modelling stages, are examined.

MODELLING AS A PROCESS

Sampling

The most effective acquisition of formant frequencies is to accomplish in a computer's memory coupled with an automatized formant frequency extraction yielding high precision readings. This stage deals also with statistical estimates of the formant data derived and with an evaluation of representing centroids (centers of gravity) of soundtypes as well. Measuring formant frequencies in spectrograms causes errors, is quite laborious and should preferably be avoided. However, problems of linguistic selection and phonetic realization of samples have undoubtedly to prevail at this stage.

Option and Construction of Models

There are three special kinds of modelling the F-structures of VSTS, producing correspondingly three types of models. Option of a particular model type depends on its purpose. Thus, the typology of the models in question covers the following types: (I) models of single soundtypes and of their systems through single models formed with approximating polinomials of 1st and 2nd degree; (II) models of VSTS formed with vector-to-point soundtype representation; (III) models of VSTS formed through axonometric constructions.

The modelling consists in formation of closed convex images in a multidimensional modelling space under employment of geometrical methods. In principle, a topological approach is also possible. However, geometrical constructions are important means of activating and stimulating the intuitive euristic image-bearing thinking.

Let us introduce a formant space of n dimensions with the Euclidean metrics therein. Then, the distance between two soundtypes X and Y with the formant frequency values X('F1, 'F2, ...'Fn) and Y ("F1, "F2, ..."Fn) is expressed as

$$L=(('F1-"F1)^2+('F2-"F2)^2+ \dots \dots+('Fn-"Fn)^2)^{1/2} \qquad (1)$$

The modelling F-space is necessarily isometrical if the coordinate axes therein are linearly scaled. With this goal in mind, both the natural frequency values as well as their logarithms linearly scaled are applicable. Since the image clarity and the complicacy of models are conflicting claims, subspaces of less than n dimensions are to be introduced. Thus, introducing, for instance, subspaces of 2 dimensions in the F-space of n dimensions, we have P subspaces which are actually modelling F-hyperplanes:

$$P= \sum_{q=1}^{q=n-1} (n - q). \qquad (2)$$

Single Soundtype Models

These models as well as such of VSTS through single models reflect the distribution of formant frequencies in the modelling F-space under condition of plural realization of soundtypes. Construction of models consists in an adequate linear or/and unlinear approximation of soundtypes in the F-space with closed convexes and in working out equivalent sets of linear or/ /and unlinear inequalities.

Linear Approximation. The problem of linear approximation of sourdtypes in the modelling F-space consists in forming convex areas by means of hyperplanes of less dimensions than the F-subspaces are in compliance with the formant data of soundtypes. Hence, e.g. a 3-D F-space contains, in accordance with (2), three 2-D F-subspaces and 0-D, 1-D hyperplanes as simplexes embodied in points and vectors, s. Table 1. Apparently, simplexes may be formed with hyperplanes of no more than n-1 dimensions, cutting out the simplexes in the F-space. Any hyperplane can be described then as

$$a_1F_1+a_2F_2+\dots$$
$$+\dots a_{n-1}F_{n-1}+a_nF_n+a_{n+1}\geqslant 0\ . \quad (3)$$

When reffering to modelling 2-D F-subspaces and 1-D hyperplanes, a model will be formed with 1-D simplexes given through algebraic sets compraising inequalities of the following form:

$$a_{11}F_k+a_{12}F_{k+1}+a_{13}\geqslant 0. \quad (4)$$

Forming a v-dimensional simplex of (v-1)-dim. hyperplanes, a minimal number of hyperplanes, forming the simplex, can be written in the form

$$h_{min}\geqslant v+1 \quad (5)$$

The coresponding number of inequalities, describing this soundtype, amounts then to:

$$H_{min}\geqslant P(v+1) \quad (6)$$

The sign of inequality in (3) and (4) indicates the sharing of the F-space into two F-subspaces: the polynomial has negative values in one F-subspace and positive ones or zero in the other. A set of inequalities, in this way, is capable of describing a complex of limited convex F-subspaces, forming a polyhedral (if v=2, then it is a polygonal domain) closed convex area of solutions of the set modelling a soundtype given. Obviously, a more extended set of modelling inequalities contributes to a better approximation of soundtypes.

Unlinear Approximation. This type of approximation is based on the formation of n-1 dimensional closed convex (hyper)surfaces of the second degree, or quadrics, to be described through quadratic inequalities and enclosing the point sets, corresponding to soundtypes given, in the F-space of n dimensions. It includes several conoidal types: hyperboloidal, paraboloidal, and ellipsoidal approximations. Preferably, an ellipsoidal approximation should be used as yielding closed convex hypersurfaces. However, this procedure is rather laborious, so that a meaningful use of 2-D subspaces, consistently with (2), instead of the n-dim. modelling F-space has to be preffered. The quadrics used will have then the index 1-D, and the approximation of soundtypes will make use of second degree curves: 1-D ellipsoids (ellipses) or even 1-D spheres (circles). Pascale and Brianchon's theorems are helpful when constructing approximating ellipses. Parabolic or hyperbolic approximations are also applicable.

A second degree inequality in a most general form is as follows:

$$\sum_{i,k=1}^{n} a_{ik}F_iF_k + 2\sum_{i=1}^{n} b_iF_i + C\geqslant 0 \quad (7)$$

First, it can be reduced to the form

$$w_1{}'F_1^2+w_2{}'F_2^2+w_3{}'F_3^2+\dots$$
$$\dots+w_n{}'F_n^2+w_{n+1}\geqslant 0 \quad (8)$$

Table 1

ELEMENTARY LINEAR MODELLING CONSTRUCTS
(SIMPLEXES)

Number of soundtypes in simplex	Simplex	Dimension of simplex	Formal characteristic parameters
1	Point	0	-
2	Vector	1	Vector length, distance between vector's ends
3	Triangle	2	Side length, area, angles, gravity center position
4	Tetrahedron	3	Edge length, side area, volume, gravity center position

Table 2

UNLINEAR CLOSED CONVEX IMAGES
APPROXIMATING VOCALIC SOUNDTYPES

Modelling geometrical image	Dimension of modelling F-(sub)space	Approximating inequality in general form	Formal characteristic parameters
Ellipse	2	$a_{11}F_k^2+2a_{12}F_kF_{k+1}+a_{22}F_{k+1}^2+$ $+2a_{13}F_k+2a_{23}F_{k+1}+a_{33}\leqslant 0$	Half-axis length, contraction factor, area, gravity center position
Ellipsoid	3	$a_{11}F_1^2+a_{22}F_2^2+a_{33}F_3^2+2a_{12}F_1F_2+$ $+2a_{23}F_2F_3+2a_{13}F_1F_3+2a_{14}F_1+$ $+2a_{24}F_2+2a_{34}F_3+a_{44}\leqslant 0$	The same; volume

When reffering to 2-D and 3-D modelling (sub)spaces, (8) is reduceable to the forms indicated in Table 2. In the case of an n-dimensional ellipsoidal approximation, we have:

$$\frac{'F_1^2}{t_1^2}+\frac{'F_2^2}{t_2^2}+\frac{'F_3^2}{t_2^2}\cdot\cdot+\frac{'F_n^2}{t_n^2}-1\leqslant 0 \quad (9)$$
$$t_n^2=+(w_{n+1}/w_n)$$

Plane VSTS Models

Structural models of VSTS are constructed by means of elementary linear constructs (s. Table 1) when higher forms of abstraction are substituted for lower ones. It is natural to construct a structural model through a transformation of single soundtype models and by reducing them to 0-D simplexes in the n-dim. modelling space. This implies that higher dimension simplexes of single soundtype models are substituted by lower dimension simplexes of structural models. These latter incorporate into sets of higher dimension simplexes, thus forming spatial models of the F-structures of VSTS. It is obvious that in 2-D (sub)spaces the use of modelling simplexes of 0 to 2 dim. and in 3-D (sub)spaces that of simplexes of 0 to 3 dim. is possible. In any case, a structural model is cut out through hyperplanes as a multidimensional polyhedron with every vertex representing a single soundtype. A geometrical interpretation of the F-structures of VSTS and formal characteristic parameters as well allow some acoustic problems in the n-dim. F-space to be solved by means of numerical and graphometrical methods.

Axonometric VSTS Models

Axonometric models are essential for a spatially condensed picturelike portraying of the F-structures of VSTS in the n-dim. F-space (n 3). Though the problem of visual support through graphic aids in science is rather vague, it is not reasonable to underestimate a contribution of image-bearing thinking and euristic factors in research work and education. However, the solution of metric problems with axonometric models becomes too complicated so that their use is limited by illustrative and demonstrative goals. Also, there are some specific problems in application of multidimensional descriptive geometry methods to the modelling of the F-structures of VSTS. Rather promising in this respect seems to be the construction of axonometric models on the basis of a pair of usually available, reciprocally orthogonal, plane modelling images.

VERIFICATION AND IDENTIFICATION OF MODELS

These modelling stages are required to prove the agreement between the soundtypes to be modelled and their models. The modelling process is considered as being completely finished if only the following chain has been preserved: (a) soundtype X, (b) model A, B..., (c) soundtype Y... The relationships between the links of the chain are of fundamental importance and manifest themselves in the course of the verification of the dyad (a) and (b) or/ and the identification of the dyad (b) and (c). As a result, an agreement or a disagreement between the soundtypes X and Y can be stated, the relationships between the soundtypes and the models being supposed those of a structural analogy.

Principally, the following relationships between soundtypes and their models are to be expected: (I) reflexivity, (II) transitivity, (III) antisimmetry. These qualities of a model are usually combined with each other, but if being absent, they imply the presence of a converse quality as it will be clearly shown below. The model qualities mentioned above signify the following: (I) a soundtype is a model of its own; (II) a model's model is a model of the prototype; (III) a model in general is a homomorphic image of a soundtype given and can be substituted for the latter within the limits of its characteristics of significance which permit to state a structural analogy. Thus, verification and identification both are counterpart processesand qualify the relationships between soundtypes and their models as follows: (i) b:a=a:b, b:a≠a:b (verification, antisimmetry/simmetry); (ii) b:c=c:b, b:c≠c:b (identification, antisimmetry/simmetry; (iii) if b:a=a:b and b:c=c:b, then a=c (transitivity/antisimmetry)

In this way, 1) if the model A is separately adequate to soundtypes X and Y, then X=Y (transitivity, antisimmetry); 2) if the model A is not adequate to at least one of soundtypes X and Y, then X≠Y (absence of transitivity, simmetry); 3) if the model A is adequate to a soundtype X while the model B is adequate to a soundtype Y, and A=B, then X=Y (reflexivity, transitivity, antisimmetry); 4) if in the preceding item A≠B, then X≠Y (absence of transitivity, simmetry).

Summing up, we may state that the above-mentioned relationships as well as the deformations of models are subject to investigation by means of:A) characteristic parameters (s. Tables 1 and 2); B)geometrical affine transformations of the models /including 1) parallel tarnsfer of the F-structure in the F-space, 2) rotation of the F-structure in the F-space; 3)contraction or expansion of the structure along the coordinate axes in the F-space/.

SUMMARY

A brief account of means and ways of the mathematical modelling of the acoustic structures of vocalic soundtype systems by methods of multidimensional analitical and descriptive geometries, theories of convexes and inequalities has been presented. The modelling stages may well involve the use of computers and graph plotting devices as working tools. In general, the geometrical approach traced proves to be an effective means of the mathematical modelling of vocalic soundtype systems in research work, demonstration and illusration processes.

BIBLIOGRAPHY

Batoroyev K.B. Analogii i modeli v poznanii.-Novosibirsk, 1981.

Bonnesen T., Fenchel W. Theorie der konvexen örper.-Berlin,Heidelberg, New--York, 1974 (Berlin, 1934).

Borsuk K. Multidimensional analytic geometry.- Warsaw, 1969.

Chao Yuen Ren. Models in linguistics and models in general. In: Logic, methodology and philosophy of science.-Stanford, 1962.

Dambska J. Le concept de modèle et son rôle dans les sciences.-Revue de synthèse, 1959, vol. 80, n° 13-14.

Hesse M.B. Models and analogies in science.-L., N.-Y., 1963.

Kuipers A. Model en inzicht.-Nijmegen, 1959.

Serge C. Mehrdimensionale Räume.- Enzyklopadie der mathematischen Wissenschaften, 1920, Bd. 3, H. 7, SS. 769-972.

Sommerville D.M.Y. An introduction to the geometry of n dimensions.- L., 1929.

Straas G. Modell und Erkenntnis.-Jena, 1963.

Tiapkin A.D. Matematicheskoye modelirovaniye formantnykh struktur.-In: IV vsesoyuzny simpozium "Metody teorii identifikacii v zadachakh izmeritelnoy tekhniki i metrologii", tezisy dokladov, September 10-12, 1985, Novosibirsk, pp. 253, 254.

Valentine F. Convex sets.-N.-Y., 1964.

APPROXIMATION OF INTONATION STRUCTURE OF SPEECH

TAMARA BROVCHENKO

Lab. of experimental phonetics
Odessa State University
Odessa, Ukraine, USSR 270021

VLADIMIR VOLOSHIN

Lab. of experimental phonetics
Odessa State Univeristy
Odessa, Ukraine, USSR 270021

ABSTRACT

The approximated intonation contours allow one to vizualise the most typical features of the melody and energy structure of the utterance in the form, directly appliable in automatic recognition and synthesis of speech prosody.

In a series of experiments discussed in the present paper typical intonation contours of various communicative types of phrases in Russian and English expressive conversation (as compared to the monotonous one) have been determined.

The most adequate methods of approximation of intonation contours have been analysed. Analytical expression which offers opportunity for presenting each intonation contour as a mathematical model has been suggested.

INTRODUCTION

In studing the intonation structure of speech a number of problems arise. Alongside with the problem of determining the physical nature of the phenomenon under study and defining typical intonation contours it is extremely important to elaborate the form of presentation of the intonation contours which should be precise and easy to apply.

The purpose of this paper is to compare the intonation structure of phrases read with expression to those read monotonously and to make an attempt to elaborate an analytical expression of typical intonation contours of expressive speech.

INTONATION CONTOURS OF EXPRESSIVE SPEECH

In our studies, five adult male speakers of British English and five speakers of Russian recorded a set of English and Russian written dialogues read with expression, lively and animatedly and then a set of the same dialogues, read monotonously, without expression. 20 statements, 20 questions (yes, no) and 20 request were picked out of these dialogues (a total of 600 utterances) and used for this experiment. The acoustic characteristics (fundamental frequency, duration and intensity) were measured for the two sets of the data. The problem was not simple that of describing the acoustic characteristics, but it was just as important to determine which of these characteristics are significant in discriminating expressive utterance and those read monotonously.

It has been commonly assumed that any speech realisation is a random process which is described in terms of a functional dependence of the variable in time, whose parameter value can be presented with the help of the parametric equation:

$$X[t] = f[A_i B_i D_i C_i] \qquad (1)$$

where A_i – constant parameters, unchangeable in all realisations;

B_i – interfering factors, varying from one realisation to another by some unknown law of distribution;

C_i – occasional interference, varying in separate elements of the utterance and describable by normal distribution;

D_i – the unknown parameters being sought, which determine the realisation as belonging to a given linguistic phenomenon.

In case occasional interferences are minimized, they will slightly influence the characteristics of the phenomenon under study, and the parametric model may be presented as a model with additive interference:

$$X[t] = E[D_i A_i B_i] + C_i \qquad (2)$$

where $E[D_i A_i B_i]$ – range of parameters, describing the realisation being formed with no interferences present.

With various values of the parameters defined, function $E[D_i A_i B_i]$ gives a set of specific realisations as an ensemble, presenting phenomenon analysed.

The parametric model is described in the present paper in terms of discrete values of the fundamental frequency and intensity. These are associated with a definite number of points within each structural element of the utterance: 3 measurements within the initial unstressed syllables; 7 measurements within the head of the utterance (the first stressed syllable and all the stressed and unstressed syllables preceding the nucleus, 4 measurements within the nucleus and 2 within the tail. In total 16 measurements within each utterance. As a result the so called dynamic or temporal series was obtained.

Occasional interferences were reduced by the requirements of the procedure being kept fairly equal through the whole experiment. The realiability of the characteristics obtained in this experiment were ensured by statistically reliable number of speakers and amount of the experimental data. In addition only those utterances which were accurately identified by not les that 95% of the listeners were selected for further electroacoustic analysis.

The average values of the fundamental frequency and intensity were taken as a basis for a generalized intonation contour which reflects the main regularities of the phenomenon under study.

The results of these experiments have shown that the acoustic peculariities of expressive speech find vivid reflection in the dynamic series of the fundamental frequency, i.e. in the melody contour of various communicative types of phrases. It will be noted that the quality of speech (expressive or monotonous) determines the frequency level of the utterance, the speed of the fundamental frequency within the head and the nucleus, the location of the melodical peak of the utterance. These cues have been found typical of both English and Russian and it may be suggested that they are typological.

Figures 1 and 2 represent melody contour (dynamic series) of utterances in expressive and monotonous speech. The solid curve represents the average fundamental frequencies of statements, the dashed curve of questions, the dotted curve of requests. Structural elements of the utterance (P/h-initial unstressed syllables; h - head; n - nucleus, t - tail) were plotted as abscissae. Average normalized fundamental frequency as ordinates.

Average fundamental frequency values were normalized with the help of the equation:

$$X_n = \frac{X_i - X_{min}}{X_{max} - X_{min}} \cdot 15 \qquad (3)$$

where X_i - selective value of the characteristic;
X_{max}, X_{min} - limit values of the characteristic.

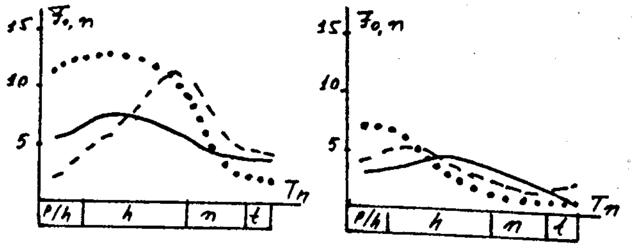


Fig. 1. Average normalized values of the fundamental frequency of utterance in expressive (left) and monotonous (right) Russian speech.

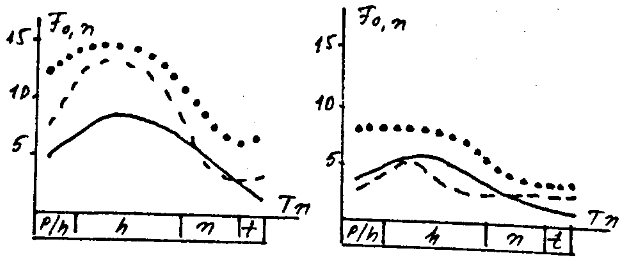


Fig. 2. Average normalized values of the fundamental frequency of utterances in expressive (left) and monotonous (right) English speech.

The experiments suggest that it is possible to establish the melody contours typical of expressive speech.

As to the values of intensity, the analysis reveals a relatively distinct difference between expressive and monotonous speech, the level of intensity both in English and in Russian being considerably higher in expressive speech.

On the other hand, the form of the intensity curve has shown remarkably little variation from utterance to utterance, from speaker to speaker, from one communicative type to another in expressive and monotonous speech. Commonly it has the shape of a gradually descending curve (fig. 3). The fact that different units of speech: a syllable, a sence-group, a phrase, etc. - are characterized by a similar envelope of the intensity makes it possible to conclude that the form of the intensity curve is of paramouth importance in organizing units of speech.

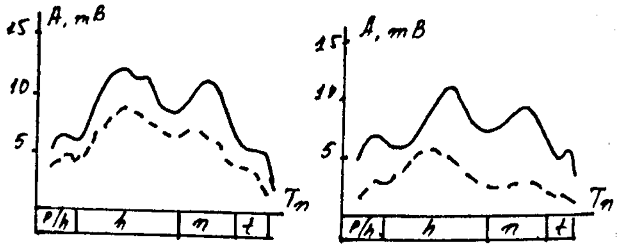


Fig. 3. Average normalized values of intensity of statements in expressive (solid curve) and monotonous (dashed curve) Russian (left) and English (right) speech.

Though the average values of the acoustic characteristics reveal the main regularities of the intonation contour it would not be sufficient to analyze only the average values. One should also study the varieties of acoustic cues in definite speech realizations.

As shown in Fig. 4-6 a number of realizations of utterances of the same communicative type make an ensemble within which it is possible to select from one to four main variants, differing to some extent in frequency, configuration of the curve, etc. In some cases the variants are equivalent and interchangeable in others they are dependent on the degree of expressiveness, on modal and emotional colouring and extralinguistic factors.

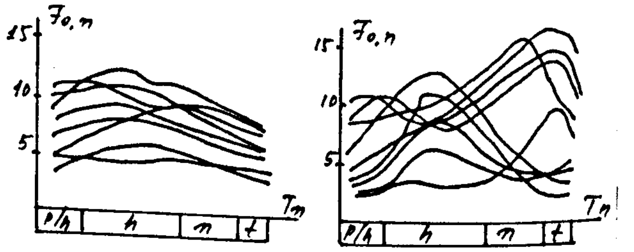


Fig. 4. Ensembles of intonation contours of statements (left) and questions (right) in Russian expressive speech (speaker RM_1).

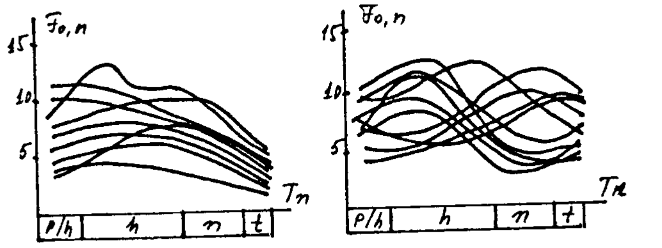


Fig. 5. Ensembles of intonation contours of statements (left) and questions (right) in English expressive speech (speaker EM_1).

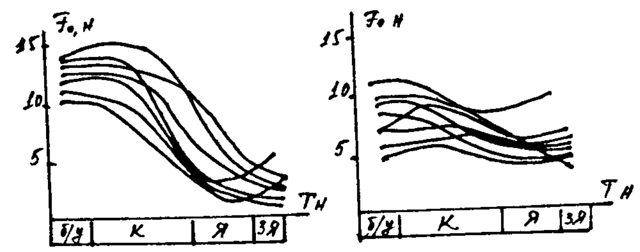


Fig. 6. Ensembles of intonation contours of requests in Russian (left) and English (right) expressive speech (speakers RM_1 and EM_1 correspondingly).

Attention should be drawn to the fact that specific intonation contours represent only one of many possibilities to make speech expressive. A quantitative study of the intonation structure of speech has suggested that besides the above mentioned acoustic cues of separate expressive utterances acoustic characteristics of the whole text might account for the difference between expressive and monotonous speech. Valid data were obtained showing that acoustic characteristics within the text provide effect of expressiveness of speech. Of particular interest in the present study, however, is that the correlation of the fundamental frequency and intensity values at the border of sense groups and phrases constituing the text, the correlation of acoustic measurements of initial and final unstressed syllables of different phrases of the text, etc. indicate whether the text is expressive or monotonous. Besides the alternation of different equivalent variants of intonation contours of one and the same communicative type of the phrase within a speech sample as well as the alternation of phrases with different level of intensity makes speech expressive.

These questions, however, are beyond the scope of the present paper.

APPROXIMATED INTONATION CONTOURS

Our final experiment aimed at the problem of approximation of the intonation contour of the utterance. There is a strong evidence to suggest that the main features of the intonation contour appear to be associated in the mind of the speaker with the communicative type of the utterance, its modality and emotional colouring, the degree of expressiveness and other linguistic and extralinguistic factors. It seems that initial and final values of the fundamental frequency and intencity, as well as the configuration of the curve are direct cues in "planning" the intonation contour of the utterance.

Taking it into consideration the values of the physical characteristics at the beginning and at the the end of the utterance, as well as the configuration of the curve were taken as a basis for approximating the intonation contours of expressive speech.

Variants of the trajectory of fundamental frequency and intensity measurements, obtained in the present study, could be readily approximated as close to the original as possible by analytical expressions, describing the intonation contour with the help of the method of least squares.

The method of analytical approximation includes: (1) establishing the character of the dependence and selection of corresponding equations: (2) minimizing trajectory deviations of the analytical expression from natural speech contour: (3) evaluating the constant coefficients that determine the trajectory of the changes in the parameters under study.

The analytical expression describing the trajectory of the fundamental frequency changes have been developed experimentally and calculated by the formula:

$$y[t] = F_{in} e^{-\alpha t^2 + \beta t} + F_{fin} e^{-K/t} \qquad (4)$$

where F_{in}, F_{fin} - values of the parameter at the beginning and the end of the speech sample:
t - successive number of time - segment values:
α, β, K - constant coefficients, selected for each realization in terms of the intonation contour.

For the analytical expression describing the trajectory of the intensity changes it is possible to express that function as follows:

$$y[t] = A_{in} e^{-\alpha t^2 + \beta t} \qquad (5)$$

where A_{in} - value of the parameter at the beginning of the speech sample.

In case of complicated curves (those having more that two turning points) the approximation is calculated by formula (4), with the beginning and the end of each structural element taken for the values of F's.

Coefficients α, β, K - determine the profile of the curve and account for the occational interferences and the parameters of the model sought for. Coefficient α varies in the range: .01 ÷ .3; β - .1 ÷ .5; K - 2 ÷ 14.

Particular values of the coefficients used in approximating each intonation contour are given in Table 1.

Table 1. Analytical expression of intonation contours approximation

	Communicative type of the phrase	Language	Analytical expression of approximation
Fundamental frequency	Statements	Russian	$y[t]=100e^{-.01t^2+.05t}+80e^{-6/t}$
		English	$y[t]=115e^{-.01t^2+.1t}+80e^{-8/t}$
	Questions	Russian	$y[t]=100e^{-.01t^2+.14t}+90e^{-5/t}$
		English	$y[t]=150e^{-.016t^2+.16t}+130e^{-10/t}$
	Requests	Russian	$y[t]=210e^{-.012t^2+.06t}+110e^{-5/t}$
		English	$y[t]=195e^{-.01t^2+.05t}+100e^{-55/t}$
	Statements (intensity)	Russian	$y_1[t]=100e^{-.06t^2+.2t}$ $y_2[t]=106e^{-.05t^2+.42t}$ $y_3[t]=175e^{-.05t^2+.21t}$ $y_4[t]=125e^{-.033t^2+.04t}$
		English	$y_1[t]=100e^{-.07t^2+.23t}$ $y_2[t]=100e^{-.03t^2+.3t}$ $y_3[t]=150e^{-.06t^2+.23t}$ $y_4[t]=100e^{-.07t^2+.23t}$

The results of calculations are plotted in Fig. 7 - 10.

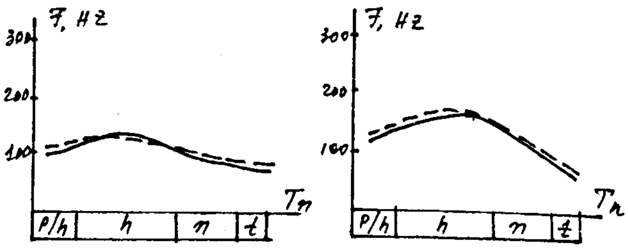


Fig. 7. Melody countours of statements in Russian (left) and English (right) expressive speech (solid curve) and their approximated variants (dashed curve).

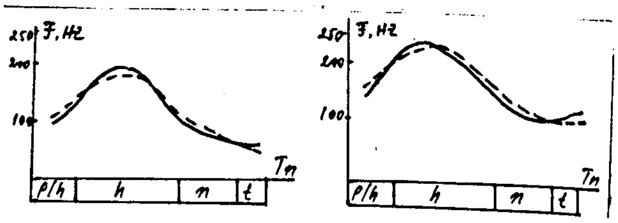


Fig. 8. Melody contours of questions in Russian (left) and English (right) expressive speech (solid curve) and their approximated variants (dashed curve).

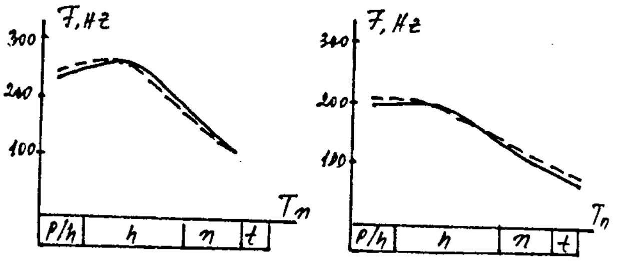


Fig. 9. Melody contours of requests in Russian (left) and English (right) expressive speech (solid curve and their approximated variants (dashed curve).

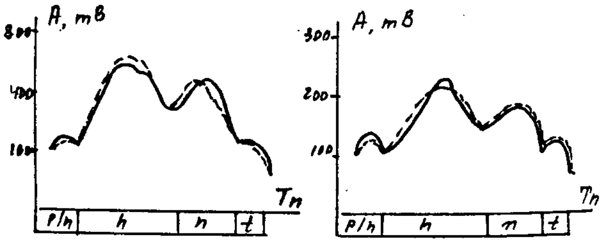


Fig. 10. Intensity contour of statements in Russian (left) and English (right) expressive speech (solid curve) and their approximated variants (dashed curve).

CONCLUSION

It appears from the foregoing analysis of the intonation structure of Russian and English utterances that differences in perception of degree of expressiveness are always associated with respective differences in the characteristics of the intonation contour of the phrase and those of larger speech units.

The analytical expression suggested enables to approximate the intonation contour of various types of expressive utterances close to the original intonation contours, preserving all their main properties. As compared to approximation by polynomial, the present method is more simple and effectual. The presentation of the intonation contour as a mathematical model makes it possible to use it directly in the synthesis of speech prosody.

REFRENCES

/1/ Bengrande K.D. Text Discourse and Process. - London: Longman, 1980. - 231 p.

/2/ Vazen M. Statisticheskaja approksimatsia. - M. :Mir, 1972. - 295 s.

/3/ Kovalev V.P. Virazitelnije sredstva hudozestvenoj rechi. - Kiev: Rad. Sohkola, 1985.-136 s.

/ 4/ Torsueva I.G. Sovremennaja problematica intonatsionnych issledovanij //Voprosy Yazicoznaniya, 1984. - N 1. - S. 116-126.

LATERAL INHIBITION AND SPEECH SIGNAL PROCESSING

DANG* V.C., CARRE R.

Laboratoire de la Communication Parlée, ICP Unité Associée au CNRS,
INPG-ENSERG, 46 Avenue Félix Viallet,
38031, GRENOBLE CEDEX, FRANCE.
*Present address : Khoa vô tuyên, DHBK, HANOI, VIETNAM.

ABSTRACT :

Lateral inhibition is a side-band effect of excitation of the auditory system by a complex signal. Indeed, single neuron response is modified by the signals issued by surrounding neurons due to the complex stimulation. In this paper, we present works on this subject using a simplified model over natural and synthetic speech sounds. A spectral lateral inhibition is used to enhance spectral peaks. Preliminary tests on temporal lateral inhibition (lateral inhibition in time-domain) show an enhancement of time-domain contrasts. This information might be used to find stable regions in the speech signal.

1. INTRODUCTION

In the past years, it has been recognized the existence of a lateral inhibition function in neuronal processings and several works have been developed on the modelling of this function (GREENWOOD & MARUYAMA - 1965, GREENWOOD & GOLDBERG - 1970, MORISHITA & al. - 1972, TOKURA & al. - 1977, CAELEN - 1979, VOIGHT & YOUNG -1980, PALMER & EVANS - 1982, MARTIN & DICKSON - 1983, SHAMMA - 1985). In short, single neuron response is modified by the signals issued by surrounding neurons due to the complex stimulation.

KARNICKAYA & al. (1973) have applied a "lateral inhibition" model on the auditory spectrum equivalent and have observed that spectral contrasts are increased. They have used a three-range window : a central positive one and two lateral negative ones, gliding in the frequency domain.

MORISHITA & al. (1977), SHAMMA (1985) have tested neuron network models.

LEBEDEV & al. (1985) have built a performant recognition system by taking into account the time-domain and frequency-domain masking effects.

In this work, a simplified inhibition model similar to that of Karnickaya's is tested, in the frequency domain and in the time domain, to point out contrast effects on the spectrum. This three-range window model can be compared with the cepstral technique where the inverse FFT + square windowing + direct FFT block corresponds to a sin(x)/x operator with a positive central lobe and two main negative lobes.

2. EXPERIMENTS

Original speech signal is low-pass filtered at 5 KHz, sampled at 10 KHz, weighted by a Hamming window and processed via FFT. The spectral components e(t,j), at the FFT output(t = time, j = frequency) are then processed by a lateral inhibition system.

In the frequency domain, the spectral lateral inhibition filter contains one central region and two lateral inhibition regions, as represented in figure 1.

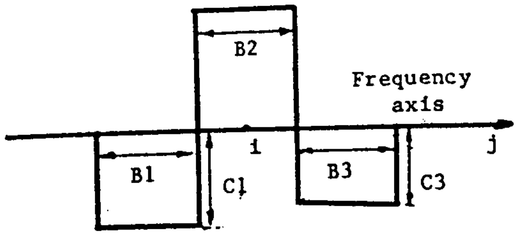


Figure 1. Spectral lateral inhibition filter.

The filter output S(t,i) is the weighted sum of the inputs e(t,j) in the central region minus the sums of the two lateral inhibition regions :

$$S(t,i) = -C1\sum_{j\in B1} e(t,j) + \sum_{j\in B2} e(t,j) - C3\sum_{j\in B3} e(t,j)$$

C1, C3 are amplitude constants.

The three range filter is applied on the FFT spectral components and the filter is gliding on the frequency scale. B1, B2 and B3 are set in a Bark scale.

In the time domain, two filters have been tested :

Type 1 : the time domain filter is exactly corresponding to the frequency domain filter described above :

$$S(t,i)' = -D1\sum_{t\in T1} e(t,j) + \sum_{t\in T2} e(t,j) - D3\sum_{t\in T3} e(t,j)$$

D1, D3 are amplitude constants and the T constants are duration constants. Figure 2 represents the time-domain lateral inhibition law, which is a modification of Lebedev's time masking curve (LEBEDEV & al. (1985)).

Type 2 : the output element $S_2(i)$ in a time-domain lateral inhibition filter is computed over the sum of the absolute values of differences between the spectral components processed at time i and i-1 (the output of the first element S(i,j)). Equations a and b define $S_2(i)$

a) $$S_1(i) = \sum_{j=1}^{N} |S(i,j) - S(i-1,j)| \;;$$

b) $$S_2(i) = -D1\sum_{j\in T1} S_1(j) + \sum_{j\in T2} S_1(j) - D3\sum_{j\in T3} S_1(j)$$

i, j, T1, D1, T2, T3, D3 are defined on figure 2.

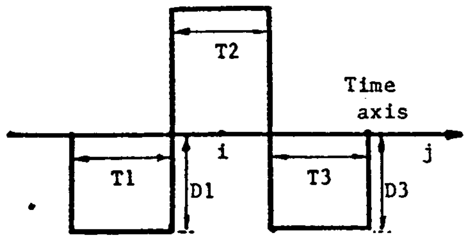


Figure 2. Time-domain lateral inhibition law.

3. RESULTS AND DISCUSSIONS

Spectral lateral inhibition.

a) Study of the model parameters.

In order to study the role of the model parameters. /a/, /i/, /u/ vowel spectra were calculated for different values of one parameter among the others. The objective was to find the right value corresponding to a better contrast effect on the spectrum. The optimal values for the window ranges are 1 Bark, and the amplitudes C1, C2, C3 are -0.7, 1, -0.3. These values are closed to those proposed by KARNICKAYA.

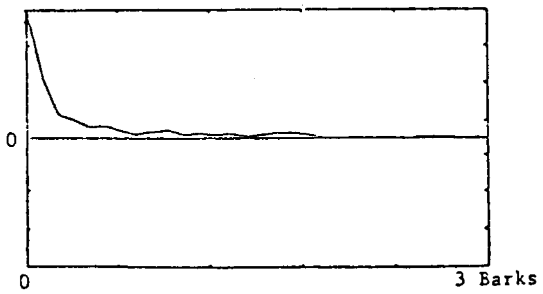


Figure 3. Evolution of the spectral distance (vowel /a/) for different B1 values.

Such parameter values were tested to verify the good stability of synthetic vowel spectrum representation. The euclidian distance between two successive spectra was calculated for different values of each parameter (figure 3). This distance has a first minimum when B values equal 1 Bark and when C values are around -0.7, 1, -0.3.

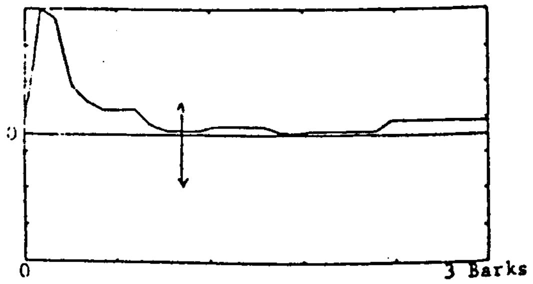


Figure 4. Distance between the formant values (for synthetic vowels) and the spectral peak values, for different B1 values.

The parameters were also tested to verify the good acuracy of the spectral peaks. For different synthetic vowels with specified formant frequencies, the distances between these frequencies and the peak frequencies of the spectral representation were calculated. Again the distances have a first minimum at around 1 Bark for the B values and at -0.7, 1, -0.3 for the C values (figure 4).

b) Results on synthetic vowel signals.

Figure 5 shows the spectral representation obtained by FFT (curve 1), FFT + 1 Bark integration (curve 2), cepstral technique (curve 3) and FFT + lateral inhibition.

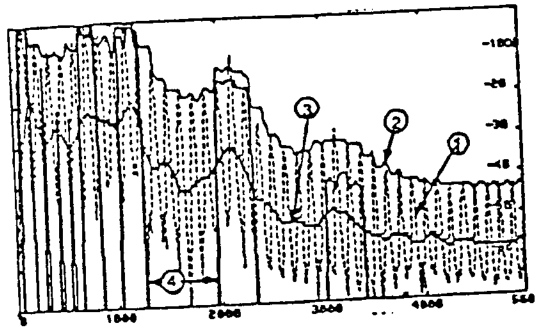


Figure 5. Spectral lateral inhibition for /a/ vowel (curve 4).

The parameters of the lateral inhibition model are : 1 Bark for the B values and -0.3, 1, -0.7 for the C values. The spectral contrast is clearly increased.

In the case of noisy vowels, spectral peaks are better represented in the case of lateral inhibition processing (Figure 6).

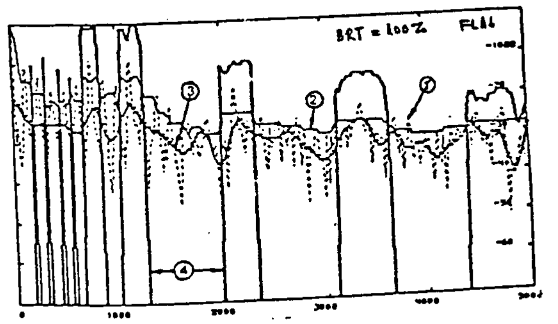


Figure 6. Spectral lateral inhibition for noisy /a/ vowel (noise level 100%) - curve 4 . Curve 1 : FFT, curve 2 : FFT + 1 Bark integration, Curve 3 : cepstral representation.

c) Results on natural CVCVC sounds.

The use of the spectral lateral inhibition clearly increases the spectral contrast (figure 7). When the central range of the model is only one FFT value, the contrast is more important and the harmonic structure appears mainly for low frequencies (figure 8).

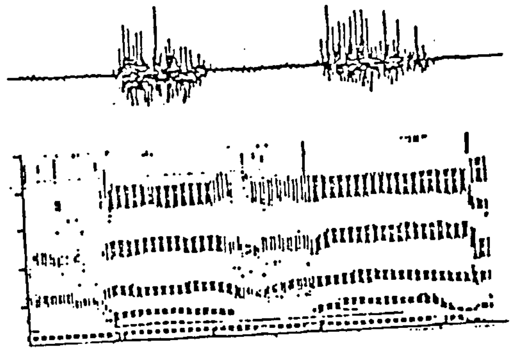


Figure 7. FFT + spectral lateral inhibition (CVCVC : /babab/).

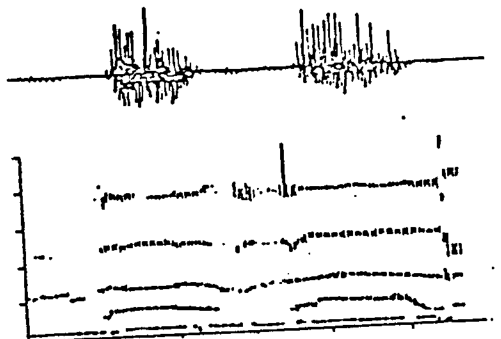


Figure 8. FFT + reinforced inhibition (CVCVC : /babab/)

Temporal lateral inhibition.

The type 2 representation is given on figure 9 with duration ranges of 5 ms. Lateral inhibition gives peaks at the place of temporal transitions. This system could be used for event detection.

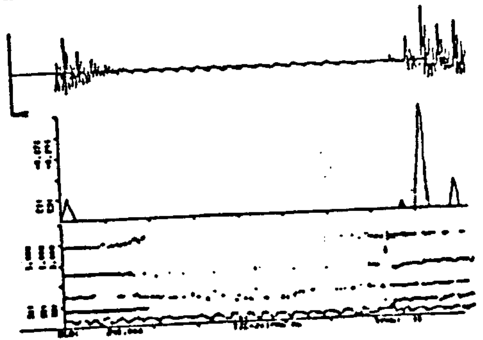


Figure 9. Time domain lateral inhibition. Temporal range = 5 ms, CVCVC : /babab/, /aba/ part.

4. CONCLUSIONS

The results obtained show that lateral inhibition is able to increase temporal and spectral irregularities. Increased spectral irregularities enhance the spectral peaks. Thus, the speech spectrum is simplified. According to the parameter values of the model, the low frequency harmonic structure can be observed.

In the time domain, according to the parameter values of the model, lateral inhibition enhances either the boundaries of the stationary sounds or small temporal events.

5. REFERENCES

(1) Chistovich L.A, Lublinskaya V.V., Malinnikova T.G., Ororodnikova E.A., Stoljarova E.I and Zhukov S. Ja. (1982).
Temporal processing of peripheral auditory patterns of speech.
In "Representation of speech in the peripheral auditory system", ed. Carlson R. and Granstrom B., Amsterdam, 165-181.

(2) Dang V.C., Carré R. and Tuffelli D. (1986).
Research on preprocessing by a lateral inhibition.
12th. Int. Cong. on Acoustics, Montreal.

(3) Karnickaya E.G, Mushnikov V.N., Slepokurova N.A. (1973).
Auditory processing of steady-state vowels.
Symp. on Auditory Analysis and Perception of speech, Leningrad.

(4) Lebedev V.G. and Zagoruiko N.G. (1985).
Auditory perception and speech recognition.
Speech communication, **4**, 97-103.

(5) Morishita I. and Yajima A. (1972).
Analysis and simulation of networks of mutually inhibiting neurons.
Kybernetic, **11**, 154-165.

(6) Tokura T. and Morishita I. (1977).
Analysis and simulation of double-layer neural network with mutually inhibiting interconnections.
Biol. Cybernetics, **25**, 83-92.

(7) Shamma S.A. (1985).
Speech processing in the auditory system, II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve.
J. of Acoust. Soc. of Am., **78**, 1622-1632.

SPEECH SOUNDS IN FREQUENCY FOLLOWING RESPONSES OF THE AUDITORY SYSTEM

ELENA A. RADIONOVA

Lab. of Hearing Physiology, I.P.Pavlov Institute
of Physiology, Leningrad, USSR, 199034

ABSTRACT

Speech sound signals are much better reproduced in the summed neuronal activity than in the activity of single neurons. The most complete reproduction is observed at the lower levels of the auditory system. At higher levels only information concerning the signal periodicity may be partly retained.

At present it is well known that sound signal parameters may be but poorly reflected in the impulse responses of single neurons of the auditory system. This is the case even for simple pure-tone signals. For instance, as it was found for the cat, at least about half of neurons from the cochlear nucleus (which is the first auditory brain level, receiving the whole auditory information from the ipsilateral cochlea of the inner ear via the auditory nerve fibers) show a pronounced nonlinear relation between the signal frequency and the impulse response value: at signal intensity of 50-80 dB this function has several (up to about ten) maxima separated by different frequency intervals over the frequency range from about 1 to 20 kHz. Besides, the time patterns of a single neuron responses often show very slight differences (if any) over a wide range of signal frequencies. At higher levels of the auditory system the correspondence between the sound signal parameters and the single neuron responses declines even to a greater extent. The above properties as well as some others, make each neuron, when alone, unable to indicate what kind of a sound signal comes to the ear.

Meanwhile the summed response of a number of neurons may give appropriate information about the signal presented to the ear. This possibility was first proposed as the so called "volley" principle /1/ and was then supported by the experiments with the "Frequency Following Response" (FFR) registered from the lower levels of the auditory system as the summed response of a group of neurons with the near spatial positions within a given brain region.

The FFR is the result of the activity of a number of neurons whose impulse responses are synchronized with a certain phase of the tonal signal. The upper frequency limit of this synchronization was reported as at least 5 kHz for the auditory nerve fibers, about 6-6.5 kHz for the cochlear nucleus level, with a pronounced diminution of this value at the higher auditory centers: to about 1.5 kHz at the midbrain level (inferior colliculus) and to about 1 kHz or even less resp. at the medial geniculate body and the auditory cortex levels /2/. Thus, the widest frequency range reproduced in the FFR is related to the lower levels of the auditory system.

It was found that FFR not only followed the tone frequency but could also reproduce the wave form of rather complicated sound stimuli. This was especially well observed at the cochlear nucleus level with complex harmonic signals containing 2 to 6 harmonics, as well as with the sound speech signals. For instance, when two-tone complex of the second and the third harmonics (or of some others) was presented while varying the signal waveform through variation of the phase of the higher harmonic, the FFR evoked by this complex signal usually reproduced rather precisely the waveform of the signal, with almost all the details: for each oscillation in the complex signal wave a cooresponding deflection in the FFR could be observed.

However, in some cases nonlinear phenomena took place, when one of the signal harmonics, usually the lower one, was fully suppressed in the FFR. It seems of interest that this suppression depended on the phase of the higher harmonic and could be observed within a certain phase range only.

Inspite of some nonlinear features, the FFR of the cochlear nucleus can reproduce rather well the sound speech presented to the ipsilateral ear. The speech reproduced in the FFR is well distinguished, the masculine and feminine voices can be well distinguished as well and even the person can be sometimes identified according to the voice reproduced in the FFR. The prosodic characteristics of the speech are also well reproduced in the FFR of the cochlear nucleus.

Thus, it may be concluded that at the cochlear nucleus level the sound speech signal can be rather fully described in the summed activity of the population of the neurons. The characteristic frequencies (CFs) of neurons forming such a population should be relatively similar, since these neurons are obvioslyy positioned rather near each other as their summed activity, the FFR,is recorded with the help of the same electrode. Besides, for the best reproduction of the speech signals, the CFs of the neurons whose activity forms FFR should lie at the upper frequency limit of the speech sound range (for instance, of about 4 kHz) or higher: such neurons would respond to the wide frequency range below the values slightly higher than the CFs. The neurons with the lower CFs would not respond to high components of the speech signals. Therefore only populations of neurons with sufficiently high characteristic frequencies would reproduce sound speech in their summed FFR.

At the higher level of the auditory system, namely, at the central nucleus of the inferior colliculus (IC) speech sounds may be reproduced in the FFR much more roughly than at the cochlear nucleus level. The sound speech, as it is reproduced in the FFR of the inferior colliculus, is not distinguishable now, though speech prosodic characteristics are well pronounced when the FFR is listened to through the audio reproduction system. These restrictions in the description of the sound speech in the FFR are obviously connected with the restricted frequency range (not higher than 1.5-1.7 kHz) reproduced in the FFR at the level of the inferior colliculus. This, in turn, may be connected with the properties of single neurons forming the FFR: the neuronal impulse activity is much lower than the activity at the cochlear nucleus level, its variability is greater, as well as nonlinear effects due in particular to neuronal inhibitory interconnections.

When analysing the inferior colliculus FFR to complex sound signals containing 2-6 harmonics it may be seen that, unlike the cochlear nucleus, the FFR from the IC does not describe the complex waveform in detail. Usually only the periodicity of the signal waveform is reproduced in the FFR quite well, especially the periodical sharp changes in the signal amplitude. Other, more delicate, though rather essential changes in the signal waveform produced for example by phase variations of the signal components may still be reflected in the FFR-waveform, but with many details lost.

As to the highest, cortical level of the auditory system, the FFR may be registered here only in a narrow low frequency range, for instance, of about 100 Hz /3/. Cortical FFR can reproduce only the low frequency envelope of complex sound signals, the speech sounds including, without any details concerning their waveform. The fact is, that neurons of the highest levels of the auditory system, the medial geniculate body and the auditory cortex, are greatly specialized concerning different sound parameters. Besides, their responses are variable and nonstationary to a great extent, with inhibitory effects well pronounced. These neuronal properties result in great restriction of both the FFR amplitude and frequency range, which makes the sound speech reproduction impossible in the summed neuronal activity at these auditory regions.

Thus, rather complete analogue description of speech signals in the neuronal FFR can be observed only at lower levels of the auditory system. At higher levels a restricted description based on definite signal features seems to substitute step-by-step the full description of the signals: now only pronounced changes in the signal envelope or transients may be reflected in the summed neuronal responses. Meanwhile it may be thought that at higher levels of the auditory system there would be a possibility to extract in a way the information from the lower levels concerning more complete description of complex signals, the speech signals including. It is not clear yet how it could be done. Still it may be supposed that the main function of the higher auditory centers (together with some other brain high centers) would be to form general ideas or at least sound images, i.e. mental pictures of the human voice, of the animal cry, of the step sound etc., which should be necessarily connected with a loss or diminution of the information relating to particular details of the real sound signals.

REFERENCES

/1/ E. G. Wever, "Theory of Hearing", Wiley, 1949.

/2/ E. A. Radionova, "Neurophysiological Studies of the Monaural Phase Sensitivity of the Auditory System". In: Sensory Systems, "Nayka", 1982, 72-86 (in Russian).

/3/ M. Steinschneider, J. Arezzo, H. G. Vaughan, Jr., "Phase-locked Cortical Responses to a Human Speech Sound and Low-frequency Tones in the Monkey". Brain Res., 1980, v. 198, 75-84.

THE ASSOCIATIVE BIONIC APPROACH TO THE DEVELOPMENT OF THE SPEECH SIGNAL PROCESSING CENTRAL MECHANISMS

A.A.Kharlamov

Moscow

The goal of the report is to show possibilities of the associative bionic approach to construction of the model of the speech signal processing central mechanisms. The basis of this approach is the data processing in the dinamic associative memory block (DAMB) and its usage for constructing hierarchical structure (HS) for phonetic, lexical and syntactical processing of speech.

INTRODUCTION

The analysis of the data representation methods that has been used in the speech recognition system shows that graphs (matrix) representation is the best. For example in the form of an evidently given graph or a hidden Markow model. The methods in question have some disadvantages: data representation inflexibility, graph labour consuming forming or need in large computer power. Hardware realisation of graph (matrix) data representation by the system of DAMB is free of these disadvantages.

The DAMB construction is based on some biological facts about structure and properties of neurons and their pulls. The uniform structure and a simple data processing algorithm allow to produce the DAMB using the microintegral technology as an integral system on the sylicon plate.

THE MODEL OF THE SPEECH SIGNAL PROCESSING CENTRAL MECHANISMS.

The main functions of DAMBs and HS formed of them are: storage of data with compact packing, reproducing it with the help of context association, and the statistical processing of the input data by picking out the number of different occurency frequency elements (i.e. vocabularies); extraction of vocabulary word relations in the input data - that allows to reconstruct the inherent input data structure. The above functions give us the possibility to model phonetic, lexical and syntactical levels of data processing by the HS of DAMB. It is supposed that the acoustical speech signal is preprocessed, optimal in each specific case, which is not discussed here.

The model in question consists of two data processing chanals: the coarse one and the precise one. When training the precise chanal performs compilation of a phonotype vocabulary {P} and on its base the vocabularies of lexical level sublevels: i.e. the word vocabulary {L} and the morph one { M } . When training the coarse chanal forms the syllable-phoneme vocabulary {SP} . The unit segmentation of the corresponding levels is performed by DAMBs as the natural feature of the data processing in them. The type of a unit is determined by the DAMB parameters (of a hiperqube dimension).

In the process of recognition the data in the input of the lexical level is represented in terms of syllable-phoneme vocabulary, i.e. syllable representation, as a number of syllable type sequences in the corresponding words, or morphs L^{sp}_{j} = = (SP_I, SP_2, ... , SP_i). The whole number of lexical level input is divided into the subvocabularies according to the equal syllable representation principle, these subvocabularies are indexed by this representation L^{SP}. If the vocabulary consists of only one lexical unit or there are high level constraints (contextual), which allow us to choose the necessary alternative, the recognition process is stopped. Let us assume $L_k \equiv L^{sp}_{j}$.

If it is necessary to choose a lexical unit from the subvocabulary {L} L^{sp}_{j}, which is indexed by the given syllable representation L^{sp}_{j}, the precise chanal is used. In this subvocabulary the lexical units L_k and L_1 (L_k = (P_I, P_2, ... , P_m, ...)) are divided by one or more phonetic element types P_m. To divide phonotypes the preprocessing form is used which is associatively related to the phonetic element type P_m. This division uniquelly determines the lexical level unit L_k.

The higher levels (from the lexical point of view), the syntactical, for instance, bring in additional (contextual) constraints on lexical level unit L_k choosing. In the process of training in the syntactical level the words and morphs relations (i.e. inflections) vocabulary {F} and the type phrase vocabulary{Pr} are compiled.

The above mentioned structure can be realized as a HS from DAMB.

THE DAMB FORMALIZATION.

The DAMB is a net of neuroliked elements (NE) with the input signal time summaring, which is accomplished by shift register of n cells - the model of the generalized dendrit . The DAMB consists of 2^n NEs and each of them models one of the n-dimensional unitary hiperqube node in the signal space.

The binary sequence $A = (\ldots, a_{-I}, a_0, a_I, \ldots, a_i, \ldots; a_i \in \{0, I\})$, the input sequence for the DAMB, is mapped into the hiperqube as a directed sequence of the nodes - trajectory $\hat{A} = F(A)$. Each n of the symbols from the sequence $(a_{i-n-I}, a_{i-n-2}, \ldots, a_i)$ corresponds to the node $\hat{a}_i$ with the given coordinates. The initial sequence A can be restored from this trajectory $A = F^{-I}(\hat{A})$. F mapping has the property of associative addressing to the data with the help of context association (n sequential symbols).

The DAMB can operate in one of the following three modes: (I) training or perception; (2) reproduction; and (3)structural processing.

The training mode.

In the process of training NEs of the DAMB change their inner state under the influence of an input sequence. This changing means that the memory function H is inserted to the nodes of the hiperqube. That is why the trajectory is stored (i.e. in the current node the sequence next symbol is stored in case the record is autoassociative, or the informational sequence current symbol is stored in case the record is heteroassciative).

Let us provide function H with the thresholding properties. This allows to process the data that is stored in DAMB, statistically. The processing allows to compile the vocabulary of ivents $\{\hat{B}\}$, from the input of the DAMB as the number of sequences $\{A\} : \{\hat{B}\} = F(\{A\})$. Under the influence of the threshold value h of function H words of the vocabulary are either the union of input ivent trajectories (in that case the whole data is stored), or fuzzy or precize intersection (in this case more or less common part of data is stored). The identical parts of the sequences or events are mapped into the same chain, and different parts are mapped into the different ones. As a result a directed metrized graph or a graph-word is formed at the nodes of the hiperqubes. The trajectory attenuating models the forgetting process.

The reproduction mode.

The stored data reproduction using of F^{-I} mapping allows to recognize the input sequence by comparing it with the reproducing one according to a measure system. Only the pretrained DAMB can operate under reproducing. The n-member segment of the DAMB input addresses to one of the hiperqube nodes (to one of the NE) where some data is stored (the inner states were changed). The trained NE answer, added to the input n-member segment, determines a new address and thus a new node is addressed. And so on.

The structural processing mode.

Under the structural processing the input sequence is compared with the compiled in the DAMB vocabulary with the sequence segments changed by zero sequences, if these parts of sequences, corresponding to the parts of trajectories, coincide with the node sequences (chains of the vocabulary graph-words). Thus the special F_c^{-I} mapping allows to eliminate the vocabulary data from the input sequence, and to preserve only the relations of the vocabulary words. The abbreviation sequence (AS) $C = F_c^{-I}[F\ (A)\ ,\ \{\hat{B}\}]$ is formed. The mechanism of the AS forming allows to use the DAMBs for structural data processing within the DAMB hierarchical structure, as some rarefied parallel data flows can be united into one flow without losses.

The hierarchical structure of the DAMB.

There are some parallel processes $\{A\}^I \oplus \{A\}^2 \oplus \ldots \oplus \{A\}^q = \{\mathfrak{A}\}$ - that is the situation - in the HS input. The vocabularies $\{\hat{B}\}$ of the most frequently occured situation $\{\mathfrak{A}\}$ events are formed in the first level DAMBs. After the vocabularies compilation, if sequences $\{A\}^q$ are given in the first level DAMB inputs, the AS are formed in their outputs. These AS converge in the second level DAMB inputs and compile the vocabularies $\{\hat{D}\}^r$, $(r = I, \ldots, R;\ R < Q)$ in the DAMBs. Thus the input situation model is formed in the HS as a repeatedly enclosed directed metrized graph. In that graph the graph-words of the low level vocabularies are enclosed into the corresponding parts of the high level graph-words. The HS curtale the input data in the down-up direction and vice versa. That HS property allows to reproduce the stored situation with the help of association both from high level and low level. (Thus the HS can be used as an analyser and an effector as well).

CONCLUSION

The report is devoted to the development of a theoretical model of data processing at the phonetic, lexical and syntactical levels. That model allows to create a device for structural speech signal processing. That device automatically performs compilation of vocabularies of those level units, reconstruction of those level grammars, and recognition of the input events by matching them with the compiled vocabulary units.

INTERVAL OF SPECTRAL INFORMATION ACCUMULATION IN PERCEPTION OF NON -STATIONARY VOWELS

INNA CHISTOVICH TAISA MALINNIKOVA ELENA OGORODNIKOVA

Lab.of Speech Physiology, Pavlov Institute of Physiology
Leningrad, USSR, 199164

ABSTRACT

The results of identification experiments indicate that the interval of the auditory spectrum accumulation exceeds 20 ms. The data is compatible with the supposition that the accumulation interval is comparable to the duration of the vowel.

INTRODUCTION

This work is a development of the study of the spectrum shape processing started by L.Chistovich. She suggested a new approach to this problem which allowed to demonstrate that the information about spectrum shape was accumulated over the vowel length, but the data concerning the accumulation mechanism was rather contradictory (see /1/ for a review).

The fact of accumulation can be explained by either one of the following hypotheses:

1. The running auditory spectrum is considerably smoothed in time before extraction of the phonetically relevant parameters.
2. The parameters characterizing spectrum shape are extracted from practically unsmoothed auditory spectrum and then are accumulated. It is evident that in this case the extracted parameters depend strongly on the sampling instants.

The choice between these hypotheses will influence the direction of future studies. If hypothesis 2 is correct, it probably means that the sampling is synchronized to the fundamental tone, and it is necessary to investigate the synchronization mechanism. If hypothesis 1 is correct, the sampling with constant interval or the sampling at the ends of the segments (synchronized to segmentation marks) is to be considered.

We discuss here the previously obtained data /2,3/ and present new experiments designed to test these hypotheses (some of the experiments were suggested by L.Chistovich).

In all the experiments discussed here the same type of the signals, specially designed to have no dynamic cues (formant transitions), was used /4/. To simplify their description we shall introduce some designations.

The signal is a train of n formant pulses. One-formant pulse s_i is a short tonal pulse with triangular time envelope, F_i is the tone frequency, L_i is its intensity. $v_{ij}=s_i+s_j$ is a two-formant pulse, $w_{ijk}=s_i+s_j+s_k$ is a three-formant pulse. The stationary signals, consisting of identical pulses, are denoted S_i, V_{ij} and W_{ijk} respectively. Signals (S_iS_j) contain s_i and s_j; (S_iV_{jk}) contain s_i and v_{jk}; n_i is the number of s_i pulses in such signals. T_0 is the interval between the onsets of two identical pulses, T is the interval between the onsets of any two successive pulses, t is the interval between the onsets of s_i and s_j (or v_{jk}).

The results compatible with hypothesis 1 were obtained in several experiments /1, 3,4/. The most striking is the fact that increasing n_i in (S_iV_{ij}) causes the same changes in identification as increasing L_i in V_{ij} /3/. The main result against hypothesis 1 was obtained in the experiment on identification of S_i, V_{ij} and (S_iS_j) for $T \approx 10$ ms /2/. Signals (S_iS_j) were not identified with the same phonemes as V_{ij}. What is more, (S_iS_j) were mostly identified with either the same phonemes as S_i and S_j, or with [ɨ]. Obviously such result is possible only if hypothesis 2 is correct and the auditory spectrum is so little smoothed that a formant pulse does not affect the next pulse after 10 ms delay. The great number of [ɨ] responses was explained by the fact that Russian subjects often use [ɨ] as a label for indefinite vowels.

As this is the only experiment directly contradicting hypothesis 1, we tried to check its results in Experiments 1 and 2.

EXPERIMENT 1

In this experiment we obtained the identification data on signals S_i and (S_iS_j) for a wide range of F_iF_j. First, we wanted to check if (S_iS_j) would be identified as S_i, S_j or [ɨ] for other values of F_i,F_j($F_i << F_j$) than those used in /2/. Then, there were some indications in /2/ that the "local center of gravity effect" (LCGE) could be observed on (S_iS_j). LCGE manifests itself in the fact that a signal with formant frequencies F_1,F_2, $F_2-F_1 < 3\div4$ Bark,is phonetically similar to a one-formant signal with formant frequency F, $F_1 < F < F_2$ /1/. If a) hypothesis 2 is correct, and b) LCGE is a result of smoothing of the auditory spectrum in the frequency domain, LCGE should disappear when the formants are sufficiently separated in time.

Signals of Tests 1,2,3: n=12, T=20 ms or 14 ms, F_i=0.3,0.65,1.15,1.9,3.0 kHz. In (S_iS_j) j=i+2, n_j=4,6,8.

Signals of Test 4: n=8, T=20 ms, F_i=0.3, 0.45,0.65,0.85,1.15,1.5 kHz. In (S_iS_j) j=i+2, n_j=2,4,6.

The results of Tests 1,2,3 were combined, as no significant differences were found between the tests. The results of Tests 1,2,3 do not agree with /2/. All the 3 subjects responded to (S_iS_j) quite differently than to S_i and S_j. Subjects A and B practically always identified S_1, S_3 and S_5 as [u], [a], [i], (the corresponding response rates for 90 trials are 1., 1., 0.99 for A; 1., 0.96, 1., for B). Maximal (for 3 values of n_3) rate of (neither [u] nor [a])responses to (S_1S_3) is 0.43 for A, 0.62 for B. Maximal rate of (neither [a] nor [i]) responses to (S_3S_5) is 0.87 for A, 0.46 for B. Only subject C frequently identified (S_iS_j) with [ɨ] ; A and B practically never used this phoneme.

In respect of LCGE the results were qualitatively the same as for stationary signals. LCGE was observed in Test 4, where $F_j-F_i \approx 3\div3.5$ Bark: (S_iS_j) were perceived as similar to S_{i+1}. The square distance between the response distributions served as a measure of similarity. In 8 cases out of 12 (3 subjects x 4 F_i, F_j combinations) at least one of three (S_iS_j) with n_j=2,4,6 was nearer to S_{i+1} than to S_i or S_j. In the 4 remaining cases the distances from (S_iS_j) to S_{i+1} and to S_i or S_j were approximately equal (and small). Thus, LCGE does not disappear when the formants are separated in time.

EXPERIMENT 2

In this experiment we tried, using the same F_1,F_2 combination as in /2/, to find the minimal time lag t at which (S_1S_2) begins to be perceived as a mixture of S_1 and S_2 and not as V_{12}.

Signals of Test 1: F_1=0.75 kHz. For S_1,

S_2, V_{12} n=6, T_0=20 ms. For (S_1S_2) $n_1=n_2=6$, $T_0=|t|+20$ ms, t=±5, ±10, ±15, ±20 ms. We found that for all t values the (S_1S_2) response distribution is not a mixture of responses to S_1 and S_2. All the 5 subjects identified S_2 with [i] (response rate $p_{[i]} \geqslant 0.93$); for all $(S_1 S_2)$ $p_{[i]} \leqslant 0.125$. 3 subjects identify S_1 with [O] ($p_{[O]} \geqslant 0.95$) and never use [O] in responses to (S_1S_2). Only E.Z. gave a lot of [ɨ] responses to (S_1S_2), but she also responded to V_{12} with $p_{[ɨ]}$=0.5.Other subjects had $p_{[ɨ]} \leqslant 0.125$ for all signals. The responses of two subjects were almost independent of t: $p_{[ɛ]}$ fluctuated from 0.58 to 0.87 for |t| =0 ÷ 20 ms. Others exhibited a strong dependence of identification on t. Increase of |t| increased $p_{[e]}$ and decreased $p_{[ɛ]}$ for S.Zh; increased $p_{[ɨ]}$ and decreased $p_{[ɛ]}$ for E.Z; T.M. changed responses from [ɑ] to [ɛ] and then to [æ]. Thus,the results of one subject (E.Z.) only are similar to those obtained in /2/. The dependence of identification on t is, we suppose, really the dependence on duration or/and pitch, which were not constant. The results of Test 2 support this supposition. Signals of Test 2: F_1=0.75 kHz, F_2=2.5 kHz, T_0 =16 ms, $n_1=n_2=12$, t=0, ±4, +8 ms. Four of the subjects of Test 1 took part in Test 2. The table shows the variation of $p_{[ɛ]}$ when t was varied from -5 ms to +5 ms in test 1 and from -4 ms to +8 ms in test 2.

	T.M.	S.Zh.	E.Z.	I.Ch.
Test 1	0.37	0.38	0.2	0.23
Test 2	0.2	0.17	0.07	0.1

As can be seen, though the t range for Test 2 is larger, variation of $p_{[ɛ]}$ is always smaller when duration and T_0 of signals are kept constant.

EXPERIMENT 3

The goal of this experiment was to find out if (S_2V_{13}) could be identified with the same phonemes as W_{123}, and if varying n_2 in (S_2V_{13}) would lead to the same changes in identification as varying L_2 in W_{123}. It is only possible if hypothesis 1 is correct and the auditory spectrum is integrated over several formant pulses. Such an effect was observed for V_{ij} and (S_iV_{ij}) /3/. As (S_2V_{13}) contain no three-formant pulses, the equivalence of varying n_2 and L_2 would be even a stronger argument for hypothesis 1 than /3/.
Signals: F_1=0.3 kHz, F_2=1.1 kHz, F_3=3 kHz, n=12, T=14 ms. For W_{123} $L_1=L_3$, $\Delta L=L_2-L_1$=±20, ±10, 0 dB. For (S_2V_{13}) $L_1=L_2=L_3$, n_2=3,6,9.
The responses to W_{123} strongly depended on Δ L. When Δ L decreased from + ∞ (S_2) to - ∞ (V_{13}) the obtained sequences of most probable responses were [ɑɛi] for T.M., [ɑɛɨ] for E.Z., [ɑɛɨi] for E.K. and I.Ch., [ɑɛɨu] for S.Zh. All the subjects identified (S_2V_{13}) with the same phonemes as W_{123}, and increasing n_2 in (S_2V_{13}) had the same effect on the identification as increasing Δ L in W_{123}.
To evaluate this effect quantitatively we approximated the (S_2V_{13}) response distribution P_n by the weighted sum of two (closest to P_n) W_{123} response distributions: $P_n=k_1P_1+k_2P_2$. The obtained k_1, k_2 and residual error d^2 are shown in the table. Indices of k indicate Δ L of corresponding W_{123}.
It can be seen from the table that d^2 are quite small. Increasing n_2 is equivalent to increasing Δ L, but Δ L range corresponding to variation of n_2 from 3 to 9 is different for different subjects (from 0 ÷ 10 dB for I.Ch. to -10 ÷ 20 dB for T.M.). Thus, all the 3 experiments are compatible with hypothesis 1 and contradict hypothesis 2. The duration of the time window used for smoothing of the running auditory spectrum should, according to Experiment 2, exceed 20 ms.

	n_2	k_{+20}	k_{+10}	k_0	k_{-10}	d^2
E.K.	9	0.09	0.98			0.004
	6		0.44	0.69		0.042
	3			0.76	0.33	0.042
E.Z.	9	0.07	0.88			0.003
	6		0.17	0.90		0.016
	3			0.61	0.37	0.001
T.M.	9	0.96	0.02			0.001
	6		0.29	0.88		0.086
	3			0.12	0.92	0.024
I.Ch.	9		1.			0.091
	6		0.52	0.60		0.094
	3		0.04	0.97		0.016
S.Zh.	9	0.09	0.85			0.008
	6			1.		0.082
	3			0.67	0.47	0.054

The results of Experiment 3 corroborate the data of /3/ and suggest the duration of time window comparable to the duration of the signal. If this is the case, some sort of amplitude compression or normalization must precede the smoothing, as the identification of (S_iV_{ij}) very weakly depends on the amplitude of V_{ij} pulses /3/. All our results concern only the spectrum shape processing. The formant transitions are probably processed by the system with quite different temporal properties.

REFERENCES

/1/ Chistovich L.A. Central auditory processing of peripheral vowel spectra. - J.Acoust.Soc.Amer.,1985, v.77, pp.789-805.

/2/ Chistovich L.A., Ogorodnikova E.A. Temporal processing of spectral data in vowel perception. - Speech Communication, 1982, v.1, pp.45-54.

/3/ Chistovich L.A., Malinnikova T.G. Processing and accumulation of spectrum shape information over the vowel duration. - Speech Communication, 1984, v.3, pp.361-370.

/4/ Чистович Л.А., Чихман В.Н., Огородникова Е.А. Новый подход к определению фонетической близости стимулов и его проверка в автоматизированном эксперименте. - Физиол.журн. СССР, I98I, т.67, с.7С4-7IС.

SOUND SHAPE OF LANGUAGE AND CEREBRAL ASYMMETRY

VADIM DEGLIN OLGA TRACHENKO TATIANA CHERNIGOVSKAYA

Lab.of Brain Functional Asymmetry, I.M.Sechenov Institute
of Evolutionary Physiology and Biochemistry, Acad.of Sciences, Leningrad, USSR 194223

ABSTRACT

The purpose of the study was to analyze cerebral asymmetry in speech sound processing. It is suggested that difference in hemispheric ability is of a qualitative nature: left hemisphere provides for correct phonemic analysis while right hemispheric competence is in prosodic arrangement of speech material, its quick global recognition. The research was performed in normal subjects and in patients of psychiatric clinics after unilateral electroconvulsive therapy.

In the beginning of the century the outstanding neurophysiologist I.P.Pavlov and the no less prominent hand in the science of language L.V.Shcherba were surprisingly unanimous in suggesting that tc know the laws of a functioning system one must examine its disturbances. I.P.Pavlov's words refer to the complex forms of brain activity, while the words of L.V.Shcherba - to language. Emerging in the middle of the century, neurolinguistics seems to integrate both applications of the idea. On the one hand, while studying disorders of speech processing caused by pathology it reveals cerebral organization of speech functions, on the other - the data obtained in this way explain many disputable questions of linguistic system structure. Among the founders of neurolinguistics one can name two eminent experts of science of this century - philologist R.Jakobson and neuropsychologist A.Luria. It was they who demonstrated the great value of "negative data" - both linguistic and cerebral.

Aphasiological tradition has postulated that all linguistic skills are the functions of the left hemisphere (LH), while the right hemisphere (RH) has nothing to do with language. The last decades produced a lot of data undoubtedly prov ing the fact of RH involvement in speech processing. Nowadays it is a generally accepted thesis though accompanied by alternative viewpoints: (1) the abilities of RH are duplicating those of LH, the difference lying in the degree of functions duplication (in full or in part); (2) the difference in hemispheric abilities is of qualitative nature - each contributing to speech activity. We adhere to the second viewpoint. The purpose of the present research was to reveal the involvement of each of the two hemispheres in phonetic material processing. Two procedures were used.

I. Monaural testing of normal subjects. The method enables one to see the hemispheric dominance for verbal processing (perception). Lists of words and nonsense words were presented monaurally to both left and right ear in turn. Reaction time (latent period between stimulus and response) was registered. A hemisphere was decided to be dominant for the analysis if reaction time for the stimulus heard from contralateral ear was shorter.

II. Testing of linguistic skills after unilateral electroconvulsive therapy, used in psychiatry. The seizures were administered to patients of psychiatric clinics. By this means develops a situation when for 30-50 min one hemisphere of the patient is suppressed and incapable of normal activity while the other one is intact and even reciprocally aided. Every patient has been subjected to both right- and leftsided shocks; it was possible to juxtapose the suppression effect of LH and RH in one and the same subject, as well as to compare it with speech functions in patient's normal conditions. The table below illustrates monaural testing that points to the fact that there are no significant ear differences in reaction time for presented nouns, adjectives and verbs.

STIMULI	MEAN REACTION TIME		P
	RIGHT EAR	LEFT EAR	
NOUNS	914 ± 5	918 ± 5	>0,05
ADJECTIVES	784 ± 5	789 ± 4	>0,05
VERBS	778 ± 3	773 ± 4	>0,05
Mean reaction time	828 ± 4	827 ± 4	>0,05
NONSENSE WORDS	1022 ± 4	1043 ± 5	<0,001

This suggests that the degree of each hemispheric involvement in meaningful words analysis is the same. The perception of nonsense words produces completely different results: the reaction time is much shorter when these are presented to the right ear, which show s dominating role of LH. It should be mentioned that the reaction time needed to process nonsense words is much longer than that for the meaningful words. The data oftained show two main differences in processing words and nonsense words: (1) words are equally well processed by both hemispheres, while nonsense words involve LH to a much greater degree. (2) To analyze nonsense words one needs more time. What lies at the basis of such a difference? Let us consider the data obtained after the suppression of one of the hemispheres.

The examination of verbal material discrimination revealed that after LH suppression the comprehension of words, logotomes and phonemes (both consonantal and vowel) is impaired. This phenomenon is in no way due to hearing disturbances: the sensitivity tests show no auditory deficit depending on the side of the hemispheric suppression.

Consonant and vowel discrimination analysis gives grounds for understanding the reason of discrimination impairment after LH suppression. Fig.1 demonstrates typical failures in recognition of speech sounds, i.e. phonemic substitutions.

It can be seen that in their normal conditions patients substitute back vowels by front vowels, voiceless consonants by voiced ones, dentals by bilabials, velars by dentals. Errors in speech sound recognition are, therefore, in no way due to chance, quite the reverse, they demonstrate a kind of regularity - we see neutralization of some phonemic oppositions. After LH suppression the amount of errors increases along with the widening of the range of errors: front vowels are now confused with back vowels, velars with bilabials, dentals with velars etc. Accordingly LH inactivation leads to a considerable decline in discrimination ability caused by phonological system disorder and incapability of distinctive feature analysis. What is important is that the neutralization of phonemic oppoaitions observed after LH suppression, is never seen in patents' normal conditions. The change in the ability of speech sounds recognition after RH suppression is of a different nature: its facilitation is illustrated by fig.1. Misinterpretations concern only conso nant voicing and mixing of high back or front vowels. Such a facilitation of functions after RH suppression is due to LH reciprocal activation.

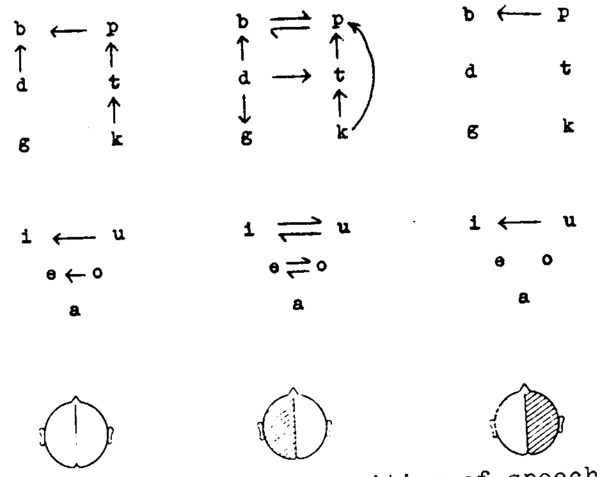


Fig.1. Failures in recognition of speech sounds after unilaterial ECT. Most frequent types. The schemes below represent the state of subjects (inactivated hemisphere is black).

Let us consider now the investigation of phoneme boundaries for stationary vowels. We used 46 vowel-like stimuli with constant F3 and F4 and variable F1 and F2. The subject had to classify each presented stimulus as one of the phonemes. The control testing revealed the same general regularities as already observed in investigation on normal subjects. Neither LH nor RH inactivation affected the average formant position. On the other hand there occured remarkable differences with regard to magnitude of uncertainty(fig.2)

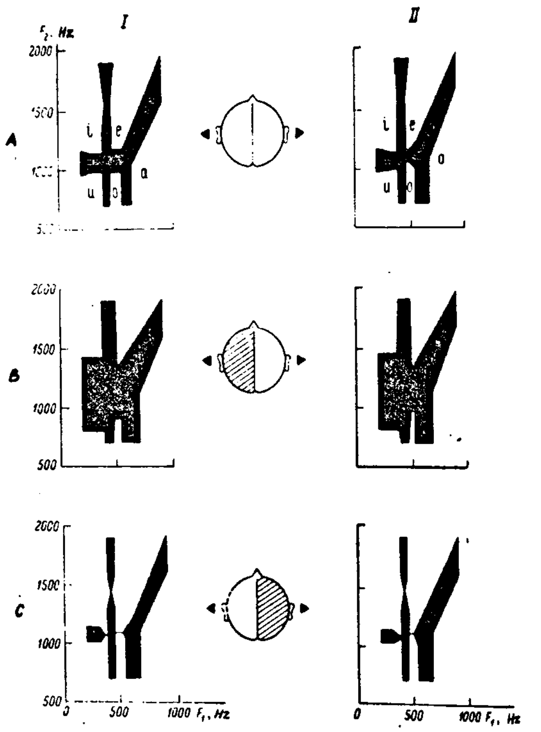


Fig.2. Zones of uncertainty ($\bar{x}\pm\sigma$) in the regions of phoneme boundaries after unilateral ECT. (I) and (II) Performance by the left and right ear respectively. The schemes are the same as in Fig.1.

Thus, after LH suppression the range of areas of uncertainty grew considerably; they were most outspoken in the regions close to F2. In our opinion it is the result of phonemic classification impairment. Knowing that F2 is closely correlated to the dimension front-back we can understand the impairment of front-back vowels discussed earlier. The suppression of RH leads to narrowing (or even disappearance) of the areas of uncertainty, i.e. to the phonemic categorization impairment.

On the whole LH inactivation results in phonological system disorder, reverts the hearing to an infrahuman state when the ability to interpret the significance of F2 is lost. RH inactivation and LH functions improvement leads to phonological coding facilitation even if compared with patients' natural conditions. Thus the research shows that phonological coding is the function of LH. It becomes clear why LH is preferable for perception of nonsense words: to discriminate them one needs the most accurate phonological encoding, since there is no other way for the perception of nonsense words. Then we must assume that RH's spech perception is of a different kind: it proceeds without phonemic encoding. In what way, then? The most probable procedure is to take the word as a whole unit, to use a kind of global, Gestalt perception strategy. There is certain evidence to prove it. The research points to the fact that while discrimination of words and syllables after RH inactivation improves, the number of mistakes of certain types drastically increases: phonemes, syllables and accents could be misplaced and the former ones even totally omitted. Similar mistakes could be found in patients' spontaneous speech production; these are: wrong rhythmical patterns both of words and sentences, monotonous or, vice versa, irrelevantly accentuated speech. Experiments show prosodic perception impairment: with disfunctional RH the identification of intonational patterns - both rendering grammatical meanings - interrogative, imperative and declarative patterns, or, especially, emotional moods - decreases considerably. Under these conditions discrimination of male/female, young/old, familiar/unfamiliar voices becomes impaired.

Fig.3. illustrates the perception of synthetically produced phonemes/a/and/i/ with two varying (high and low) fundamental frequencies.

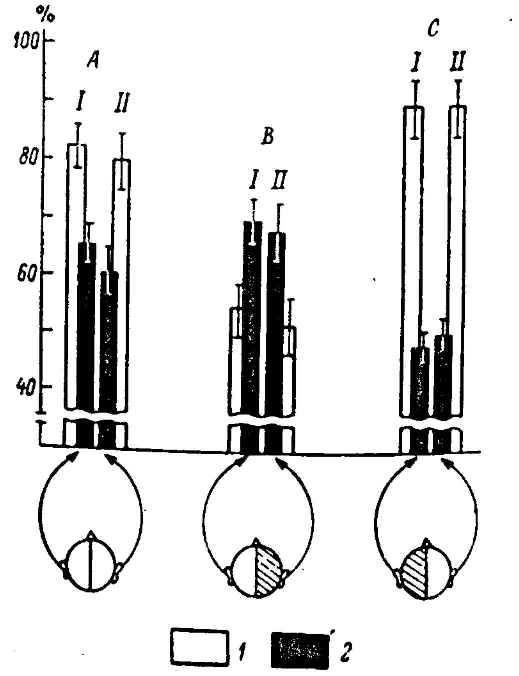


Fig.3. Discrimination of synthetic vowels/a/and/i/ as to their pitch (1) or phonemic quality (2) after unilateral ECT. (I) and (II) and the schemes are the same as in Fig.2.

After RH inactivation the subjects could not determine whether the stimuli were produced by a male or a female but easily identified phonemic quality of the stimuli. On the other hand after LH inactivation the phonemic quality identification was impaired while pitch recognition became more accurate in comparison with patients' normal conditions.

The experiment demonstrates how hemispheric functions specialize even in dealing with the smallest sound segment. We can suggest therefore that it is the RH that is responsible for paralinguistic and prosodic perception. It is well known that prosodic - suprasegmental-features play prominent role in the sound shaping of words - accent contours distinguish individual words, whereas intonation contours distinguish different sentence types. Prosodic features arrange elements to form the units of a higher order: phonemes - to form a word, words - to form a sentence. Consequently the global Gestalt way of perception must be realized by RH structures. However, such a strategy could be used only for previously familiar speech material. It is impossible to discriminate nonsense words using this way of perception.

In relation to the theoretical issues considered in this paper it is obvious that both cerebral hemispheres take part in forming sound shape of language. LH provides for correct phonemic analysis, enabling to reduce sound continuum to functionally relevant segments. The role of RH is to realize global or so called template recognition.

To sum up, the results of the present study suggest that brain has different mechanisms for speech perception. RH mechanism provides for quick orientation in familiar speech material. LH mechanism secures accuracy of discrimination as well as processing of unfamiliar speech samples; but loses in speed of perception. Under usual communicative conditions both mechanisms function simultaneously resulting in optimum speech perception.

МОДИФИЦИРОВАННАЯ ПОЛЮСНАЯ МОДЕЛЬ РЕЧЕВОГО СИГНАЛА И ЕЕ ПРИМЕНЕНИЕ ДЛЯ ПОДАВЛЕНИЯ ШУМА

ЛЮДОВИК ЕВГЕНИЙ КУЗЬМИЧ

Институт кибернетики им. В.М.Глушкова АН УССР
Киев, СССР 252207

АННОТАЦИЯ

Дополнение традиционной полюсной модели (модели линейного предсказания) речевого сигнала моделью квазипериодического сигнала возбуждения, некритичной к наиболее частым ошибкам в выделении основного тона, позволяет использовать периодичность вокализованной речи для ее выделения из смеси с шумом.

ВВЕДЕНИЕ

Традиционная полюсная модель отражает квазипериодичность сигнала возбуждения на вокализованных интервалах сомножителем $(1 - z^{-L})$, стоящим в знаменателе передаточной функции речевого тракта. Таким образом, сигнал возбуждения считается строго периодическим с периодом основного тона L. Известно, однако, что свойство периодичности наблюдается в реальном сигнале лишь в определенной мере, иногда больше, иногда меньше. С одной стороны, форма сигнала изменяется от периода к периоду, с другой - изменяется и сам период.

Вследствие такого рода отличий реального квазипериодического сигнала от модельного строго периодического в процессе анализа возникают ошибки в определении основного тона. Все ошибки могут быть разбиты на три класса:

I) "малые" ошибки в пределах I0-I5% от истинного значения периода;
2) ошибочные значения, кратные периоду или частоте основного тона;
3) грубые ошибки, не коррелирующие с истинным значением периода.

Помимо указанных проблем на этапе анализа неадекватность строго периодической модели вызывает некоторую неестественность синтезируемого вокализованного сигнала.

Попытки /I/ использования на основе этой жесткой модели свойства квазипериодичности для коррекции вокализованных речевых сигналов, искаженных аддитивным шумом, наталкиваются на следующие трудности:

I) необходимо точно определять период основного тона и признак тон/шум по зашумленному сигналу, что и в отсутствие шума является сложной задачей;
2) непонятно, как устанавливать соответствие между отсчетами сигнала из разных периодов и с какими весами следует их усреднять.

Навязывание строгой периодичности с зачастую ошибочным периодом приводит к "смазыванию" динамики сигнала и снижению разборчивости.

Таким образом, имеется потребность в модели квазипериодичности, которая отражала бы изменчивость сигнала от периода к периоду, изменение самого периода, а также была бы некритична к ошибкам в определении периода основного тона.

МОДЕЛЬ

Первый шаг в направлении такой модели можно усмотреть в работе /2/, в которой выражение $(1 - z^{-L})$ заменяется на $(1+g_{L-1}z^{-(L-1)}+g_L z^{-L})$, а параметры g_{L-1}, g_L и L определяются по речевому сигналу. В /2/, однако, эта замена использована фактически только для усовершенствования корреляционного метода выделения основного тона и уточнения амплитуды периодического сигнала возбуждения. На самом же деле такой вариант позволяет учесть в модели изменчивость периода и формы сигнала, но не снимает проблемы создания модели, некритичной к ошибкам.

В настоящей работе предлагается заменить выражение $(1-z^{-L})$ на выражение $(1+\sum_{i=0}^{1} g_{L-i}z^{-(L-i)} + \sum_{i=0}^{2} g_{2L-i}z^{-(2L-i)})$, причем параметры g и L должны определяться по речевому сигналу.

Во временной области предлагаемая модель голосового источника имеет вид:

$$w_n=-\sum_{i=0}^{1} g_{L-i}w_{n-L+i}-\sum_{i=0}^{2} g_{2L-i}w_{n-2L+i}+e_n, \quad (I)$$

где e_n - входной сигнал типа белого шума,
w_n - квазипериодический сигнал голосового возбуждения.

Поскольку в предлагаемой модели учитывается периодичность и с периодом L и с периодом $2L$, вероятность грубых ошибок, некоррелированных с истинным периодом снижается. Кроме того, здесь малосущественно, равен ли период L или $2L$. Таким образом, модель некритична по отношению к довольно частым ошибкам типа удвоения периода или удвоения частоты основного тона.

Результирующая модифицированная полюсная модель получается, если квазипериодический сигнал w_n подать на вход обычной линейной прогнозирующей модели:

$$x_n= -\sum_{i=1}^{m} a_i x_{n-i} + w_n . \quad (2)$$

Передаточная функция модифицированной полюсной модели имеет вид:

$$H(z) = \frac{1}{(\sum_{i=0}^{m} a_i z^{-i})} \times$$

$$\times \frac{1}{(1 + \sum_{i=0}^{1} g_{L-i}z^{-(L-i)}+\sum_{i=0}^{2} g_{2L-i}z^{-(2L-i)})}. \quad (3)$$

ИДЕНТИФИКАЦИЯ МОДЕЛИ

Задача оценивания параметров модели по отрезку речевого сигнала x_n на основе метода максимальнгго правдоподобия или метода линейного предсказания сводится к минимизации следующего критерия:

$$F(a,g,L)=\sum_{i,j} R_{i-j}\sum_{r} a_r a_{r+j}\sum_{s} g_s g_{s+i} , \quad (4)$$

где $R_i = \sum_{n} x_n x_{n+i}$, $a_0 = g_0 = 1$,

$g_i=0, \quad i \notin \{0,L-1,L,2L-2,2L-1,2L\}$.

Для решения этой задачи предлагается итерационный алгоритм, каждая итерация которого состоит из двух этапов.

На первом этапе при фиксированных значениях параметров модели голосового источника (на начальной итерации полагаем $g_i=0, i\neq 0$) осуществляется минимизация по параметрам a, определяющим резонансные свойства речевого тракта, что сводится к решению традиционной для метода линейного предсказания системы уравнений:

$$\sum_{j=1}^{m} R^g_{i-j} a_j = -R^g_i , \quad 1 \quad i \quad m , \quad (5)$$

где $R^g_i=\sum_{j} R_{i-j}\sum_{s} g_s g_{s+j}$ - корреляционная функция сигнала x_n, из которого путем обратной фильтрации устранена информация об основном тоне.

На втором этапе при фиксированных значениях параметров a осуществляется минимизация по параметрам модели голосового источника g и L. При этом для каждого возможного значения L находится минимум критерия (4) по параметрам g что сводится к решению следующей системы уравнений:

$$\sum_{j=L-1}^{L} R^a_{i-j} g_j + \sum_{j=2L-2}^{2L} R^a_{i-j} g_j = -R^a_i , \qquad (6)$$

$$i = L-1, L, 2L-2, 2L-1, 2L,$$

где $R^a_i = \sum_j R_{i-j} \sum_r a_r a_{r+j}$ – корреляционная функция сигнала, из которого путем обратной фильтрации устранена формантная информация.

Значение L , наилучшее для фиксированного значения параметра a , определяется путем перебора.

Поскольку на каждом этапе отыскивается глобальный минимум по соответствущей обобщенной координате, значение критерия монотонно уменьшается от итерации к итерации, или же не изменяется, если найдена точка локального минимума. Незначительное изменение критерия в результате выполнения очередной итерации и неизменность параметра L могут быть приняты за условие останова итерационного алгоритма.

ИДЕНТИФИКАЦИЯ МОДЕЛИ ПО ЗАШУМЛЕННОМУ РЕЧЕВОМУ СИГНАЛУ. КОРРЕКЦИЯ СИГНАЛА

Рассмотрим теперь задачу коррекции вокализованного сигнала, искаженного аддитивным шумом. Пусть на полезный сигнал x_n , порожденный модифицированной полюсной моделью, наложен аддитивный шум d_n с известной спектральной плотностью, так что фактически наблюдаемым является сигнал y_n :

$$y_n = x_n + d_n \ .$$

Задачу выделения полезного сигнала x_n поставим как задачу отыскания максимально правдоподобных оценок сигнала x_n и параметров модели a, g, L .

После перехода в спектральную область и выполнения простых преобразований функции правдоподобия получаем критерий, подлежащий минимизации:

$$F(X,a,g,L,\sigma_e) = 2N^2 \ln \sigma_e + \sum_n |X_n|^2 |A_n|^2 |G_n|^2 / \sigma_e^2 + \sum_n |Y_n - X_n|^2 / |D_n|^2 , \qquad (7)$$

где X_n и Y_n, $0 \leqslant n \leqslant N-1$, – дискретные спектры Фурье искомого и наблюдаемого сигналов соответственно,

$|D_n|^2$ – известный энергетический дискретный спектр шума,

σ_e^2 – дисперсия сигнала возбуждения на входе модифицированной полюсной модели, подлежащая определению,

$$A_n = \sum_{s=0}^{m} a_s e^{-jsn2\pi/N} ,$$

$$G_n = \sum_s g_s e^{-jsn2\pi/N}, \ s \in \{0, L-1, L, 2L-2, 2L-1, 2L\} .$$

Особенностью выведенного критерия является наличие подстраиваемого спектра G_n, обратного спектральной характеристике голосового источника.

Для минимизации критерия предлагается итерационный алгоритм, каждая итерация которого состоит из четырех этапов, причем на каждом этапе определяется глобальный минимум по одной из четырех обобщенных переменных X_n , a , g, L и σ_e при фиксированных значениях остальных.

I-й этап. Минимизация по X_n:

$$X_n = Y_n \frac{\sigma_e^2}{|A_n|^2 |G_n|^2} \Big/ \left(\frac{\sigma_e^2}{|A_n|^2 |G_n|^2} + |D_n|^2 \right)$$

П-й этап. Минимизация по a сводится к решению системы уравнений (5) с коэффициентами

$$R^g_i = \sum_n |X_n|^2 |G_n|^2 \cos(2\pi in/N).$$

Ш-й этап. Минимизация по параметрам источника осуществляется точно так же, как в описанном выше алгоритме идентификации модифицированной полюсной модели по незашумленному сигналу, при этом

$$R^a_i = \sum_n |X_n|^2 |A_n|^2 \cos(2\pi in/N).$$

IУ-й этап. Минимизация по σ_e :

$$\sigma_e^2 = \sum_n |X_n|^2 \cdot |A_n|^2 \cdot |G_n|^2 / N^2 \ .$$

Как и в случае первого алгоритма критерий монотонно уменьшается от итерации к итерации, либо остается неизменным, если значения обобщенных переменных соответствуют локальному минимуму. Условие останова такое же, как и в первом алгоритме.

На основании полученной оценки спектра незашумленного сигнала X_n путем обратного преобразования Фурье можно вычислить оценку отсчетов исходного речевого сигнала во временной области.

ЗАКЛЮЧЕНИЕ

Предложенная модель может быть полезна при анализе речевых сигналов, поскольку она позволяет осуществить развязку формантной информации и информации, связанной с квазипериодическим сигналом голосового возбуждения.

Достоинством модели является ее некритичность к ошибкам в определении периода основного тона типа удвоения периода или удвоения частоты, а также отсутствие необходимости принимать решение о признаке тон/шум.

С точки зрения задачи коррекции зашумленных вокализованных сигналов в этой модели важно то, что степень периодичности и веса при усреднении различных периодов определяются самим зашумленным речевым сигналом, в отличие от других подходов, включающих элемент произвола.

Литература

/I/. Лим Дж.С., Оппенхайм А.В. Коррекция и сжатие спектра зашумленных речевых сигналов. ТИИЭР, т. 67, № I2, 1979.

/2/. Kwon S.Y., Goldberg A.J. An Enhanced LPC Vocoder with No Voiced/Unvoiced Switch. IEEE Transactions on ASSP, vol ASSP-32, N 4, 1984.

POLYTONAL ANALYSIS-SYNTHESIS SYSTEM FOR SPEECH ENHANCEMENT

V. MAKHONIN

Institute for problems of information transmission
USSR Academy of Sciences
101447 ГСП-4 Moscow USSR

ABSTRACT

Most of techniques for speech enhancement using noise suppression suppress discontinuously modulated speech oscil - lation too.At result of discontinuously modulated oscillations suppression the quality of selected speech signal is decreased.One way of enhancing speech in an additive noise is to perform a functional decomposition of a frame of noisy speech and to attenuate a particular transformed component depending on how much the measured component pulsation power exceed an estimate of the background noise.Using a Walsh-Hadamard blocks connected by strings of zeroes results in a new class of suppression curves which permits a tradeoff of noise suppresion against speech distortion.The algorithm has been implemented in "Eclips C-330" minicomputer.

INTRODUCTION

The security of vocoders,speech recognition devises and speech synthesizers is less than natural one . It was found by speech testing that technicians ignore high frequency part of telephone spectra / 1 /. The level of high frequency oscillations is low,but these oscillations transmit an important information for human hearing becouse of voise modulations.

The signal preemphasis before an A/D convertion and the discontinuous demodulation of speech oscillations select speech data for proper display and hearing expe - riments.

Different techniques have been compared for enhancing the noisy speech.The results seem to point out the superiority of block-cascade Walsh-Hadamard transformations especially concerning heavy noise environment. Lines of Hadamard matrix are strings and elements of these strings are +1 and -1. Some of these lines seem like clipped harmonics , while others seem like clipped phase shift keyed modulated oscillations.

Thus one string of ±1 had been taken from Hadamard matrix repeated periodically together with string of zeroes produce sequence of block clipped waves. The set of such sequenses and a frame of samples have been taken from speech signal produce the set of scalar products.

Next step in frame processing is neglecting of oscillation components those pulsations power are less then corresponding background thresholds. Neglecting reduces noise components , while strongest poly - tonal components represent voised speech signal and its waveform microvariations by sequence of such accords.

The work presented here is continuation of works have been carried out ten years ago during a stage in ULB / 2 / and later / 3,4 /. Concerning speech signal microvariations representation few researches have been done / 5,6 /. One was performed by M.Rohtla et al. / 7 /.

OUTLINE OF THE POLYTONAL TRANSFORMATIONS

The described computational tehnique has been employed according to the frame by frame processing mode. While samling frequency = 20 kHz and frame duration = 51.2 msec each frame consist of 1024 samples.Frames are overlapped on half of frame size,i.e. 25.6 msec.

The frame is transfrrmed into set of scalar products of speech signal samples and elements of an elementary subsequence. While transformation is block-cascade Walsh-Hadamard transformation , the elementary subsequence consist of ± 1 inside and 0-es outside of blocks , accordingly.

For example , if tone period = 128 samples , line seven from an Hadamard transform matrix formed of eight lines and eight columns has been chosen and the chosen block is second , the elementary subsequence is represented as follows:

0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,-1,-1,
+1,+1,-1,-1,+1,+1,+1,+1,-1,-1,+1,+1,-1,-1,
0,0,0,(in all 128-16=112 zeroes),-1,-1,
+1,+1,-1,-1,+1,+1,+1,+1,-1,-1,+1,+1,-1,-1,
0,0,0,(in all 112 zeroes again), -1,-1,
et cetera up to 1024th.

After transformations of 1024 samples into 1280 scalar products (20 tones*8 lines * 8 blocks = 1280 elementary subseq-uences) have been fulfilled the second cascade of transformations to be fulfilled. By those second transformations one get a set of estimations pitch pulsations sincronuously to tones scanned. Estimations taken from this set are compared with corresponding thresholds , those are greather then thresholds to be selested,others to be neglected and their values be - come equal zeroes.

Next steps are related with choises of tones from scanned scale. To choose tones estimations corresponded to scanned tone collected together to transform in logarithmic scale and to compare those partial pulsation levels with corresponding thresholds and select tones those levels are greather... and so on.

It is possible to print the results of computations in format as data represented in Tables N° N° 1,2,3,4,5,6 or to rewrite a linear combinaton of selected oscillations on the disk memory and to convert records by D/A converter for hearing experiments.

So ,a usage of described computation process permits to select so many tones as nessesary to represent signal microvaritions and to enlarge frame size to suppres noise oscillations.

COMMENTS TO TABLES

A table of results of polytonal analysis consists of two parts,left with "stars" instead of zeroes and integer nombers I2, which represent pulsation level distribution between lines of Hadamard matrix. Signal time is growing from row to row on the amount of overlap between frames,i.e. 25.6 msec.

SUMMARY

A polytonal analysis-synthesis system has been described , which has appliation to robust speech processing. The experi - ments using the new model of speech sig - nal indicate its power in synthesizing natural sounding voised signals.

REFERENCES

/1/ Речевые тесты и их применение. Изд-во МГУ. I986г.,стр.64

/2/ Makhonine V. On the representation of discontinuous speech acoustical events. Rapp.d'activ.de l'inst.de phon.RA12/1 Brux.1978.

/3/ Махонин В.А. Изучение микровариации речевого сигнала.Франко-Советский симпозиум по исследованию речевой информатики. Гренобль I98Iг. стр.303-3I2

/4/Д.Отесер и др. Экспериментальная методика наблюдения микровариации.Советско-Французский симпозиум"Акустический диалог человека с машиной.М.I984г.стр.III-II6

/5/Пирогов А.А.Предисловие к переводу . ТИИЭР,Том 73,№II ,М.,"Мир" I986г.,стр.3

/6/Арнольд В.И. Теория катастроф. Изд-во МГУ ,I983г.,стр.24-25

/7/ Рохтла М.,Раудсепп М. Зависимость качества синтезированной речи от тонкой структуры изменения основного тона.Тезисы АРСО-I4.ЧастьI.Изд-во КПИ,Каунас I986г. стр.57-58.

/8/ H.Nagabuchi,T.Kobayashi,F.Itakura Effect of the comb-filtering noise reduction method in speech analysis-synthesis processing in noisy environments. Rep. Acoust.Soc.Jap.1980. S80-54.

APPENDIX

Table 1
File: "HA" female speaker
Length of minimal pitch period=77
Step of pitch period increasing=1
Threshold for modul.selection=0.25

**************9*****	95	2	0	0	0	0	0	0
9*******************	94	2	0	0	0	0	0	0
92*****6************	95	2	o	0	0	0	0	0
943*723*31***1	89	6	0	1	0	0	0	0
91*74**51*******	88	5	0	1	0	0	0	0
9*338**5********	88	6	0	1	0	0	0	0
938***********	88	5	0	1	0	0	0	0
92*6*351******	90	5	0	1	0	0	0	0
********9*74********	81	12	1	2	0	0	0	0
*******9************	87	7	2	0	0	0	0	0
*********96*********	86	8	1	1	0	0	0	0
**********9*********	82	12	0	2	0	0	0	0
**********9*********	88	7	1	1	0	0	0	0
**********9*********	86	8	1	1	0	0	0	0
**********9246******	33	50	5	5	0	0	1	1
**********94**4*****	61	25	3	6	0	0	0	1
******92************	86	7	3	0	0	0	0	0
****94*52***********	76	15	2	2	0	0	0	0
917*************	73	20	0	3	0	0	0	0
92**************	88	7	0	1	0	0	0	0

Table 2
File: "MA" female speaker
Length of minimal pitch period=77
Step of pitch period increasing=1
Threshold for modul.selection=0.25

92*55*********	94	3	0	0	0	0	0	0
*9******************	89	5	0	0	0	0	0	0
********97*3********	87	6	0	2	0	0	0	0
********95*3********	90	5	0	1	0	0	0	0
********966*********	91	5	0	1	0	0	0	0
*********91*********	85	9	0	2	0	0	0	0
********966*********	78	14	1	3	0	0	0	0
*********91*********	76	15	1	3	0	0	0	0
********92**********	90	6	0	1	0	0	0	0
*********97*********	85	5	4	1	0	0	1	0

```
**********9*********  76 16 1 3 0 0 0 0
**********92********  87  7 2 1 0 0 0 0
***********91*******  88  7 0 1 0 0 0 0
**********9*********  89  5 2 0 0 0 0 0
*********945********  54 32 4 4 0 0 1 1
******9*************  88  6 2 1 0 0 0 0
**98****************  80 12 3 2 0 0 0 0
9*******************  85  9 1 1 0 0 0 0
9*******************  78 15 1 2 0 0 0 0
```

Table 3

File: "MA" male speaker
Length of minimal pitch period=110
Step of pitch period increasing=2
Treshold for modul.selections=0.1

```
****943*************  89 66 0 1 0 0 0 0
******915***********  91  5 0 1 0 0 0 0
*****9**************  82 11 0 1 0 0 0 0
****91**********3***  74 13 4 1 0 1 0 1
****9***************  70 13 5 2 0 0 5 0
**93*3*7************  44 3111 2 1 1 5 1
**988***************  44 2314 4 1 7 2 0
****91*4************  45 1910 9 1 4 5 1
*****978************  51 1323 1 1 2 4 0
*******98***********  68  9 7 2 1 3 5 0
**********96*5******  56 20 9 2 0 1 6 0
***************94327  61 17 4 6 1 4 1 1
```

Table 4

File "МИ " male speaker
Length of minimal pitch period=112
Step of pitch period increasing=2
Treshold for modul.select.=0.1

```
*****94*************  92 4 0 0 0 0 0 0
***********91*******  88 7 0 1 0 0 0 0
****9***33241*******  90 6 0 1 0 0 0 0
*******9552*********  89 6 0 1 0 0 0 0
*****967************  91 5 0 1 0 0 0 0
****9283******5*****  85 8 0 2 0 0 0 0
**9*87**************  83 7 0 2 0 1 1 0
****97**************  8310 0 2 0 0 0 0
****98**************  85 7 0 1 0 1 0 0
*****97*************  8712 0 2 0 0 0 0
*****9452***********  8211 0 2 0 0 1 0
******96**5*********  8211 0 2 0 0 0 0
***********944******  86 8 0 2 0 0 0 0
****************9*4*  87 8 0 1 0 0 0 0
```

Table 5

File " НАДУВАТЬ ",male speaker,heavy noise environment.
Length of minimal pitch period=166
Step of pitch period increasing=3
Treshold for modul.select.=0.25

```
9***5***************  60 13 9 8 2 2 0 0
***97***************  60  7 5 0 4 4 4 0
*****96*************  36 33 511 7 2 1 1
**7*4********5******  52 25 8 3 2 2 2 1
*94*8***************  46 31 5 7 2 2 2 1
9**3****************  74  9 3 5 3 0 0 0
9862*51*1***********  64 15 4 4 2 2 2 2
96*7************2***  55 18 8 6 4 1 2 2
92****83************  58 21 8 4 2 1 1 0
976****************2  61 19 8 4 1 0 1 2
98*4****************  61 21 4 4 2 1 2 1
96****6*************  55 15 5 5 4 6 2 2
*********91*********  64 14 9 3 1 1 2 1
```

```
98**4********4******  38 27 17 5 2 3 3 1
957*2***************  40 33 11 3 1 2 3 1
92******************  38 27 21 2 0 1 6 0
9******3************  12 53 10 8 5 2 2 2
9*********8*3*******  23 28 18 126 5 3 1
```

Table 6

File "ПЛЯСОВОЙ ",male speaker
Length of minimal pitch period=170
Step of pitch period increasing=3
Threshold for modul.select.=0.15

```
***9****************  90  1   1 1 1 0 0 1
98******************  94  1   0 0 0 0 0 0
9*******************  60 17   510 1 2 1 1
95*******3**********  57 25   7 2 1 2 1 0
9438*2**************  72 16   3 2 1 0 0 0
9***5****2**********  61 22   2 7 0 0 0 1
9*******************  54 30   2 6 0 1 0 1
984****22***********  69 17   3 4 1 1 1 0
********************   0  0   0 0 0 0 0 0
********************   0  0   0 0 0 0 0 0
********************   0  0   0 0 0 0 0 0
9*******************  42 12 1323 0 1 1 3
9*****************7*  24 40  19 5 2 2 2 1
9*******************  53 33   5 3 0 0 1 0
9*******************  22 44  18 7 0 0 3 1
9******7************  33 43   9 6 501 0 1
96*****3****2****2**  51 31   6 5 0 1 1 1
*954*2**********3**4  56 26   8 3 0 1 1 0
9****1***********7**  15 41  30 2 0 2 4 0
*9******************  6  51   3 2 0 3 5 0
***9******54********  14 63   9 3 0 5 1 1
*****9**************  21 14  48 2 0 3 7 0
*********9**********  55 23   9 4 1 0 2 0
**************9467**  51 31   7 5 0 0 1 1
***************9****  46 39   3 7 0 0 0 1
```

EXTRACTION OF SPEECH IN ACOUSTICAL NOISE BY MARKOV FILTERING

Y.N.PROKHOROV

A.V.MININ

Moscow Telecommunication Institute, USSR, 111024

ABSTRACT

This paper presents a general approach to the improvement of speech intelligibility in broad band acoustical noise. By using the methods of Markov filtering the digital processing algorithms of noise-added speech are being synthesized and their experimental study is being carried out.

INTRODUCTION

The telephone communication systems and the systems of automatic man-machine communication by voice often operate in a severe broad band acoustical noise situations. The organising protective measures and the compensation techniques do not always provide the effective noise suppression. In such cases the signal-noise ratio (SNR) of the microphone output may be 0-3 db, and the intelligibility S may be 40-50% /1,2/. The special digital processing for noise reduction is applied but it doesn't allow to increase intelligibility sufficiently so far /1,2/. The aim of this paper is to develop the effective processing methods by using Markov filtering.

FORMULATION OF THE PROBLEM

In interval the duration of which is about 20 - 50 ms the mixture of signal and noise is

$$z_t = x(\vec{\lambda}, t) + n_t\ ,\ t = 0, \pm 1, \pm 2, \dots, \qquad (1)$$

where $\vec{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_m)^T$ – is a vector of parameters describing the articulation apparatus state (ArA); $x(\vec{\lambda}, t), n_t$ – the sample sequences of speech signals (SS) and noise. Because of the low accuracy of articulatory organs the parameters take a continuous set of values.

For the automatic recognition (reception) of corrupted speech the values of parameter $\vec{\lambda}$ should be classified. However, in the process of extraction it is quite enough while using z_t to formulate such signal $u(\vec{\lambda}^*, t)$ on hearing of which the maximum intelligibility is achieved: $S^{(z)}_{max} = S^{(u)}$, where $S^{(z)}, S^{(u)}$ – is the intelligibility of signals z_t and $u(\vec{\lambda}^*, t)$ respectively. Since a vector $\vec{\lambda}^*$ or an unknown function $g(\vec{\lambda}^*)$ is classified in the process of human perception, the value of $\vec{\lambda}^*$ should be chosen in such a way that $\varepsilon^2_\lambda = E[\vec{\lambda}^* - \vec{\lambda}]^T Q_\lambda [\vec{\lambda}^* - \vec{\lambda}] = min$ where E – mathematical expectation operation, Q_λ – a weighted coefficient matrix. The minimum attainable value ε^2_λ is defined by Kramer-Rao's inequality.

So the problem of speech extraction is interpreted as the construction of the best estimation of $\vec{\lambda}$ and the creation of the signal $u(\vec{\lambda}^*, t)$ with $S^{(u)}=max$. A general diagram of the extraction device is given in fig.1, where CD is a controlling human ear perception system device.

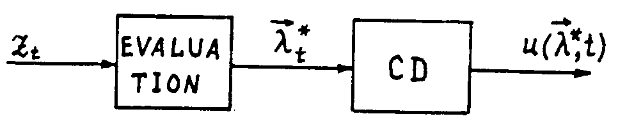


Fig.1. A General Diagram of Speech Extraction

SIGNAL AND NOISE MODELS

For the best evaluation of $\vec{\lambda}$ the adequate models of signal and noise are required. The simplest model of the broad band noise is a Gaussian sequence n_t with $En_t=0, En_t^2=\sigma_n^2, En_{t1}n_{t2}=0, \forall\,(t_1 \neq t_2)$. The more precise model is the process of autoregression

$$v_t=\sum_{i=1}^{L} \alpha_i v_{t-i}+\mu_t, \quad L=2\div 10, \qquad (2)$$

where α_i is evaluated a'priory by the noise realization by means of a least-square technique with limitations. One can use the orthogonal projections as the forms of limitations/3/.
The signal is modelled by a nonlinear autoregressive process

$$y_t=\beta\varphi(y_{t-1})+b_y\eta_{t-1}, \qquad (3)$$

$$x_t=\sum_{i=1}^{m} a_i x_{t-i}+cy_{t-i}+b\xi_{t-1}, \qquad (4)$$

$$\vec{\lambda}=(\beta, a_1, a_2, \ldots, a_m)^T, \quad m=2\div 10.$$

The function φ is found on the synthesis stage from the condition:

$$\min E\left|\varepsilon_x(t)\varepsilon_x(t-\tau)\right|,$$
$$\varepsilon_x(t)=x_t-\beta\varphi(x_{t-1}), \quad \tau=const.$$

The result is shown in fig.2.

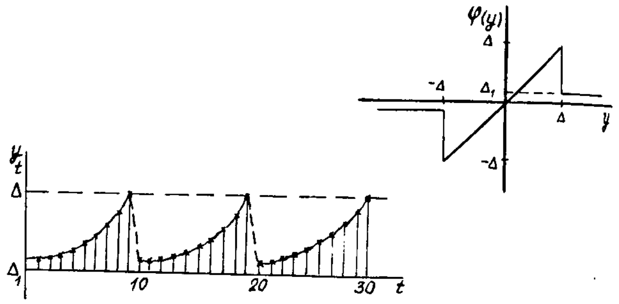


Fig.2. The funotion of Excitation

To reduce the number of the parameters evaluated model (3), (4) may be written in the following way:
$x_t=\sum_{j=1}^{N}\lambda_j\psi_j(x_{t-1}, x_{t-2}, \ldots, x_{t-m})+b\xi_{t-1}$, where $N=2-6$ and the orthogonal functions ψ_j are found experimentally. To do this on the speech signal of a concrete speaker the set of the parameters values is defined in models (3),(4), then a set of autoregression functions is given
$\psi(x_{t-1}, x_{t-2}, \ldots, x_{t-m})$,
and the Karunen-Loev basis is built for it.

THE EVALUATION OF THE POSSIBLE INTELLIGIBILITY

The intelligibility $S^{(z)}_{max}$ may be evaluated in the presence of noise with an average flat spectrum. Consider

$$z_t^{(1)}=x(\vec{\lambda},t)+n_t, \quad z_t^{(2)}=f[x(\vec{\lambda},t)]+n_t.$$

Chose the function f so that $S^{(z_2)}\approx S^{(z_1)}_{max}$ For example f may be the central clipping function. According to the articulatory tests $S^{(z_1)}$, $S^{(z_2)}$ may be found for different SNR_1, SNR_2 and the threshold of the clipping x_{thr} Putting down the Kramer-Rao's inequalities the formulas for $\varepsilon^2_{\lambda 1}(SNR_1)$ and $\varepsilon^2_{\lambda 2}(SNR_2)$ can be obtained. In the situation where $\varepsilon^2_{\lambda 1}(SNR_1)= =\varepsilon^2_{\lambda 2}(SNR)$ we can find a family of curves with the equal reception accuracy:
$SNR_1=\varphi_\varepsilon(SNR_2)$ with parameter x_{thr}. Now the estimation $\hat{S}^{(z_1)}_{max}$ can be obtained in the following way: by using SNR_1, x_{thr} and function φ_ε we can find SNR_2, where $\varepsilon^2_{\lambda 1}(SNR_1)=\varepsilon^2_{\lambda 2}(SNR_2)$. Then $\hat{S}^{(z_1)}_{max}= =S^{(z_2)}(SNR_2)\leq S^{(z_1)}_{max}$ and $\Delta S^{(z_1)}=\hat{S}^{(z_1)}_{max}- -S^{(z_1)}$ is a possible benefit in intelligibility in digital processing of the corrupted speech $z_t^{(1)}$.

THE FILTERING ALGCRITHMS

For the simplification we can take $g(\vec{\lambda})$ as a mutually unique continuous (unknown) function. It can be shown that in this case $u(\vec{\lambda}^*,t)=x(\vec{\lambda}^*,t)$ and instead of CS we may use a speech synthesizor operating according to (3), (4). Thus, the extraction of speech is performed according to the algorithm No.1: "estimation of $\vec{\lambda}$ - synthesis $x(\vec{\lambda}^*,t)$ " (analysis-synthesis). If $bE\xi_t^2 \ll Ex_t^2$, then the algorithm No.1 is very close in effectiveness to the mutual evaluation of $x, \vec{\lambda}, y$ or to the "adaptive filtering" - the algorithm No.2. If there are pauses in conversation and the consonants in speech, then the algorithm No.3 "a mutual evaluation of parameters, filtering of speech and classification of tone-consonant-pause" is quite optimal.
The above mentioned algorithms are synthesized by the maximum of a'posteriory probability criterion in /4, 6/ by using Markov filtering technique /5/.

EXPERIMENTAL RESULTS

Testing of algorithms No.1-3 are performed on the speech signal with the sampling frequence 15 kHz and with the number of quantizing levels 2^{12}. In fig.3 the power spectral densities of the initial (G_x), the processed (G_x^*) signals and the noise-added speech (G_z) are shown for the word "geücmbue" (algorithms No.3, 1). In fig.4 the curves of likelihood function Λ and the current signal power Ex_t^2 recieved on the articulatory tables of syllables without any pauses are shown. The probability of a classification error of tone-consonant-pause is about $3\cdot 10^{-2}$ with a zero threshold. In Table No. 1 the results of tests are shown, where ΔSNR is a benefit

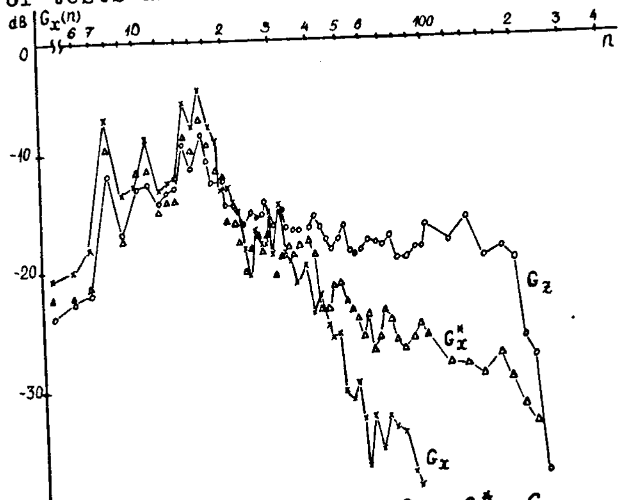


Fig. 3. Signal spectrum G_x, G_x^*, G_z

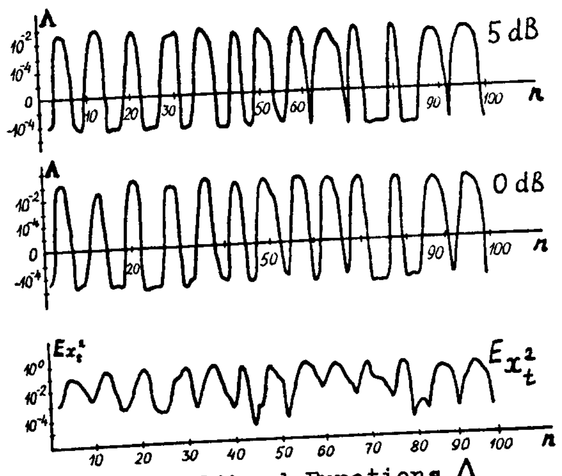


Fig.4. Likelihood Functions Λ

Table 1

No.	3-1	2	3
ΔSNR dB	7	6-7	10-12

in SNR , the number of algorithm 3-1 is a sequential application of the algorithms N3 and N1. These results are achieved for the noise with close to an average flat spectrum and model (3), (4). In pauses the mixture of z_t is multiplied to a coefficient $q < 1$. The coefficient of noise power in filters is chosn experimentally.

In Table No.2 the signal-error prediction ratio (SER) for models (4), (5) which is achieved on the initial speech signal is given. There the results of algorithm No. 2 with (5) in noisy environment because of the engines operation.

In Table П is percentage of the real favours given by the listeners to the processed signal. The number of listeners is 20-25.

Table 2

Model	SER, dB, for N			П	ΔSNR
	2	4	6	%	dB
(4)	21,3	24,7	26,5	-	2-3
(5)	26	27	27,3	85-88	3-5

CONCLUSION

The method of intelligibility improvement in noisy environment is worked out. The theoretical benefit of the digital processing for noise with long-term average flat spectrum is evaluated. By using the Markov filtering techniques the algorithms of mutual speech filtering, the parameters evaluation and the classification of tone-consonant-pause are developed. The algorithms provide the improvement of the corrupted speech intelligibility in broad band noise and can be technically done on the mikroprocessor devices Am 2900.

BIBLIOGRAPHY

1. H.Suzuki, J.Igurashi, Y.Ishii."Extraction of Speech in Noise by Digital Filtering", J.A.S. of Japan, v.33, No.8, 1977, pp.405-411.

2. D.Graupe, J.Grosspietsh, S.Basseas. "Self Adaptive filtering of Environmental Noises from Speech", Proc. AIAA/IEEE 6-th Dig. Avionics Syst. Conf., Bultimor, MD, 1984, N.Y., 1984, pp.263-269.

3. А.В.Минин, Ю.Н.Прохоров. "Оценка параметров речевых сигналов методом наименьших квадратов с ограничениями", Электросвязь, №3, I986, с.26-29.

4.Ю.Н.Прохоров. "Статистические модели и рекуррентное предсказание речевых сигналов", Радио и связь, I984. - СТС, вып.20.

5. A.P.Sage, J.L.Melse. Estimation Theory with Application to Communication and Control, N.Y., McGraw-Hill, 1972.

6.М.В.Назаров, Ю.Н.Прохоров. "Методы цифровой обработки и передачи речевых сигналов", Радио и связь, I985.

QUALITY CONTROL OF SPEECH BY MODIFYING FORMANT FREQUENCIES AND BANDWIDTHS

Hisao Kuwabara

ATR Interpreting Telephony
Research Laboratories
Twin 21 Bldg. MID Tower
2-1-61 Shiromi Higashi-ku
Osaka 540 Japan

Tohru Takagi

NHK Science and Technical
Research Laboratories
1-10-11 Kinuta Setagaya-ku
Tokyo 157 Japan

ABSTRACT

An analysis-synthesis system which is capable of independent manipulation of acoustic parameters has been developed to investigate the contribution of these individual parameters to the speech quality. Formant frequencies and their bandwidths were used as the acoustic parameters to characterize the vocal tract configuration, and pitch frequency as the voice source. This paper describes a way how to control the voice quality of natural speech by manipulating the formant frequencies. Formant trajectories extracted from a natural speech were modified to alter their up-and-down oscillation to some extent, and the resultant speech wave was synthesized by the above mentioned method to present to listeners for the judgment of voice quality. It was found that the speech intelligibility was improved to some extent when the movement of time-varying formant pattern was slightly emphasized, but too much emphasis would cause degradation of the voice quality.

1. INTRODUCTION

This paper deals with a way of controlling the voice quality of natural speech. An analysis-synthesis method has been developed which is capable of independent manipulation of such acoustic parameters as formant frequencies, their bandwidths and pitch frequency [1]. Using this system, voice quality of natural speech has been controlled by changing formant trajectories that are supposed to have a close relation to such voice qualities as intelligibility, clearness and so on.

According to our previous study [2], acoustic characteristics of professional announcers speech, which is considered to be the most intelligible or the clearest, lies in the dynamics of pitch and formant frequencies. The dynamic range of these features for the announcers speech is significantly large compared to that for the non-professional speakers. Correlation analysis between psychological and acoustic distances reveals that the formant trajectory has the largest correlation with the voice quality of the announcer's speech sounds, followed by pitch frequency. This result suggests that the quality of speech sound of non-professional speakers may possibly be improved by altering the dynamics of formant trajectory patterns.

Based on the experimental evidence mentioned above, an experiment has been performed to change and improve the quality of natural speech making use of the analysis-synthesis system. Formant trajectories are extracted first from voiced portions by LPC method and the dynamics of these trajectories are altered depending on the formant pattern itself. The method for altering the formant pattern is the same as that we have proposed earlier for the normalization of vowels in connected speech [3]. This method is applied to the formant trajectories extracted from natural speech, and the quality-controlled speech sounds are synthesized using the analysis-synthesis system to present to listeners for perceptual judgment.

2. ANALYSIS-SYNTHESIS SYSTEM

Fig. 1 illustrates the block diagram of the analysis-synthesis system. Low-pass filtered input speech was digitized in 12 bits at a rate of 15 kHz. A short time LPC analysis based on the autocorrelation method was performed to obtain LPC coefficients and the residual signals. Formant frequencies and their bandwidths were estimated by solving a polynomial equation. A modification of the spectral envelope is equivalent to a manipulation of the coefficients that would result in a frequency response of the filter equal to the modified envelope. These acoustic parameters

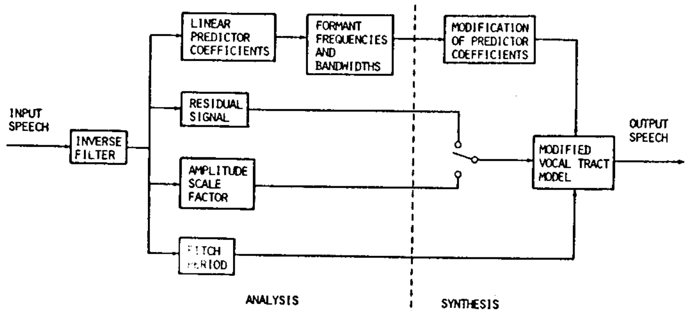

INPUT SPEECH
INVERSE FILTER
LINEAR PREDICTOR COEFFICIENTS
FORMANT FREQUENCIES AND BANDWIDTHS
MODIFICATION OF PREDICTOR COEFFICIENTS
RESIDUAL SIGNAL
AMPLITUDE SCALE FACTOR
PITCH PERIOD
MODIFIED VOCAL TRACT MODEL
OUTPUT SPEECH
ANALYSIS
SYNTHESIS

Fig. 1 Block diagram of the analysis-synthesis system to modify formant frequencies.

(pitch periods, LPC coefficients, formant frequencies, bandwidths, residual signals) were stored for later synthesis.

Let $z_i = r_i \exp(j\omega_i)(i = 1,2.....,p)$ stand for the roots corresponding to the formants to be changed. Formant frequencies and/or their bandwidths are modified by changing related angular frequencies ω_i and/or the factors r_i. LPC coefficients are modified so that the modified poles $\tilde{z}_i$ become the roots of a new polynomial,

$$z^p + \tilde{a}_1 z^{p-1} + \dots + \tilde{a}_{p-1} z + \tilde{a}_p = 0. \qquad (1)$$

Calculation of $\tilde{a}_i(i=1, 2, \dots, p)$is performed simply by comparing terms of the same order on both sides of the following equation,

$$(z-\tilde{z}_1)(z-\tilde{z}_2)\dots(z-\tilde{z}_p)= z^p + \tilde{a}_1 z^{p-1} + \dots + \tilde{a}_{p-1} z + \tilde{a}_p . \qquad (2)$$

The modified vocal tract model $\tilde{V}(z)$ is then given by,

$$\tilde{V}(z) = 1 / (1 + \sum_{i=1}^{p} a_i z^{-i}) \qquad (3)$$

where $\{\tilde{a}_i\}$ are the solutions of equation (2). The modified vocal tract model $\tilde{V}(z)$ has the desired frequency characteristics. If the spectral manipulation is too large, some discontinuities are found to occur at the boundary of each frame, which eventually cause a typical buzzing. To cope with this, a simple time domain manipulation has been performed. In this experiment, half the analysis window is set as the period of frame shift. Output speech wave from the modified vocal tract model $\tilde{V}(z)$ for each frame is multiplied by a triangular time window. The amplitude of this triangular window is composed so that the sum of the gain at any instant within the overlapped portion between two successive frames becomes 1. The resultant speech is obtained by adding successively speech waves between adjacent two frames. This process is illustrated in Fig. 2.

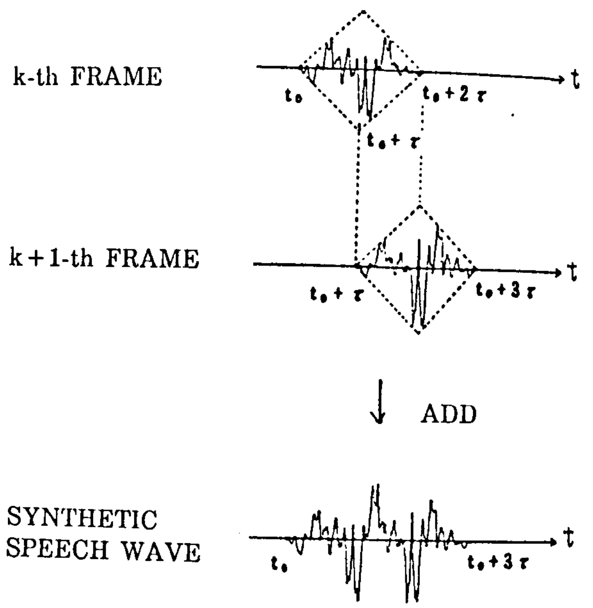


Fig. 2 Method of obtaining high quality synthetic speech

3. ALTERATION OF FORMANT TRAJECTORY

3.1 Formant Trajectory Extraction

Low-pass filtered input speech was digitized at a rate of 15kHz, and the linear predictive analysis was made to find formant frequencies. Autocorrelation method for the inverse filter was adopted with order 14, the analysis window 20 ms, and the frame period 10 ms. Silent intervals and voiced/voiceless distinctions were made based on the speech power and the first order PARCOR coefficient, respectively. Formant frequencies for each frame were extracted from a set of 7 poles using the method proposed by Kasuya et al [4] and a smoothing was made by averaging formant data over three consecutive frames.

3.2 Formant Trajectory Change

Modification of formant trajectory was conducted in such a way that the preceding and succeeding acoustic features contributed to the present value with the same weight if the time differences from the present were equal, and that the amount of contribution was proportional to the difference from the present acoustic feature [3]. This process is illustrated in Fig. 3. Suppose the curve x(t) be an actual time-varying pattern of a formant frequency, the new value y(t) is defined as the sum of the original value x(t) and the additional term of contribution by contextual information. The contribution is assumed to be a weighted sum of differences between values at the present time t and at different time $t \pm \tau$. Thus, y(t) is given by

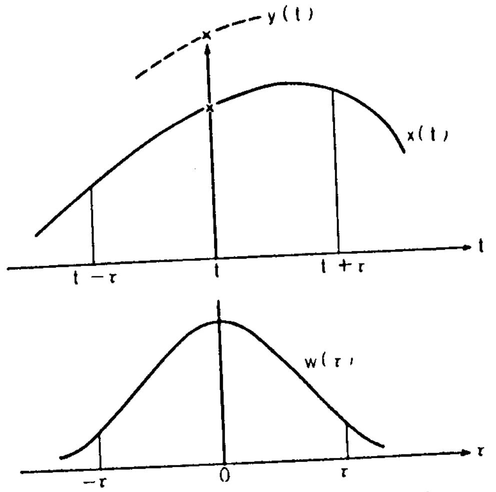


Fig. 3 Illustration of how to change the formant trajectory.

$$y(t) = x(t) + \int_{-T}^{T} w(\tau)(x(t)-x(t+\tau))d\tau \qquad (4)$$

where $w(\tau)$ is the weighting function which is given as

$$w(\tau) = \alpha \cdot \exp(-\tau^2/2\sigma^2). \qquad (5)$$

In this study, T=150ms and σ =52ms were experimentally decided. Given α >0, the dynamics of the original formant trajectory is emphasized, while for α <0, it becomes de-emphasized.

Equation (4) is applied to each of the three formant trajectories without vowel/consonant (but except voiceless consonant) distinction.

4. PERCEPTION OF QUALITY-CONTROLLED SPEECH

As described in the previous section, dynamic movement of formant frequency is one of the most important acoustic factors that characterize clear and intelligible voice. Change of voice quality to improve the clearness of intelligibility should, therefore, be done by modifying the dynamics of formant frequency. In this section, we describe a perceptual experiment on voice quality for formant-modified speech.

4.1 Synthesis Method

Based on the analysis-synthesis system described above, several formant-modified speech signals were obtained to present to listeners for quality judgment. Following is the process of speech synthesis.

(1) Speech waves are digitized with 15kHz sampling rate and 12bits accuracy. Analysis is made based on the system shown in Fig. 1 with a 20ms analysis frame multiplied by the Hamming window and 10ms frame period. The orders for analysis are 14 for male and 10 for female voices, and the predictor coefficients and the residual signals for each frame are stored.

(2) Formant frequencies of the first three are calculated from the predictor coefficients for each frame, and their trajectories over the entire word are estimated using a tracking algorithm [4].

(3) Equation (4) is applied independently to each formant trajectory and new frequencies down to each frame are calculated, and the resultant new coefficients are obtained by the method described in section 2. However, formants higher than the fourth and voiceless consonants remain unchanged.

(4) A vocal tract model is formed using the new predictor coefficients, given in equation (3), and the speech signals are obtained by inputting the residual signals to the model.

A nonsense word /a o i u e/ which consists of a concatenation of five Japanese vowels was used as the speech material. As mentioned before, some discontinuity would occur at the boundary between two successive frames if we simply connect speech signals from the two frames without overlap, which may cause degradation of speech quality. Fig. 2 shows a method how to avoid this sort of degradation.

In equation (5), constant α represents a scale factor which controls the amount of formant modification when it is applied to a formant trajectory as in equation (4). The dynamic pattern of formant movement is emphasized for α being positive, unchanged for α=0, and de-emphasized for negative value. Fig. 4 represents an example of formant trajectories of a speech sample used in the perceptual experiment before and after applying Eq.(4).

4.2 Result of Perceptual Experiment

The above mentioned nonsense word was used as the speech material and two speakers, male and female each, read the word with a normal speed. Seven different values, ranging from -15.3 to 15.3 including zero were selected as the factor α to

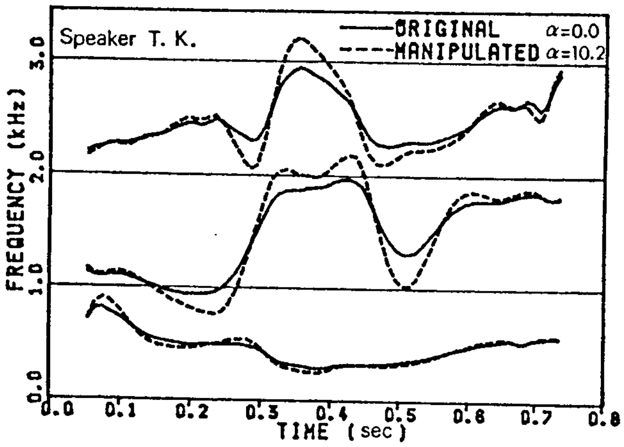


Fig. 4 An example of formant trajectory modification: original(solid lines) and modified(dashed lines).

get synthetic speech samples to be examined. Five female listeners, who never heard the speakers voices before, participated in the experiment. For each speaker, seven speech samples were paired and the listeners were asked to judge which one of a dyad sounded more intelligible or clear by comparison.

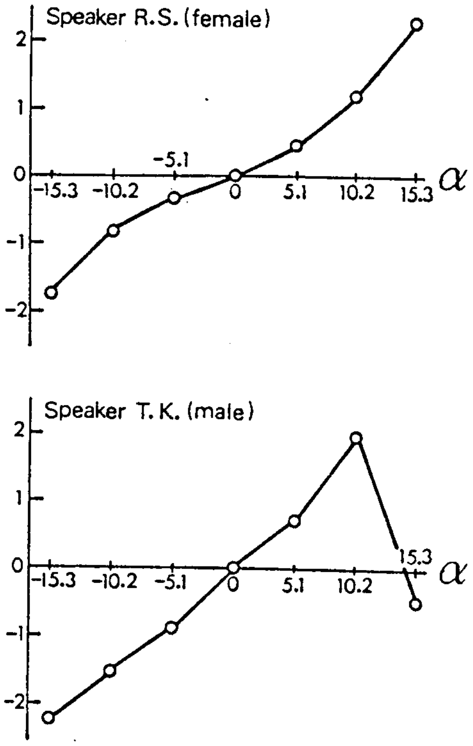


Fig. 5 Results of perceptual experiment on voice quality for male and female speakers.

Fig. 5 shows the result for speech samples of each speaker. The abscissa represents the factor α and the ordinate is a psychological distance. This distance is similar to JND (Just Noticeable Difference) distance, and 1 means that the perceptual difference between the two stimuli is greater than 50 percent chance level.

Being α =0 the reference of comparison, the results show that, in general, the voice quality becomes intelligible as the factor α increases. For male speaker's voice, however, it goes maximum when α =10.2 and goes down rapidly for larger α. This speaker dependency is caused by the degradation of quality by emphasizing the frequency movement too much and partially losing the phonetic quality.

In general, voice quality was found to be improved for the factor somewhere between 5 to 10. The factor greater than 10, however, sometimes gives the speech an improved quality but sometimes degraded quality depending on speakers.

CONCLUSIONS

Time-varying dynamic pattern of formant frequencies which is the main factor to contribute to the clearness or intelligibility has been modified using an analysis-synthesis system and perceptual experiment has been performed on the voice quality. It was found that the voice quality was improved to some extent when the dynamics was properly emphasized.

References

[1] H. Kuwabara, "A pitch synchronous analysis/ synthesis system to independently modify formant frequencies and bandwidths for voiced speech," SPEECH COMMUNICATION, Vol.3 (1984) pp.211-220

[2] H. Kuwabara, K. Ohgushi, "Acoustic characteristics of professional male announcers' speech sounds," ACUSTICA, Vol.55 (1984) PP.233-240

[3] H. Kuwabara, "An approach to normalization of coarticulation effects for vowels in connected speech," J. Acoust. Soc. Amer., Vol.77 (1985) pp.686-694

[4] H. Kasuya et al, "An algorithm to choose formant frequencies obtained by linear prediction analysis method," Trans. IECE Japan Vol.J66-A (1983) pp.1144-1145

SPEECH ENHANCEMENT*

by

Jae S. Lim

Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
Cambridge, Massachusetts, USA

ABSTRACT

There has been considerable interest in recent years on the problem of enhancing degraded speech. This interest is motivated by several factors including a broad set of important applications and the apparent lack of robustness in recent speech compression and recognition systems. One objective of this paper is to provide an overview of various techniques that have been proposed for enhancement of speech. Another objective is to suggest some directions for future research in the speech enhancement problem.

I. Introduction

The objective of speech enhancement may be to improve the overall quality, to increase the intelligibility, to reduce the listener fatigue, etc., and there exists a wide variety of contexts in which speech enhancement is desirable. For example, environments such as offices, streets, and motor vehicles in which the interfering background noise has been introduced are common, and the interfering noise generally degrades the intelligibility and quality of speech. Other examples in which the need for speech enhancement arises include correcting for reverberation, correcting for the distortion of the speech of underwater divers breathing a helium-oxygen mixture, correcting for the distortion of speech due to pathological difficulties of the speaker, and improvement of normal undegraded speech for people with impaired hearing.

Engineers and researchers in various disciplines have shown considerable recent interest in speech enhancement. Among these are engineers working on speech communication problems such as developing robust vocoders and audiologists helping people with impaired hearing. This recent interest is due in part to rapid advances in hardware technology that allow sophisticated signal processing algorithms to be implemented in real time. This interest is likely to continue as speech enhancement systems find more practical applications. One main objective of this paper is to provide a review and survey of past and current research on speech enhancement.

The approach to speech enhancement taken varies considerably depending on the context in which the problem arises. For example, the type of processing indicated for enhancing speech degraded by additive noise is different from that suggested for enhancing speech degraded by echoes. This paper addresses speech enhancement in three different broad contexts which were selected for their common occurrence in practice and for the existence of some major research results. Section II considers the problem of enhancing speech which has been degraded by additive noise. Even though this problem has received considerable attention in recent literature and is rich with sophisticated signal processing, major unsolved problems offer considerable room for further research. Section III considers the problem of enhancing speech degraded by reverberation or echoes. Systems that are successful in reducing room reverberation or telephone network echoes have been developed and discussed in this section. Section IV considers the problem of slowing down or speeding up the apparent rate of speech. Potential applications exist in which even undegraded original speech is enhanced by such processing. For example, people with impaired hearing or who are learning a foreign language may prefer the slowed-down speech to the original undegraded speech. Section V concludes this paper with an attempt to identify some of the potential future research topics on the speech enhancement problem.

II. Enhancement of Speech Degraded by Additive Noise

The problem of enhancing speech degraded by additive noise received considerable attention in the literature in the past ten years and a variety of systems have been proposed. Such an interest in this problem was motivated partly by the desire to improve the robustness of vocoders such as linear prediction vocoders which degrade quickly in performance as noise is added and partly by the impression that reduction of additive noise in speech appeared to be a relatively simple problem. In this section, we discuss some of the representative speech enhancement systems which attempt to reduce the additive noise. We first discuss the case when the degradation is due to additive random noise and then the case when the degradation is speechlike noise.

Let $s(n)$,$d(n)$, and $y(n)$ denote speech, additive noise, and degraded speech, respectively, so that

$$y(n) = s(n) + d(n) \qquad (1)$$

where $d(n)$ is uncorrelated with $s(n)$. One approach to restore $s(n)$ from $y(n)$ is to exploit the long-term characteristics of $s(n)$ and $d(n)$. Specifically, the average speech spectrum decays with frequency at approximately 6 dB/octave and assuming that the power spectrum of the background noise is known or can be estimated such as from the silence intervals of the degraded speech, a time-invariant Wiener filter may be used to estimate $s(n)$ from $y(n)$. The Wiener filter is the best linear filter in the sense that no other linear filter leads to a smaller mean square error between $s(n)$ and $\hat{s}(n)$, the estimate of $s(n)$, under the assumption that $s(n)$ and $d(n)$ are samples of stationary random

[*] This paper was previously published as a pre-conference lecture paper for ICASSP 86 held in Tokyo, Japan, in April 1986.

processes. The frequency response, $H(\omega)$, of the non-causal Wiener filter is given by

$$H(\omega)=\frac{P_s(\omega)}{P_s(\omega)+P_d(\omega)}. \qquad (2)$$

In equation (2), $P_s(\omega)$ and $P_d(\omega)$ represent the power spectrum of the signal and the additive random noise uncorrelated with the signal respectively. The Wiener filter can be quite effective in applications in which the spectrum of the signal and the background noise do not overlap significantly or the background noise is narrow-band such as in the case of sinusoidal interferences.

Another approach to speech enhancement is to exploit some perceptual aspects of human speech. One such system was proposed by Drucker [1]. Based on some perceptual tests, Drucker concluded that one primary cause for the intelligibility loss in speech degraded by wide-band random noise is the confusion among the fricative and plosive sounds which is partly due to the loss of pauses immediately before the plosive sounds. By high-pass filtering one of the fricative sounds, the /s/ sound, and inserting short pauses before the plosive sounds, Drucker claims a significant improvement in intelligibility. The system considered by Drucker assumes that the locations of the plosive and fricative sounds are accurately known, which may not be a reasonable assumption for degraded speech.

Another class of speech enhancement systems exploits the notion that it is principally the short-time spectral magnitude rather than phase that is important for speech intelligibility and quality. In this class of systems, the degraded speech is first windowed, the short-time spectral magnitude of speech is estimated from the windowed degraded speech, and then enhanced speech is obtained by inverse-transforming the estimated short-time spectral magnitude combined with the phase of the windowed degraded speech. A number of different methods to estimate the short-time spectral magnitude of speech from the windowed degraded speech have been developed both theoretically and heuristically. In one method referred to as "power spectrum subtraction", the short-time spectral magnitude of speech $|S_w(\omega)|^2$ is estimated by

$$|\hat{S}_w(\omega)|^2 = \begin{cases} |Y_w(\omega)|^2 - E[|D_w(\omega)|^2] & \text{for } |Y_w(\omega)|^2 > E[|D_w(\omega)|^2] \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

In equation (3), $|\hat{S}_w(\omega)|^2$ is an estimate of $|S_w(\omega)|^2$, $|Y_w(\omega)|$ and $|D_w(\omega)|$ are the Fourier transform magnitudes of the windowed noisy speech and the windowed additive noise, respectively, and $E[|D_w(\omega)|^2]$ denotes the average $|D_w(\omega)|^2$. A speech enhancement system based on a generalization of equation (3) is shown in Figure 1. In the figure, if the result after subtraction of $E[|D_w(\omega)|^2]$ is less than zero, it is set to zero. When the constant "a" in the figure equals 2, the system corresponds to the power spectrum subtraction method. The system in Figure 1 was evaluated by [2] using nonsense sentences as test material when the degradation is wide-band random noise for a=2, 1, 1/2, 1/4. The results of the test show that the intelligibility is not improved at the S/N ratios at which the intelligibility scores of unprocessed nonsense sentences range between 20 and 70 percent. However, processed speech with a=1 or 1/2 sound distinctly "less noisy" and of "higher quality" at relatively high S/N ratios. The system in Figure 1 with a=1 was also evaluated [3] when the degradation is due to helicopter noise. The results based on Diagnostic Rhyme Test indicate that at the S/N ratio at which the intelligibility score of unprocessed speech material is about 84

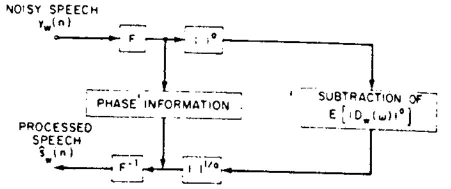


Figure 1. Generalization of Power Spectrum Subtraction Method for Speech Enhancement

percent, the system does not improve intelligibility, but improves quality. Other methods of estimating the short-time spectral magnitude of speech have not been carefully evaluated using a subjective test, but appear to have similar performance to that of the system in Figure 1.

Another approach to speech enhancement is to exploit the observation that waveforms of voiced sounds are periodic. Specifically, the periodicity of a time waveform manifests itself in the frequency domain as harmonics with the fundamental frequency corresponding to the period of the time waveform as shown in Figure 2. In Figure 2(a) is shown a segment of a periodic time waveform, and in Figure 2(b) is shown the associated magnitude spectrum. As is evident in Figure 2(b), the energy of a periodic signal is concentrated in bands of frequencies. Since the interfering signals in general have energy over the entire frequency bands, to the extent that accurate information of the fundamental frequency is available, a comb filter as shown in Figure 2(c) can reduce noise while preserving the signal. An adaptive filter which is based on the comb filtering concept and which partially accounts for the fact that voiced speech is only approximately periodic has been developed by Frazier, et al. [4]. This algorithm with a small improvement was evaluated in [5] using nonsense sentences as test material when the degradation is due to wide-band random noise.

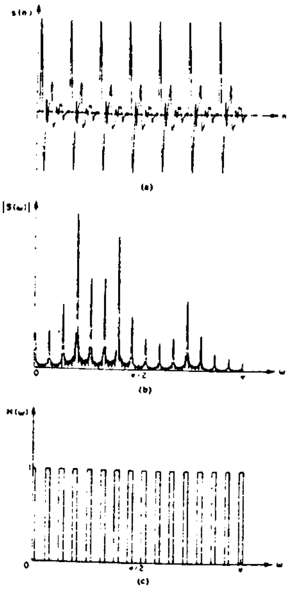


Figure 2. (a) A Periodic Time Waveform

(b) Spectral Magnitude of the Waveform in (a)

(c) Frequency Response of an Ideal Comb Filter

The pitch information used in the processing was obtained from the noise-free speech. The results of the test show that even with accurate pitch information, the adaptive filtering technique tends to decrease the intelligibility at various S/N ratios. Despite the decrease in intelligibility, speech processed by an adaptive filter sounds "less noisy" due to the capability of the system to increase the S/N ratio.

Another approach to speech enhancement attempts to exploit the underlying model for speech production. In this approach, speech is typically modelled by the response of a linear system, representing the vocal tract, driven by an excitation function which is a periodic pulse train for voiced sounds and wide-band random noise for unvoiced sounds, as is illustrated in Figure 3. Since the vocal tract changes its shape as a function of time, the digital filter in Figure 3 that represents the vocal tract is in general time-varying. However, over a short interval of time, the digital filter may be approximated as a linear time-invariant system. In a speech enhancement technique that exploits the underlying model of speech, the parameters of the speech model are first estimated and then speech is generated by a synthesis system based on the same underlying speech model or by designing a filter with the estimated model parameters and then filtering the noisy speech. Several different speech enhancement systems have been developed by using this approach with the vocal tract modelled by an all-pole or pole-zero system and with the speech model parameters estimated by the maximum likelihood method that accounts for the presence of noise. The performance of these systems has not been evaluated by a subjective test. Informal listening, however, indicates that the quality of speech is improved while the improvement in speech intelligibility is not clear.

The speech enhancement systems discussed above are applicable to the case when there is one degraded input. When more than one input is available for processing, further enhancement may be possible. For example, each of the individual inputs may be processed separately using the speech enhancement systems discussed above and then appropriately combined. In addition to processing different inputs separately, signal processing algorithms have been developed in which the correlation of noise in several inputs is exploited and dramatic improvement is possible in some limited applications. One such algorithm is the adaptive noise cancelling algorithm discussed in [6]. Specifically, consider an environment in which the primary input has the signal $s(n)$ and noise $d(n)$ uncorrelated with $s(n)$ and the reference input has noise $r(n)$ uncorrelated with $s(n)$ but correlated in some unknown way with noise $d(n)$. The adaptive noise canceller adaptively filters the reference input $r(n)$ to estimate $d(n)$ and this estimate is subtracted from the primary input to form the signal estimate. The adaptive noise-cancelling concept is illustrated in Figure 4. The adaptive noise-cancelling filter which is typically a tapped-delay line (or finite impulse response) filter adapts the filter coefficients by minimizing the power in $\hat{s}(n)$. It can be shown that minimizing the power of $\hat{s}(n)$ in fact minimizes the mean square error between $s(n)$ and $\hat{s}(n)$ and algorithms [6] have been developed to estimate the filter coefficients. The adaptive noise-cancelling algorithm has been applied to a simulated environment in which a person spoke into a microphone in a room where strong acoustic interference was present. The signal at this microphone formed the primary input. A second microphone was placed in the room away from the speaker and close to the source of the acoustic interference and the signal in the second microphone formed the reference input. The S/N ratio improvement achieved in this experiment using the adaptive noise-cancelling technique is quite dramatic. The noise canceller has been demonstrated [6] to reduce the output power of the interference, which otherwise makes the speech unintelligible, by more than 20 dB, rendering the interference in the primary input barely perceptible. Despite such a dramatic improvement in performance and the system's capability to adapt itself to changing noise statistics and movements of microphones, the adaptive noise-cancelling technique is limited in practice since the reference input typically contains the signal $s(n)$ as well as the noise, in which case the noise canceller will attempt to cancel the signal as well as the degrading noise. Various attempts to improve the performance of adaptive noise-cancelling techniques are currently in progress. Some researchers attempt to develop new algorithms for adaptive noise cancellation. Some researchers attempt to identify environments where existing noise-cancelling techniques may be used with minor modification. The results of these current research efforts are expected to be available in the open literature in the near future.

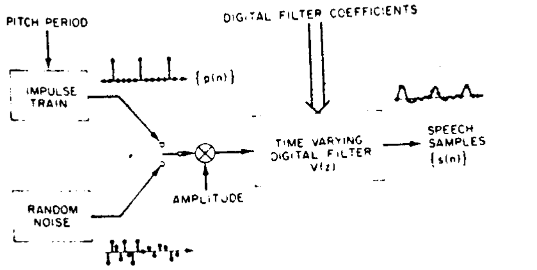


Figure 3. A Speech Production Model

In the above, we have discussed speech enhancement systems applicable to the case when the degradation is due to additive random noise. The problem of enhancing speech degraded by speechlike noise such as in the presence of competing speakers is in general considerably more difficult than the additive noise degradation case for various reasons. The speechlike noise has the long-time spectral characteristics similar to those of the speech and consequently systems such as the Wiener filter which exploit the differences in the long-time spectral characteristics of speech and the background noise are not effective. In addition, the speechlike noise varies rapidly in its characteristics as a function of time and estimating the characteristics of the degrading noise is quite difficult. Since speech enhancement systems which attempt to estimate the short-time spectral magnitude of speech of an underlying speech model generally require a good estimate of the characteristics of the degrading noise, they can not be used effectively to combat the speechlike noise.

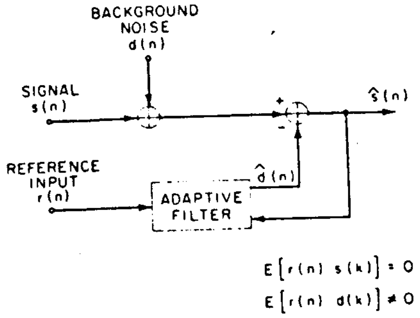


Figure 4. An Adaptive Noise Cancelling Algorithm

One approach which has been developed to combat specifically the interference from a competing speaker attempts to exploit the periodicity of voiced speech and has been developed by Parsons [7]. In this system, voiced speech is windowed and a high-resolution short-time spectrum is obtained. In the short-time spectrum, the periodicity of speech exhibits itself as local spectral peaks some of which are due to the main speaker and some others of which are due to a competing speaker. Parsons developed a technique in which each of the local spectral peaks in the high-resolution short-time spectrum is distinguished between the main speaker and the competing speaker. Then speech is generated based on the spectral content that corresponds to the peaks of the main speaker. Since the essence of Parsons' system is location and selection of speech harmonics of a speaker from the high-resolution spectrum of degraded speech, it can be approximately viewed as a frequency domain implementation of a pitch information extractor and an adaptive filter by Frazier. Even though the system by Parsons has not been evaluated by a subjective test, the adaptive filter by Frazier has been evaluated by Perlmutter [8] using nonsense sentences as test material when the degradation is due to a competing speaker. The results indicate that even with accurate pitch information, the adaptive filtering technique decreases the intelligibility at the S/N ratios at which the intelligibility of unprocessed nonsense sentences ranges between 20 and 70 percent.

The adaptive noise-cancelling system may also be used when the degradation is due to a competing speaker. Assuming that a reference input contains only the speech of the competing speaker, it is expected that the competing speech can be significantly reduced.

III. Enhancement of Speech Degraded by Echoes

In this section, we discuss some of the representative systems which attempt to enhance speech degraded by echoes. One approach which has been applied to remove echoes in signals is based on the homomorphic system theory by Oppenheim, Schafer, and Stockham [9]. In this approach, a signal combined by a convolution of two components is first transformed so that the two components become additive and then a linear filter is applied to separate one component from the other. Specifically, let $s(n)$ and $h(n)$ denote a signal and a train of pulses. Then $y(n)$, the signal degraded by echoes, can be represented by

$$y(n)=s(n)*h(n) \tag{4}$$

where "*" represents the convolution operation. For example, when $y(n)$ is a sum of $s(n)$ and its delay, then $y(n)$ can be expressed as

$$y(n)=s(n)+\alpha\cdot s(n-n_0)=s(n)*(\delta(n)+\alpha\cdot\delta(n-n_0)) \tag{5}$$

where $\delta(n)$ is a unit sample sequence. By z-transforming both sides of equation (4), applying the logarithmic operation, and then inverse z-transforming, equation (4) can be expressed as

$$\hat{y}(n)=\hat{s}(n)+\hat{h}(n) \tag{6}$$

By linearly filtering $\hat{y}(n)$, this approach attempts to recover $\hat{s}(n)$, from which $s(n)$ is recovered. For a typical signal $s(n)$ such as speech and for a rather restricted class of $h(n)$ such as when $h(n)$ is a minimum phase signal with a large equal spacing between the two consecutive pulses, a good estimate of $s(n)$ has been demonstrated. For example, for speech artificially degraded by equation (5) with $\alpha=0.5$ and n_0 corresponding to 50 msec., a significant echo suppression has been demonstrated. Even though this approach is theoretically interesting, its applicability is limited to a rather restricted class of problems.

Another approach to suppress echoes in speech has been developed specifically for the purpose of suppressing echoes in long distance telephone communications. A reasonable model of speech degradation due to echoes in long distance telephone communications is given [10] by

$$y(n)=s_d(n)*h(n)+s(n) \tag{7}$$

where $s(n)$ is the speech signal to be recovered, $s_d(n)$ represents the speech of another speaker, $h(n)$ represents the impulse response of the echo path, which may be varying in time, and the echo canceller has access to $s_d(n)$ and $y(n)$. In this approach, the echo path impulse response is approximated by a tapped delay line filter $h'(n)$ and the filter coefficients of $h'(n)$ are constantly updated by attempting to reduce the error between $y(n)$ and $s_d(n)*h'(n)$ during the intervals $s(n)$ appears to be absent. The enhanced speech is, then, obtained by subtracting $s_d(n)*h'(n)$ from $y(n)$. The success of this algorithm for the specific purpose it was developed is evidenced by the fact that a single chip VLSI echo canceller that implements the algorithm has been fabricated [10]. The chip measures 313 by 356 mils and contains 35,000 devices.

When speech is degraded by room reverberation, the degraded speech $y(n)$ can again be expressed by equation (4) with $h(n)$ representing the room impulse response. Unfortunately, homomorphic processing discussed above cannot be applied to this problem, since the room impulse response $h(n)$ does not belong to the restricted class for which homomorphic processing is applicable. Among various different approaches considered to solve this problem, one approach which appears to be quite successful exlpoits the notion that the room impulse response $h(n)$ has different characteristics when the signal is picked up at different locations and requires signals from two microphones. More specifically, let the signal at the second microphone be denoted by

$$z(n)=s(n)*g(n) \tag{8}$$

By representing $h(n)$ and $g(n)$ in terms of earlier arrivals $h_e(n)$ and $g_e(n)$ and later arrivals $h_l(n)$ and $g_l(n)$, $y(n)$ and $z(n)$ can be expressed as

$$y(n)=s(n)*h_e(n)+s(n)*h_l(n) \tag{9}$$

$$z(n)=s(n)*g_e(n)+s(n)*g_l(n) \tag{10}$$

By exploiting the empirical observation that there is a strong correlation between $s(n)*h_e(n)$ and $s(n)*g_e(n)$, but little correlation between $s(n)*h_l(n)$ and $s(n)*g_l(n)$, an algorithm that reduces $s(n)*h_l(n)$ and $s(n)*g_l(n)$, but combines $s(n)*h_e(n)$ and $s(n)*g_e(n)$ in an appropriate manner has been developed [11]. The performance of this algorithm has been evaluated by Bloom [12] for people with normal hearing and hearing impairment in a very reverberant classroom environment. Preliminary results of the test indicate that intelligibility is not improved. Empirical listening to the processed speech clearly demonstrates, however, that the echoes due to classroom reverberation have been significantly suppressed.

IV. Time Scale Modification of Speech

In the previous two sections, we discussed algorithms that account for a specific type of speech degradation, namely additive noise and reverberation. In the present section we discuss a specific class of signal processing algorithms that can potentially enhance speech in various contexts by changing the time scale of speech, slowing down or speeding up its apparent rate. Examples in which speech is enhanced by changing its time scale include slowing it down to learn a foreign language or to communicate with a person who has a hearing impairment, and speeding it up to read written material to the blind. Even though the original speech is not degraded in these examples, speech is enhanced, in the sense that the listener would prefer the processed speech, by changing its time scale.

Probably the simplest method of changing the time scale of speech is to record speech at one speed and then play it back at a different speed. Since this has the effect of scaling all the frequencies, the method is useful in practice only for a very small change in the time scale of speech. When this method is used to produce only a 10% time-scale change, the pitch change is easily perceived and speaker identification can be impaired. A time-scale change greater than 35% results in rapid deterioration of speech intelligibility.

Another simple approach is to cut speech tapes into segments, repeat or discard the segments periodically, and then rejoin the segments later. It has been reported that such methods preserve [13] both intelligibility of speech at a time-scale change of 100% or more. Retention of such high speech intelligibility is due primarily to the fact that speech has a high degree of redundancy, and the retained speech segments preserve the short-time speech spectrum to a certain extent. An ingeneous electromechanical method to periodically discard speech segments has been developed by Fairbanks, et al [13], and has been used in practice for some time. As a result, the method of periodically discarding speech segments for time compression is often referred to as the "Fairbanks method". Using the current digital technology, the Fairbanks method can be implemented in a very straightforward manner.

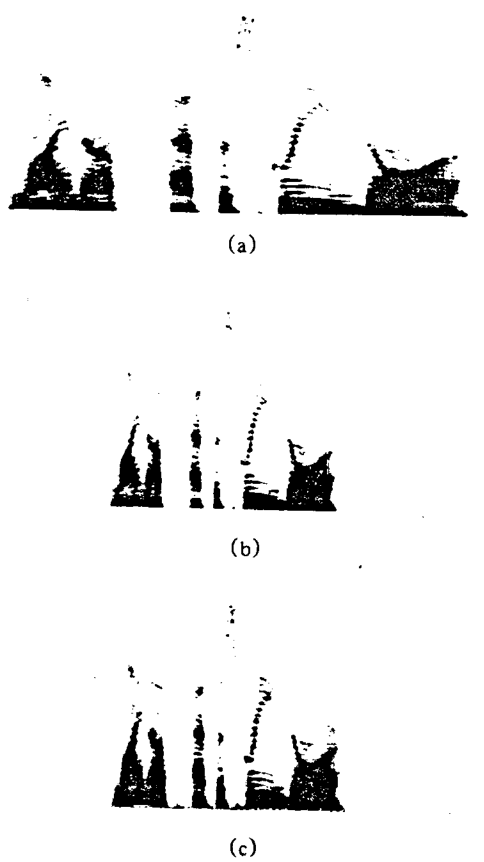
Even though the Fairbanks method preserves the intelligibility of speech at high rates of time-scale modification, the quality of speech suffers noticeably. Since speech segments are periodically discarded without any consideration of the speech waveform, the resulting speech often has discontinuities at the segment boundaries and speech is spectrally distorted. To reduce boundary discontinuity and spectral distortion problems, Scott and Gerber [14] developed a method in which speech segments are discarded or repeated pitch-synchronously. In this method, pitch information is first obtained from the speech waveform and an integer number of pitch periods are repeated or discarded. The pitch-synchronous method noticeably improves both the quality and the intelligibility of the processed speech over the Fairbanks method. Various commercial systems currently available are variations of the pitch-synchronous method.

A different approach to the time-scale modification problem is to first filter speech by a bank of bandpass filters, modify the time scale of the output of each filter, and then combine the resulting outputs. This approach has several important advantages over those discussed above. For example, any distortions caused by processing in one band of frequencies has little effect on other frequency bands, and thus the short-time spectral components important for the intelligibility or quality of speech can be better controlled. In addition, any periodic signal can be decomposed into a series of complex exponentials, and the output of each channel can be made to contain at most one exponential by properly choosing the bandwidths and center frequencies of the bank of filters. Since the time-scale modification is simpler for an exponential with one frequency than for a general speech waveform, this can be exploited in the approach. Malah [15] presents a method in which the speech is decomposed into complex exponentials, and then only the frequency of each exponential is modified by the same ratio in each channel without affecting the amplitude and time duration of the exponential. This is accomplished by a simple time-domain algorithm. When the modified exponentials are combined, the resulting speech has the same duration as the original speech but all the frequency components have been linearly scaled. The linear frequency scaling can be corrected by changing the playback speed, which results in compression or expansion of the speech time scale. This method is computationally simple and appears to have good performance.

Another approach to time scale modification of speech is to consider the problem in the short-time Fourier transform (STFT) domain. The STFT is a time-frequency representation of a signal, and its magnitude is often referred to as "digital spectrogram". Spectrograms display many features of speech such as fundamental frequency and formant frequencies as a function of time, which are known to be very important for speech perception. In one method [16], the STFT of speech is modified and speech is synthesized from the modified transform. This method is related to the method by Malah, since with proper interpretation, the STFT is equivalent to the output of a bank of bandpass filters. In this method, both the magnitude and phase of the STFT are modified. For the application to time-scale modification of speech, the required modification for the STFT magnitude is very straight-forward. The required modification for the STFT phase is quite involved and careful attention has to be paid to the modification of the STFT phase to achieve good performance.

To avoid the difficulty associated with the modification of the STFT phase, another method was developed. In this method [17], only the STFT magnitude is modified and speech is synthesized directly from the modified STFT magnitude. The modification of the STFT magnitude changes the time scale without affecting the local spectral characteristics and will tend to preserve the quality and intelligibility of speech. An example that illustrates this method is shown in Figure 5. Figure 5(a) shows the spectrogram (STFT magnitude) of a speech signal. Figure 5(b) shows the modified spectrogram obtained by compressing the time scale of the spectrogram in Figure 5(a) by a factor of 2 without changing the frequency scale. Figure 5(c) shows the spectrogram of the speech signal estimated from the modified spectrogram in Figure 5(b). This method, although considerably more expensive computationally than others, appears to have the best performance among existing algorithms. Simulation results of this method demonstrate that high-quality rate-changed speech which retains the natural quality and speaker-dependent features, with few artifacts such as glitches, burbles, and reverberation, can be generated for compression ratios as high as 2.5:1 and expansion ratios as high as 4:1. In addition, the method is robust to speech degradation by additive noise in the sense that the noise in processed speech is not perceived to increase in intensity and the noise characteristics are not perceived as different. The method has also been applied successfully to time-scale modification of the singing voice and music signals.

In addition to potential applications to the speech enhancement problem, time-scale modification of speech has a number of other applications. For example, many speech recognition systems require normalization of speech sound duration without affecting the short-time spectral characteristics of speech. Other examples include speech duration change for broadcasting and movies. The algorithm dis-



(a)

(b)

(c)

Figure 5. (a) Spectrogram (STFT Magnitude) of "Line up at the screen door."

(b) Modified Spectrogram for Time-Scale compression by a factor of 2

(c) Spectrogram of speech estimated from the Modified Spectrogram in (b)

cussed in this section are also applicable to these and other examples.

V. Areas for Future Research

In the above three sections, we have discussed some representative speech enhancement algorithms. Even though these discussions are not exhaustive, they illustrate the general approaches that have been considered and indicate some directions for future research. In this section, we discuss a few topics for future research related to the speech enhancement problem.

The objective of speech enhancement is generally an improvement in some aspects of human perception such as improvement in speech intelligibility or quality. Since the human perceptual domain is not well understood, a careful system evaluation requires a subjective test, which can be tedious and time consuming. This is one of the reasons why many speech enhancement systems have not been carefully evaluated. Further understanding of the human perceptual domain and development of simple procedures to evaluate the performance of a speech processing system will be useful not only for speech enhancement, but for speech processing in general.

Various speech enhancement systems discussed in Sections II and III appear to improve speech quality, but not speech intelligibility. Intelligibility improvement when the degradation is due to wide-band random noise or speech-like noise, in my opinion, requires a fresh new approach to the speech enhancement problem. One such approach is to exploit more information about speech. Even though some algorithms such as power spectrum subtraction method and comb filtering attempt to exploit some characteristics of speech, there is considerably more knowledge about speech signals that may potentially be incorporated in speech enhancement systems. Cooperation of researchers with signal processing background and researchers with speech background would be important for such an effort.

In the area of time scale modification of speech, the performance of existing algorithms may be further improved by exploiting the notion that when a human speaks at a slower rate, not all segments of speech are articulated uniformly more slowly. For example, unvoiced sounds, which are short in duration in human articulation, appear to be affected less than voiced sounds, which are relatively long. Even though existing algorithms are capable of changing the time scale of speech at different rates for different speech segments, the question of what rates should be applied to each speech segment to achieve a certain overall rate of time scale modification is not well understood.

In this paper, we have attempted to provide an overview of the variety of techniques that have been proposed for speech enhancement. A more detailed and complete treatment of signal processing algorithms for speech enhancement can be found in [18, 19].

References

[1] H. Drucker, "Speech Processing in a High Ambient Noise Environment", *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, pp. 165-168, June 1968.

[2] J. S. Lim, "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-26, pp. 471-472, October 1978.

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-29, pp. 113-120, April 1979.

[4] R. H. Frazier, S. Samsam, L. D. Braida, A. V. Oppenheim, "Enhancement of Speech by Adaptive Filtering", *Proceedings of the Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 251-253, Philadelphia, PA, April 12-14, 1976.

[5] J. S. Lim, A. V. Oppenheim, L. D. Braida, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-26, pp.354-358, August 1978.

[6] B. Widrow, et al., "Adaptive Noise Cancelling: Principles and Applications", *Proceedings of the IEEE*, vol. 63, pp. 1692-1716, December 1975.

[7] T. W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection", *J. Acoust. Soc. Am.*, vol.60, pp.911-918, October 1976.

[8] Y. M. Perlmutter, L. D. Braida, R. H. Frazier, A. V. Oppenheim, "Evaluation of a Speech Enhancement System", *Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 212-215, Hartford, Conn., May 9-11, 1977.

[9] A. V. Oppenheim, R. W. Schafer, T. G. Stockham, "Nonlinear Filtering of Multiplied and Convolved Signals," *Proceedings of the IEEE*, vol. 56, pp. 1264-1291, August 1968.

[10] D. L. Duttweiler and Y. S. Chen, "A Single Chip VLSI Echo Canceller," *The Bell System Technical Journal*, vol. 59, pp. 149-160, February 1980.

[11] J. B. Allen, D. A. Berkley, J. Blanert, "Multimicrophone Signal Processing Technique to Remove Room Reverberation from Speech Signals." *J. Acoust. Soc. Am.*, vol. 62, pp. 912-915, October 1977.

[12] P. J. Bloom, "Evaluation of a Dereverberation Process by Normal and Impaired Listeners", *Proc. of Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 500-503, Atlanta, GA , March 30, 31, April 1, 1981

[13] G. Fairbanks, W. L. Everitt, and R. P. Jaeger, "Method for Time or Frequency Compression-Expansion of Speech", *IRE Trans. on Audio Electroacoustics*, vol. AU-2, pp. 7-12, January 1954.

[14] R. J. Scott, and S. E. Gerber, "Pitch- Synchronous Time-Compression of Speech", *Proc. Conf. on Speech, Communications and Processing*, pp. 63-65, April 1972.

[15] D. Malah, "Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-27, pp. 121-133, April 1979.

[16] M. R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-29, pp. 374-390, June 1981.

[17] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-32, pp. 236-243, April 1984.

[18] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", *Proc. of the IEEE (invited)*, vol. 67, pp 1586-1604, December 1979.

[19] J. S. Lim, editor, *Speech Enhancement*, Prentice-Hall, Inc., 1982.

INVARIANT AUDITORY WORD PATTERNS IN SPEECH PROCESSING

NATALIE WATERSON

Department of Phonetics and Linguistics
School of Oriental and African Studies
University of London, London, England

ABSTRACT

It is proposed that one of the units of speech perception is an invariant auditory word pattern. This consists not of the whole spectrum but of a limited number of acoustic cues that are auditorily salient, together with those that are less salient but carry contrastive function in the language. Speech processing takes place by pattern recognition and pattern matching. For this two levels of representation are postulated, a phonetic level, LR1, and a lexical-phonological level, LR2. Cues are abstracted from the acoustic signal and are synthesized into patterns: these are checked against patterns at LR1. If they match, they are then matched with patterns at LR2 and indentification of the word is achieved. The organization of patterns in a network is shown for a sample of a child's phonological system, and how recognition of some words takes place is illustrated. An example of a misperception is also given to show how confusions occur between words of same patterns.

It is known that there is much redundancy in speech and that speech processing is very rapid. Context, knowledge of the language, knowledge of the topic, shared knowledge, etc., are acknowledged to play a major role in the interpretation of speech. Because speech processing is so rapid, it is clear that interpretation of the acoustic signal segment by segment is not possible. Furthermore, no one-to-one acoustic correlates have been found for phoneme segments. Word by word processing is still too slow to explain the speed with which speech is processed. It is possible that the auditory processing of speech is similar in nature to visual processing in the interpretation of written texts by reading. It is recognized that in reading attention is not given to each letter of each word, nor even to each word, and that scanning takes place—not in a serial progression but with the eyes moving back and forth as they abstract the most essential information from the text to make sense of the message. Words have visual shapes which aid recognition. It seems that one may similarly look for auditory shapes of words that can be recognized in auditory scanning of the acoustic signal. There may similarly be auditory shapes of sentences where attention is not given to the whole of the acoustic signal but abstractions are made at points (stressed high information words) which will give maximum information for the minimum expenditure of time and energy. It is proposed that it is auditory word patterns that are abstracted at such points in the acoustic signal and that such auditory patterns are invariant.
Auditory patterns are essentially acoustic skeletons composed of auditorily salient cues and such less salient cues as have contrastive function in the language. Thus the pattern will consist of part of the spectrum of the word, not the whole spectrum. The cues will involve mainly features such as intensity, e.g. peaks will indicate the number of syllables, greater intensity of some peaks and lesser of others will mark strong and weak stress; duration will mark some syllables longer than others; fundamental frequency will indicate the pitch pattern; first formant will indicate the various degrees of openness of vowel, i.e. whether more close, more open, or the same as in adjacent syllables; cues for different classes of consonants will differentiate them from each other, e.g. a fricative from a nasal, a nasal from a plosive, etc., etc. Such cues in different sequences comprise the different word patterns. The cues are relative, i.e. it is not the actual intensities, actual frequencies, actual durations, etc., that are relevant, but the relationships between the cues, which are fixed. The patterns are therefore invariant and remain the same regardless of variables such as if speech is compressed, as in fast tempo, or whether spoken at a very slow tempo; whether spoken on a man's low pitch or a woman's high pitch, and whether pronounced in the standard dialect or some provincial dialect. Words in the phonological system of a language may be described in terms of invariant auditory patterns. There may be several words belonging to a single pattern or only one or two. For instance the words: train, plane, prim, cream, tram, clan, may be classed as belonging to the same pattern, plosive with fricative release + vowel + nasal, P͡FVN. The fricative release may be lateral or central, and the vowel may be long or short, open, close, or mid with glide to close. Words like pot, kick, deep, bat, boot, belong to another pattern, plosive + vowel + plosive, PVP, and so forth.
Patterns will be organized for rapid and easy retrieval in a network which has two levels of representation (leaving aside at present the semantic and syntactic levels) the first being phonetic for receiving the acoustic signal and for synthesizing patterns and for storing patterns; the second level is for storing the phonological patterns for matching with the phonetic patterns, to arrive at the identification of words.
The adult processing system, with a vocabulary of many thousands of words is extremely complex, so for illustration, a sample of a child's very early, very simple phonological system will be used to demonstrate the proposed network and how recognition takes place. The child, aged between 1;5 and 1;6, had monosyllables and disyllables involving mostly nasals and plosives. Fig. 1 shows the network of LR1 and is mainly phonetic; Fig. 2 shows LR2 which is lexical-phonological. How such levels of representation are constructed by the child and how processing takes place in relation to these levels of representation is described in detail in [1]. In the construction of LR1, patterns are synthesized on the basis of auditorily salient features of the acoustic level. A child's limited abilities, especially in perceptual discrimination, oblige him to pay selective attention to what is most acoustically and auditorily salient at first. The patterns are stored at LR1 for future matching when other words of such patterns are recognized. Patterns of LR1 are more fully specified at LR2, and meaning is included. LR1 patterns are matched with patterns at LR2 in the process of recognition. If there is no match for the synthesized pattern, a new pattern is constructed and stored.
Fig. 1 shows the monosyllablic and disyllabic patterns of the child's LR1 and Fig. 2 shows the organization of one monosyllabic pattern, the PV pattern, in LR2. The way recognition takes place is by following pathways along which choices are made which constrain the possibilities and lead to the identification of the particular word. The pathways for the words are shown by different markings (see key on figures). It will be seen that the PV pattern words follow the same path up to the point where they divide according to the degree of vowel openness, marked by α for low vowel, ε for mid and ι for high vowel. The next choice is at the three-way contrast carried by place of articulation, viz. labial p, apical t, and dorsal k. The last choice is of contrasts carried by frontness y, backness w, and neutrality as to frontness and backness, ə, and the pattern is then identified as the particular word.
In the case of a child, the early forms are based mainly on the auditorily salient features of words which are fleshed out within his current capabilities and in a way that fits his current network. As he becomes able to give more attention to less salient features, his forms of words and patterns change and his network is therefore constantly being re-structured. For instance when [bəu] 'boat' acquires a final plosive, it moves from the PV pattern to the PVP pattern, and when [gu:] 'goose' acquires a final fricative, [gu:θ] and [gu:Φ], a new pattern has to be created and incorporated into the network, viz. PVF (plosive + vowel + fricative) to which will belong newly acquired words like [bɪf] 'beef' and [gauf] 'cow' and 'calf'. Eventually the child acquires the complex network of the adult.
This concept of invariant auditory pattern can thus offer an explanation of how the acquisition of phonology takes place.
Further evidence in support of the invariant auditory word pattern can be found in studies of misperceptions (see [1]; also for references for support from other disciplines). Examples show how much the listener contributes to the interpretation from what he thinks the intended message is, making use of the minimum of acoustic cues of the pattern and the maximum use of any other available information. A brief illustration is given of the way the interpretation of an utterance is made in terms of pattern recognition, together with use of context, shared knowledge, and other factors. It will be shown how non-linguistic information influences the interpretation of a pattern which results in the identification of the wrong word.

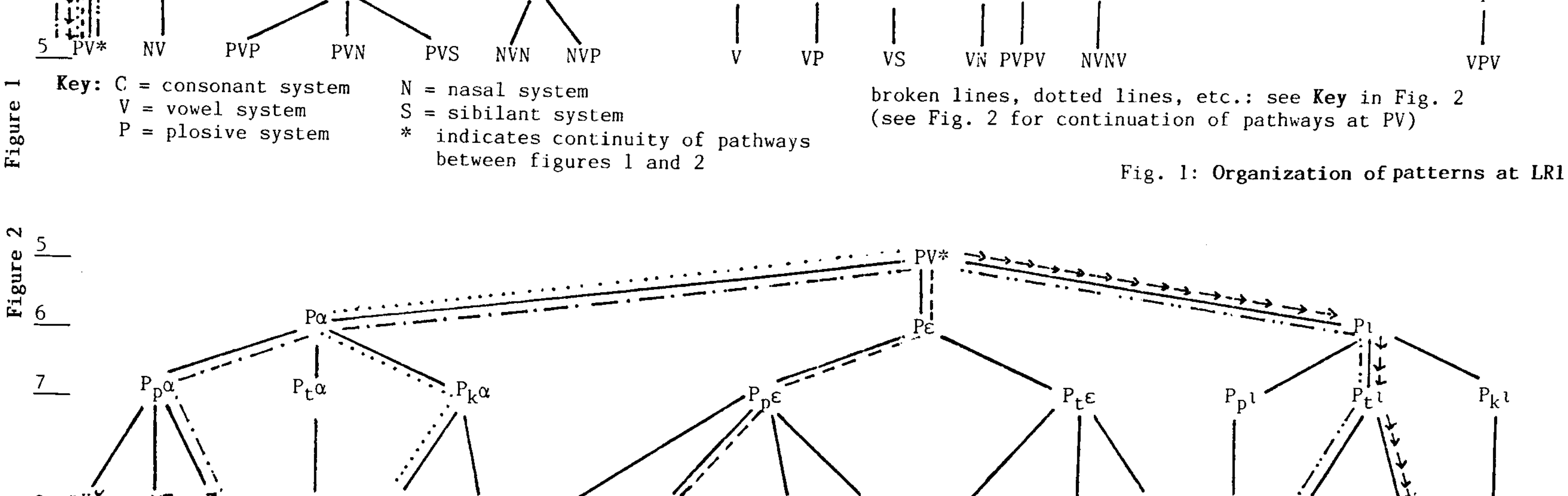
Context: Saturday morning. A and B are in the bathroom and the bath is being filled with water, so there is a loud noise of rushing water which has a masking effect. A and B had just been talking about changing the positions of their parked cars to enable A to take C to the station to get the 8.48 train. The following conversation then takes place.

B: We must get the E.45 cream today.
A: Today? Why today?
B: Why not?
A: Why on a Saturday?
B (Realizing that A has got the message wrong):
  I said 'We must get the E.45 cream today.'
A: Oh, I thought you said 'We must get the 3.45
  train today.'

A was still geared to the semantic field of trains and train times and did not realize the change of topic, and as B and her husband often came for weekends, arriving on Friday night and usually leaving on Sunday, A misinterpreted the pattern common to 'cream' and 'train', viz. plosive with fricative release + vowel + nasal, P͡FVN, as 'train'. She also interpreted 'E' [i:] as 'three' [θṛi:]. Voiceless non-salient [θṛ] would easily be masked by noise and a listener would therefore be ready to 'restore' it where needed, as here, where '[i:] forty-five' could only mean '3.45' in terms of train times. A having recognized the pattern P͡FVN, it is possible that detailed pattern matching would be skipped as the context so clearly predicted 'train'. In fact, B was referring back to the previous day's conversation (shared knowledge) about a cream called E.45 which she had recommended to A.

This example shows how the processing of the invariant auditory word pattern in combination with the use of non-acoustic information can speed up the rate of speech processing. Because of adults' huge vocabularies and complex phonetic, phonological, semantic and syntactic systems, and their fast rate of speaking, adults need to use the maximum possible short cuts in processing. The concept of invariant auditory word pattern makes it possible to explain how short cuts in processing can be made and why speech processing can be as rapid as it is.
Intonation patterns have long been described as a limited number of invariant tunes and the problem of normalization across variables as a man's use and a child's use of the tunes does not arise. Similarly, the proposed auditory word pattern is also invariant and the problem of normalization across variables such as age, sex, speech rate, and dialect need no longer arise.

Figure 1



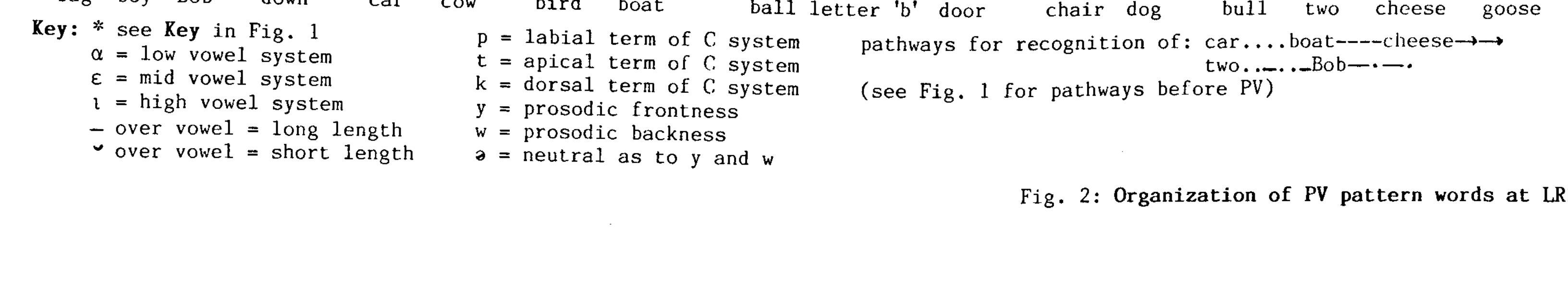
Key: C = consonant system
V = vowel system
P = plosive system

N = nasal system
S = sibilant system
* indicates continuity of pathways between figures 1 and 2

broken lines, dotted lines, etc.: see **Key** in Fig. 2
(see Fig. 2 for continuation of pathways at PV)

Fig. 1: **Organization of patterns at LR1**

Figure 2



Key: * see **Key** in Fig. 1
α = low vowel system
ε = mid vowel system
ι = high vowel system
‒ over vowel = long length
˘ over vowel = short length

p = labial term of C system
t = apical term of C system
k = dorsal term of C system
y = prosodic frontness
w = prosodic backness
ə = neutral as to y and w

pathways for recognition of: car....boat----cheese→→
two.‒.‒.‒Bob‒.‒.
(see Fig. 1 for pathways before PV)

Fig. 2: **Organization of PV pattern words at LR2**

[1] N. Waterson, 《Prosodic Phonology: The Theory and Its Application to Language Acquisition and Speech Processing》, Grevatt & Grevatt, 1987.

INTONATORISCHE MERKMALE IN DER PERZEPTION DER WORTGRENZEN IM SATZ

ZDENA PALKOVÁ

Lehrstuhl für Linguistik und Phonetik,
Philosophische Fakultät der Karlsuniversität
116 38 Prag, Tschechoslowakei

ZUSAMMENFASSUNG

Der Tonhöhenverlauf gehört im Tschechischen zu den relevanten Merkmalen, die die Wahrnehmung der Wortgrenzen bedingen. Dabei charakterisiert er das Wort als Ganzes, und nicht nur die sogenannte betonte Silbe. Diese Feststellung kann experimentell bestätigt werden.

EINLEITUNG

1.1 Unter den prosodischen Merkmalen in der linguistischen Beschreibung des gesprochenen Tschechisch ist der Wortakzent besonders durch seine Gebundenheit an die erste Wortsilbe wichtig; in bezug auf die Wortgrenzen wird seine delimitative Funktion angenommen. Es fehlt jedoch eine zuverlässige Erklärung seiner phonetisch-akustischen Merkmale, d.h. der Beziehung zwischen Sprachsignal und Perzeptionsergebnis. Zweifellos ist der Wortakzent im Tschechischen eine komplexe Erscheinung /3/. Der Einflu3 sowohl der Höhe, als auch der Stärke und Dauer des Tones auf die Perzeption des Wortakzents wurde experimentell bestätigt /1/.

1.2 Andere Experimente haben jedoch auch gezeigt, da3 die Wahrnehmung einer Silbe als akzenttragend nicht direkt mit ihren Lautqualitäten erklärbar ist, auch nicht, wenn man sie in Relation zu den beiden benachbarten Silben setzt. Wesentlichen Einflu3 haben offensichtlich der breitere Kontext und strukturelle Eigenschaften wie z.B. die Länge des Wortes /2/. In diesem Zusammenhang wurde die Lypothese aufgestellt, da3 die Klanggestalt der elementaren rhythmischen Einheiten auf der Wortebene (Takte) in gewissem Ma3e standardisiert ist, d.h. da3 einige Schallstrukturen vom Hörer mehr im Sinne einer Einheit wahrgenommen werden als andere Schallstrukturen.

1.3 Die angeführten Erkenntnisse haben wir auf der Grundlage von Material aus der natürlichen Sprachen gewonnen, wobei der Einflu3 der syntaktischen und semantischen Komponenten beseitigt wurde. Bei diesem Material sind alle Tonqualitäten veränderlich und der Versuch, eine längere Silbenkette eingehender zu analysieren, führt daher zu einer gro3en Anzahl von zu unterscheidenden Typen mit jeweils geringer Anzahl der zugehörigen Fälle. Deshalb untersuchen wir derzeit die einzelnen Schalleigenschaften getrennt unter Verwendung von synthetisch erzeugtem Material.

Das Material für die im folgenden angeführten Tests wurde in Zusammenarbeit mit Dr. Ing. M. Ptáček aus dem Forschungsinstitut für Kommunikationstechnik im Prag erstellt. Die verwendete Apparatur war der von ihm konstruierte Synthesator HOP-2.

METHODE

Wir gehen in diesem Beitrag von den Ergebnissen unseres Experiments zum Tonhöhenverlauf aus.

2.1 Grundlage dieses Experiments war eine Serie von Hörtests, in denen tschechische Muttersprachler (Philologiestudenten der Philosophischen Fakultät der Karlsuniversität Prag, Alter 18-21 Jahre) Silbenketten in die elementaren rhythmischen Einheiten auf Wortebene (Takte) zerlegen sollten. Der Tonhöhenverlauf in diesen Silbenketten wurde als Variable in fünf aufeinanderfolgenden Vierteltonstufen realisiert, was die grö3te mögliche Veränderung eines ganzen Tons darstellt. Dabei verwendeten wir zwei Arten von Material.

2.2 In der ersten Etappe (Serie A) dienten als Material Silbenketten, die durch fünfmaliges Wiederholen der synthetisch erzeugten Silbe SE gebildet wurden und in ihrer Länge kurze tschechische Sätze oder selbständige Satzteile (also in Bezug auf die Satzintonation selbständige Tongruppen) darstellen können.

Die Serie A beinhaltete 80 verschiedene Intonationsvarianten dieser Silbenketten. Dabei waren alle Typen der Veränderung von F_o, die bei einer fünfgliedrigen Silbenkette möglich sind, vertreten (unter der Voraussetzung, da3 zwei benachbarte Silben nie dieselbe Tonhöhe aufweisen). Dabei wurden die Hörer in ihrer Entscheidung durch die gegebenen Instruktionen eingeschränkt. Sie sollten sich für eine von drei Möglichkeiten entscheiden: die Gliederung der Silbenkette im Verhältnis 2:3 (also nach dem Rhythmusschema xx xxx) oder im Verhältnis 3:2 (also nach dem Rhythmusschema xxx xx) oder aber in keine der beiden Varianten.

Die Serie A dauerte 19 Minuten und wurde in zwei Gruppen von insgesamt 31 Versuchspersonen durchgeführt.

2.3 In der zweiten Etappe (Serie B) dienten als Material kurze tschechische Sätze oder selbständige Satzteile (Tongruppen) in denen die Bestimmung einer Wortgrenze bedeutungsunterscheidend ist.

Z.B. "včera to/ pili/ neradi" - wörtlich übersetzt: gestern haben sie das nicht gern getrunken;
"včera /topili/ neradi" - wörtlich übersetzt: gestern haben sie nicht gern geheizt.

Im Tschechischen verkörpert dieses Beispiel den Unterschied zwischen den Rhythmusstrukturen xxx xx und xx xxx.

Für die Serie B wurden 13 solcher Sätze mit einer Länge von 2 bis 4 Takten zusammengestellt. Aufgabe der Hörer war es, sich in jedem einzelnen Fall für eine der beiden möglichen Bedeutungen zu entscheiden.

Die Serie setzte sich aus drei voneinander unabhängigen Tests mit einer jeweiligen Länge von 18 Minuten zusammen. In einem Test war jeder Satz immer in 4 verschiedenen Varianten des Tonhöhenverlaufs enthalten. Für einen einzelnen Satz wurden also 12 verschiedene Modifikationen von F_o zur Anwendung gebracht, wobei für Sätze mit gleicherma3en variiertem Rhythmusschema dieselben Varianten des Tonhöhenverlaufs verwendet wurden. Insgesamt kamen in der Serie B 36 verschiedene Varianten des F_o-Verlaufs zur Anwendung. Ihre Auswahl erfolgte auf der Grundlage der Ergebnisse der Tests aus Serie A.

Jeder Test der Serie B wurde von 30 Versuchspersonen absolviert. Von ihnen absolvierten 15 Personen alle 3 Tests.

ERGEBNISSE

3.1 Die einzelnen Beispiele sowohl der Serie A als auch der Serie B wurden in verschiedener Weise bewertet. Der Vergleich der unterschiedlichen Hörergruppen innerhalb jeder der beiden Serien weist eine statistisch sehr signifikante Übereinstimmung auf (Wilcoxon, 0,01). In beiden Fällen war die Aufgabe also für die Versuchspersonen lösbar.

Die maximale Übereinstimmung bei der Bewertung eines einzelnen Beispiels betrug in der Serie A 83%, in der Serie B 93%. In den Tests auf der Grundlage der natürlichen Sprache hatte sie bei 94% gelegen.

In beiden Serien zeigte sich eine Neigung der Hörer, einem der Typen den Vorzug zu geben. In der Gesamtbetrachtung der Serie A registrieren wir eine Bevorzugung der Rhythmusstruktur xx xxx gegenüber der Struktur xxx xx um 9%. Demgegenüber wurde in der Serie B die Struktur xxx xx bevorzugt und zwar um fast 20%. Bei Serie B mu3 man allerdings auch den Einflu3 der Syntax und der Semantik beachten. Auch individuelle Einflüsse können nicht ausgeschlossen werden.

3.2 Die Ergebnisse der Serie A bestätigen, daß die Bestimmung der Wortgrenze durch den Hörer nicht auf Grund der Tonhöhe der ersten Silbe der bei dieser Bestimmung entstandenen Wörter erklärt werden kann. Es zeigt sich diesbezüglich eine gewisse Tendenz, wonach die erste Silbe des entstandenen Wortes um einen Viertelton höher liegt, als die vorausgegangene Silbe. Das allein ist jedoch für die Zerlegung der Silbenkette in Wörter nicht ausreichend (nur in 28% der theoretisch möglichen Fälle setzte sich diese Tendenz tatsächlich durch).

In der Serie A war bei 23% der Beispiele eine Übereinstimmung von mehr als 65% der Hörerurteile zu verzeichnen. Bei der Analyse dieser Beispiele wurden einige Tendenzen des Tonhöhenverlaufs festgestellt, die die Entscheidung, ob sich an der gegebenen Stelle eine Wortgrenze befindet, positiv bzw. negativ beeinflussen. Dabei können Fälle auftreten, in denen die negativ wirkenden Tendenzen stärker als die positiv wirkenden sind. Wichtig ist offensichtlich, daß die beiden benachbarten Wörter bezüglich ihres Tonhöherverlaufs akzeptierbar sein sollten. So war z.B. innerhalb eines dreisilbigen Wortes (zwischen der 2. und 3. Silbe) eine Veränderung von F_o um einen halben Ton für die Hörer annehmbar, zwischen zwei Wörtern war eine solche Veränderung dagegen kaum annehmbar (vgl. /4/).

3.3 Auf der Grundlage der in Serie A gewonnen Erkentnisse wurden die Varianten des Tonhöhenverlaufs für die Serie B ausgewählt. Die Auswahl war von dem Versuch motiviert, bei jedem Satz beide Varianten der Gliederung und damit der Bedeutung zu erhalten.

In der Serie B wurde sogar bei 70% der Beispiele eine Übereinstimmung von mehr als 67% der Hörer erzielt, bei 41% der Beispiele betrug die Übereinstimmung über 73%.

Die von uns theoretisch erwartete Gliederung der Sätze wurde bei der Gesamtauswertung der in Serie B beobachteten Ergebnisse eindeutig bestätigt (X^2, 0,01). Auch bei der selbständigen Bewertung der Sätze im Einzelnen traf unsere Vorhersage überwiegend zu, nur bei 3 Sätzen war die Differenzierung zwischen den beiden möglichen Gliederungsvarianten statistisch nicht signifikant.

Die Wirkung der verwendeten Varianten des Tonhöhenverlaufs unterschied sich etwas von den Ergebnissen der Serie A. So hatten von 12 Varianten, an Hand derer die in 7 Sätzen enthalteten Strukturen $\overset{'}{x}xx\ \overset{'}{x}x$ und $\overset{'}{x}x\ \overset{'}{x}xx$ unterschieden werden sollten, neun die erwartete und eine entgegengesetzte Wirkung (X^2, 0,05). Zwei Varianten führten zu einem indifferenten Ergebnis.

Die Ergebnisse bestätigen die Relevanz der Wortlänge und zeigen weiterhin, daß auch die unterschiedliche Stellung des Wortes in der übergeordneten Intonationseinheit beachtet werden muß. So ist in unserem Material z.B. eine Zerlegung in zwei aufeinanderfolgende dreisilbige Wörter leicht zu erreichen. Dagegen wird eine Unterscheidung der Strukturen $\overset{'}{x}x\ \overset{'}{x}xx$ und $\overset{'}{x}xx\ \overset{'}{x}x$ nur schwer erreicht, wenn noch ein einsilbiger Takt folgt.

SCHLUSSFOLGERUNGEN

Aus den Ergebnissen unserer Untersuchungen kann man schließen:

Der Tonhöhenverlauf innerhalb einer Silbenkette stellt einen relevanten Faktor für die Zerlegung dieser Kette in elementare rhythmische Einheiten auf Wortebene dar.

Der Tonhöhenverlauf des ganzen Taktes hat in unserem Material eine größere Bedeutung, als die melodische Charakteristik einzelner akzenttragender Silben. Es gibt Varianten des Tonhöhenverlaufs, die die Wahrnehmung einer Silbenkette als einheitliches Ganzes unterstützen, und andere, die eine solche Bewertung erschweren. Dabei kommt es auch darauf an, daß die beiden benachbarten Takte bezüglich ihres Tonhöhenverlaufs für die Hörer akzeptierbar sind.

In unserer weiteren Arbeit wollen wir nun versuchen, auf der Grundlage der erzielten Ergebnisse Formeln anzugeben, die für die automatische Synthese des Tonhöhenverlaufs in tschechischen Sätzen verwendet werden könnten.

LITERATUR

/1/ Janota P. - An Experiment Concerning the Perception of Stress by Czech Listeners, AUC - Phonetica Pragensia I, 1967, pp. 45-68

/2/ Janota P., Palková Z. - The Auditory Evaluation of Stress under the Influence of Context, AUC - Phonetica Pragensia IV, 1974, pp. 29-59

/3/ Ondráčková J. - On the Problem of the Function of Stress in Czech, ZPSK 1961, 14, 1, pp.45-54

/4/ Palková Z. - Fundamental Frequency Variation as a Factor of Word-Stress Perception, Proc. of the 23rd Acoustic Conference on Physiological and Psychological Acoustics, Acoustics of Speech and Music, České Budějovice, 1984, pp.182-185

ARTIKULATORISCHE KORRELATE DES FESTEN UND LOSEN ANSCHLUβES IM DEUTSCHEN (ANHAND DES RÖNTGENFILMS)

LARISSA PROKOPOWA

Schewtschenko-Universität Kiew
Lehrstuhl für Deutsche Philologie

Das Referat behandelt die artikulatorischen Korrelate der Sprechbewegungen der Zunge anhand des Röntgenfilms bei der Deutung solcher Erscheinungen wie der feste und lose AnschluB der Konsonanten im Deutschen.

Da der sogenannte feste und lose AnschluB der Konsonanten im Deutschen als Element der intersilbischen Struktur des starken und schwachen Silbenschnittes betrachtet wird, ist es notwendig, die Deutung der artikulatorischen Korrelate von dem Standpunkt der allgemeinen Silbenstruktur aus zu verwirklichen und zwar in Termini der Sprechbewegungen der Zunge. In die vorliegende Untersuchung geht die Annahme ein, daB die Spezifik des konsonantischen An- und Auslautes einer Silbe durch die Spezifik der Sprechbewegungen bedingt wird, wobei der feste und der lose AnschluB der Konsonanten sich ebenso aus den Besonderheiten der Sprechbewegungen ergibt. Um Informationen über die Spezifik der Sprechbewegungen zu erlangen, müBte man das Energierelief der Silbe in allgemeinen Zügen rekonstruieren. Es kommt bis jetzt in Frage, wie solche Begriffe wie Silbengipfel und Silbennaht mit zuverlässigen artikulatorischen Merkmalen zu identifizieren wären.Die Untersuchung ist gerade dieser Problematik gewidmet und aufgrund des Röntgenfilms verwirklicht (Röntgenanlage "Gigantos" von der Firma Siemens. Geschwindigkeit 50 Bilder pro Sekunde). Es wurde die Methodik der Abmessung jedes Bildes anhand des radialen Koordinatensystems von dem physiologischen Zentrum - 7,5°, 15°, 30°, 45°, 60°, 90°, 120°, 150°, 180° (Abb.1) mit dem MeBgerät von Barinowa (Abb.2) das die Zungenlage in Prozentangaben fixiert (Entfernung vom Zungenrücken bis zum Gaumen), wodurch verschiedene geometrische Konfigurationen der Mundhöhle bei verschiedenen Sprechern aufgehoben werden.
Als Sprechmaterial dienten deutsche Wörter von 3männlichen und 2 weiblichen native Sprechern gesprochen.. Es wurden die sogenannten Schlüsselvokale- a: a, u: ʊ, i: I, o: ɔ, e: ɛ in der Umgebung der VerschluB- und Engekonsonanten untersucht, also KVK, wobei der letzte Konsonant zu verschiedenen Silben angehören könnte: KVK,KV:-K, KV:K. Das bedeutet, daB alle drei Silbentypen vertreten waren: offene, geschlossene und die sogenannte quasi geschlossene Silbe. Insgesamt wurden 6000 Bilder gemessen.

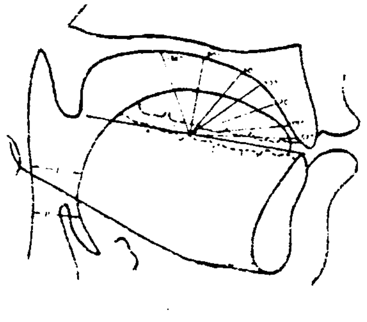


Abb.1 Das radiale Koordinatensystem

Abb.2 Das MeBgerät von Barinowa

Das Hauptziel der Untersuchung bestand darin, um den Moment des Überganges von der lösenden Bewegung des Konsonanten am Anfang der Silbe zu der schlieBenden Bewegung in drei Silbentypen festzustellen, der hypothetisch mit dem artikulatorischen Silbengipfel identifiziert wurde. Das radiale Koordinatensystem erwies sich als ausreichend: jede lösende sowie schlieBende Bewegung wurde in der Regel an zwei Koordinaten fixiert: t, d, n, ts- an den Koordinaten 7,5°, 15°; k - an den Koordinaten 90°, 120° mit Ausnahme von g und ° , welche oft nur an der Koordinate 120° fixiert wurden. Die spezifischen Hebungen des Zungenrückens bei der Artikulation der Vokale wurden auch an denselben Koordinaten beobachtet: u:,ʊ - an den Koordinaten 90°, 120°; i:, I - 7,5°, 15°. Für die Kurzvokale der mittleren Zungenlage ɔ, ɛ waren oft die Angaben an Grenzkoordinaten wichtig. Für die Vokale a: a war die Konsonantenumgebung entscheidend und zwar in der Umgebung von t, n waren die Angaben an der Koordinate 60° ausschlaggebend.

Relativ eindeutig erfolgte die Identifi-
zierung des artikulatorischen Silbengip-
felsin den Verbindungen a: a, wo die Maxi-
malwerte der Senkung den Silbengipfel an-
deuteten. Als sekundäres Merkmal diente
die Umstellung des vorderen Artikulators,
wo zuerst die VergröBerung der Werte und
dann - die allmähliche Verminderung. Der
Moment der Umstellung fiel mit dem Moment
des Maximalwertes zusammen,was einUmschlag
bestätigte. Nur bei den Kurzvokalen mit
einer geringen Dauer (40, 60 msek) fehlt
in der Regel das sekundäre Merkmal.
Ein schwerer Fall der Identifizierung des
artikulatorischen Umschlages stellt die
schlieBende Bewegung in der Verbindung
des Konsonanten mit dem homorganischen Vo-
kal - u:g, ʊk, ʊŋ , sowie i:t, It. Der
eigentliche Widerspruch besteht darin,daB
der Vokal und der homorganische Konsonant
mit einer einheitlichen Bewegung produ-
ziert werden und allmähliche Verminderung
der Werte gibt keinen AnlaB für die Fest-
legung der Umstellung. Nur an den periphe-
ren Koordinaten tritt das sekundäre Merk-
mal in den Vordergrund und erfüllt die
Funktion des primären Merkmals. Der Aus-
fall der Umstellung an der Hauptkoordina-
te wäre als Ausdruck der starken Koartiku-
lation infolge der Verschmelzung des Vo-
kals mit dem homorganischen Konsonanten
zu interpretieren. Es wäre notwendig, noch
zwei Arten der Koartikulation zu unter-
scheiden und zwar Anpassung der peripheren
Teile und allgemeine Hebung oder Senkung
der Zunge an den Haupt- und Nebenkoordina-
ten.Periphere Teile können die Bewegung
an den Hauptkoordinaten verstärken oder
schwächen. Verfolgen wir die Koartikula-
tion an der Koordinate 120° während der Be-
wegung von u:. Mit dem Strich sind die
Werte für die letzte Phase des VerschluBes
angemerkt:

60/ 40, 40, 35, 24, 15 /0 Zug(mittel)
60/ 50, 40, 28, 20, 22, 10 /0 zugehen
60/ 48, 40, 30, 30, 10 /0 Zug(ochs)
60/ 50, 50, 45, 2?, 18/0 Zugang
0/ 15, 19, 12, 19 /60 (zu)gut(ekommen)

Niedrige Werte kennzeichnen die Hebung des
Zungenrückens bei der Artikulation u: im
Wort (zu)gut(ekommen) unter dem EinfluB
des vorangehenden g, wobei ts in der Ver-
bindungen Zug- eine tiefere Zungenlage
hervorruft.
Ubrigens wären einige spezifische Besonder
heiten der Koartikulation zu vermerken.
Eine starke Koartikulation entwickelt sich
im Rahmen der Silbe und der konsonantische
Auslaut übt einen stärkeren EinfluB auf den
den Anlaut als umgekehrt, z.B. steht im
Wort mutig t unter einem starken EinfluB
des ç-Lautes, während im Wort Mutti keine
Spuren so einer Koartikulation zu finden
sind. Das bedeutet, daB die Koartikulation
nicht von dem Vokal, sondern von dem Kon-
sonanten regressiv geregelt wird.

Wie es zu vermuten wäre, vollziehen sich
die Bewegungen im vorderen Teil der Mund-
höhle bei der Artikulation der Vokale i:
I, e: im geringerem MaBstab, wodurch
die Umstellung nicht so kraB zum Ausdruck
kommt und an der Hauptkoordinate bleiben
die Merkmale meistens aus. Trotzdem tre-
ten die Merkmale der Umstellung an den
peripheren Koordinaten auf.
Bei den Vokalen e: werden die Merkmale
der Umstellung schon an den Hauptkoordi-
naten fixiert, an den peripheren Koordina-
ten treten auch sekundäre Merkmale auf,
dabei reagieren die Koordinaten 30°, 45°
sowie 90° und 120° einheitlich mit.
Die Identifizierung der Umstellung bei
o: ɔ ist ähnlich wie bei a: a. Die Ma-
ximalwerte bedeuten hier die maximale
Senkung, die sekundären Merkmale an den
peripheren Koordinaten tragen zur Identi-
fizierung bei. Der Begriff des Maximal-
wertes also ist gültig nur für a: a,o:ɔ,
für andere Vokale gelten die Umstellungs-
werte als Merkmale des Silbengipfels.
Auf der Tabelle 1 sind Ergebnisse der Pla-
zierung des artikulatorischen Silbengip-
fels dargestellt, woraus folgt, daB die
Plazierung gar nicht stabil ist und in
keinem Fall mit dem zeitlichen Mittel
punktdes Vokals zusammenfallen könnte.
Die Verteilung der Plazierung hängt von
dem Silbenschnitt ab: in der offenen Sil-
be entsteht der Silbengipfel in der er-
sten Hälfte der Vokaldauer, in der ge-
schlossenen Silbe ist der Silbengipfel
schon in der zweiten Hälfte des Vokals
plaziert. In der quasigeschlossenen Sil-
be sind gemischte Züge vorhanden. Asym-
metrischer Bau läBt vermuten, daB das
Energierelief der Silbe zur Plazierung
des Silbengipfels im Einklang stehen sol-
lte; die schlieBende Bewegung mit gröBe-
rem Energieimpuls in der geschlossenen
Silbe, also mit gröBerer Geschwindigkeit
verwirklicht werden sollte.
Die Ergebnisse der Erforschung der Ge-
schwindigkeit der schlieBenden Bewegung
in der geschlossenen Silbeim Vergleich
mit der offenen und quasigeschlossenen
Silbe sind als Zahlenangaben der Zungen-
lage in den letzten 20 msek vor dem Ver-
schluB dargestellt, Als Ausgangspunkt
diente die Annahme, daB die Uberwindung
einer gröBeren Entfernung eine gröBere
Geschwindigkeit der Sprechbewegung bestä-
tigt. Die Merkmale der Zungenlage in den
letzten 20 msek zeugten, daB es doch
einen Unterschied gäbe, der aber nicht
für alle Verbindungen gültig ist. Tatsäch-
lich, wenn die schlieBende Bewegung eine
gröBere Entfernung in Verbindungen a:t,
at, u:t, ʊt, o:t, ɔt wiedergibt, so ent-
steht das entgegengesetzte Resultat in
den Verbindungen i:t, It, was bedeutet,
daB gerade eine relativ gröBere Entfer-
nung nach dem Langvokal überwunden wird.

Tabelle 1

Plazierung des artikulatorischen Silbengipfels in drei Silbenschnitten, Mittelwerte der Zungenlage in den letzten 20 msek vor dem Verschluß und Trägheitsverzögerung

Homorganische lösende und schließende Bewegung der Konsonanten

KV:-K	52%	43,43>0	20 msek	
KVK	75%	48,59>0	40 msek	a: a
KV:K	66%	25,30>0	60 msek	
(Vati,Tat,Schatzi...)				
KV:-K	40%	26,30>0	20 msek	
KVK	85%	60,60>0	20 msek	u: ʊ
KV:K	57%	29,36>0	60 msek	
(mutig, Mutti, tut...)				
KV:-K	30%	19,20>0	20 msek	
KVK	85%	15,18>0	40 msek	i: I
KV:K	57%	29,29>0	20 msek	
(Viehzucht, Lied, litt...)				
KV:-K	30%	22,22>0	40 msek	
KVK	78%	26,29>0	20 msek	e: ɛ
KV:K	40%	12,16>0	40 msek	
(D-Zug, Beet, Bett...)				
KV:-K	-			
KVK	60%	53,56>0	40 msek	o: ɔ
KV:K	40%	50,48>0	60 msek	
(- , Pol, tot, Otto...)				

Heterorganische lösende und schließende Bewegung der Konsonanten

KV:-K	-			
KVK	70%	20,15>0	40 msek	a: a
KV:K	53%	28,29>0	60 msek	
(Tag, mißachten...)				
KV:-K	-			
KVK	60%	12,15>0	20 msek	u: ʊ
KV:K	40%	10,08>0	60 msek	
(Zugang, Zug, gut, Zucht...)				
KV:-K	-			
KVK	83%	20,20>0	40 msek	i: I
KV:K	60%	20,19>0	40 msek	
(Pieck, Pick, antik...)				
	-			
KV:-K	80%	45,48>0	20 msek	ɔ
(-, Gott, -,...)				

Ebenso zweideutig sind die Ergebnisse mit dem konsonantischen Auslaut g/k.Aber zwei Erwägungen gestatten nicht, diese Ergebnisse ohne weiteres außer Acht zu lassen. Erstens, wenn diese Erscheinung in der Kontraststellung, d.h. im Rahmen einer Silbe oder eines Wortes betrachtet wird, so tritt eine rel schnellere Bewegung auf im Vergleich mit der lösenden Bewegung. Der Unterschied gilt nicht nur für t , sondern auch für ŋ :

toll 0<40,46 53,56>0; 0<36,38 55,60>0
Gott 0<18,18 38,43>0; 0<18,0§ 25,30>0
Schatzi 0<43,46 55>0; 0·31,38 35,38>0
Zugang 0<10,20 48,42>0; 0<28,20 48,38>0

Sogar für die quasigeschlossene Silbe ist ähnliches Verhältnis oft vorhanden:

Lie 0<18,20 0 26,29>0
to 0<37,41 0 60,60>0

Weitere Beispiele in der Kontraststellung: Bettecke 0 16,20 und Bettdecke 0 20,22; Zugang 0 28,20 und Zuganker0 20,18 zeigen, daß die Sprechbewegungen fein geschliffen sind und es ist nicht ausgeschlossen, daß eine geringe Verzögerung auf die Silbennaht deutet.

Aus der Tabelle 1 ist es auch ersichtlich, daß die Trägheitsverzögerung in Betracht gezogen wurde. Diese Erscheinung folgt auch aus der Ungleichmäßigkeit der Sprechbewegungen. Abgesagt davon, daß einige Teile des Zungenrückens während der Artikulation stabil bleiben können, vollziehen sich die Bewegungen sogar an den Hauptkoordinaten nicht gleichmäßig, d.h. nicht gleiche Entfernungen in gleichen Zeitabschnitten, sondern mit Hemmungen. Einige Hemmungen sind entgegengerichtet, aber lassen sich nicht an allen Koordinaten scheinen.

Betrachten wir diese Verzögerung bei der Artikulation des Vokals e: im Wort D-Zug: Hier werden die Werte von 40 msek bis 100 msek der Dauer des Vokals angegeben:

	0°	7,5°	15°	30°	45°	60°	90°	120°
40 msek	22	24	22	24	23	20	13	16
60 msek	26	24	23	22	22	20	19	15
80 msek	$\underline{20}$	$\underline{23}$	23	$\underline{24}$	$\underline{26}$	28	20	22
100msek	24	24	23	23	23	29		30

Die unterstrichenen Werte bezeichnen die allgemeine Senkung des Zungenrückens an bestimmter Koordinaten , nach der wieder eine Hebung des vorderen Zungenrückens eintritt, die mit dem Verschluß in weiteren 20 msek endet. Diese Trägheitsverzögerung geschieht immer nur nach der Umstellung, also nach dem Silbengipfel, und unterscheidet sich im Zusammenhang mit dem Silbenschnitt. In der offenen Silbe entsteht eine Trägheitsverzögerung gegen das Ende der Vokaldauer, in der quasigeschlossenen Silbe - nach dem Mittelpunkt, in der geschlossenen Silbe kann überhaupt aus bleiben, was oft geschieht, oder in der Endphase der Dauer. Manchmal kann man die Spuren der Verzögerungsogar im folgenden Konsonanten finden, z.B. während der Artikulation ts im Wort Schatzi:

	0°	7,5°	15°	30°	45°	60°	90°	120°
60 msek	12	0	0	50	60	64	65	60
80 msek	12	0	0	$\underline{52}$	$\underline{62}$	65	$\underline{68}$	$\underline{64}$
100msek	12	0	0	50	59	65	60	62

Im Zeitabschnitt 80 msek senkt der Zungenrücken an den Koordinaten 30°45°90°120° beim vollen Verschluß und in den nächsten 20 msek erhebt sich wieder.

Es kommt in Frage, ob mit dieser Verzögerung eine Silbennaht oder eine Folge der Umstellung angedeutet wird.

Differenzierte Plazierung des Silbengipfels bestatigt die bekannte Beschreibung von E.Sievers, der den Effekt des festen AnschluBes darin sah, daB der entsprechende Vokal am Gipfel der Schallfulle "abgeschnitten worden war. Man konnte meinen, daB eine schnellere schlieBende Bewegung im Moment der naximalen Schallfülle viel zu diesem Effekt beitragen kann.

Die Ergebnisse widersprechen auch nicht den Resultaten der experimentellen Untersuchungen von O.Essen und E.Fischer-Jørgensen, daB der nach dem festen AnschluB folgende Konsonant mit gröBerer Energie erzeugt wird als nach dem losen AnschluB.

Die Diskussion uber den festen und losen AnschluB der Konsonanten scheint im Grunde genommen der Diskussion über die Bedeutung der Vokaldauer im Deutschen ähnlich zu sein. Das Merkmal der Dauer ist ebenso schwach wie das Merkmal des festen AnschluBes, weil beide keine Spur der absoluten Eigenschaften haben und nur durch den Vergleich zum Ausdruck kommen. Die endgültige Bewertung dieser Merkmale hängt von dem Aussprachestil, dem Kontext und den lokalen Besonderheiten der Sprecher ab. Als Element des Kurzvokals und des folgenden Konsonanten im Deutschen kann der feste AnschluB nur als intersilbische Erscheinung betrachtet werden.

Die Erforschung der fein geschliffenen Sprechbewegungen verlangt eine VergröBerung des Sprechmaterials, aber wie es schon in dieser Pilotuntersuchung gezeigt worden war, konnen weitere Merkmale der Silben- und Worttrennung gefunden werden.

Die Sprechbewegungen werden nicht nur von den physiologischen Faktoren bedingt, sondern auch von den spezifischen Silbenprosodik geregelt. Die Koartikulation unterliegt auch dieser Regelung. Der beste Beweis dafur wäre ein Vokal aus der umgekehrten Reihenfolge der kleinen Segmenten zu synthetisieren. Z.B. es kann kaum gelingen, aus u: im Wort Zug bei der umgekehrten Reihenfolge der Segmente ein u: im Wort gut zu produzieren.

Akustische Korrelate des festen und losen AnschluBes können durch die empirischen Perzeptionsversuche festgelegt werden.

TWO ISSUES IN ESTONIAN PHONOLOGY - QUANTITY AND PALATALIZATION

MATI HINT

Department of Estonian Language and Literature
Tallinn Teacher Training Institute

ABSTRACT

There are two problems in the Estonian phonology, solutions of which are typically non-unique - quantity and palatalization. Both contrasting quantity and palatalization occur in stressed syllables and affect morphophonology. The non-unique interpretations of quantity and palatalization reflect various phonemic qualities of these phenomena.

QUANTITY

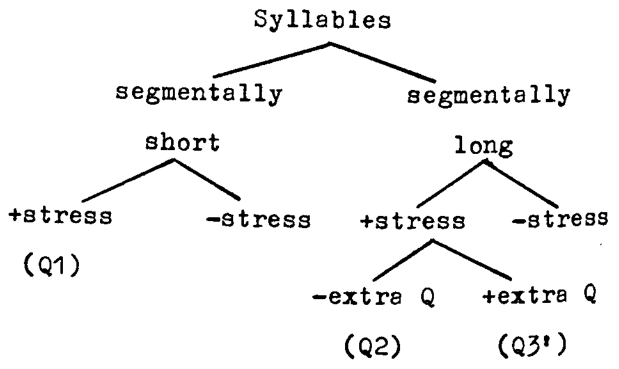
The scheme of phonological analysis should give a classification of syllables with regard to their segmental and prosodic structure.

In the recent years several new schemes and descriptions of Estonian prosody have been presented. These schemes express differing conceptions of their authors about this complex subject. Leaving aside the descriptive adequacy of different schemes, it is possible to examine their phonètic naturalness.

The following principles are involved: (1) the binary branching is more natural than a tertiary one; (2) prosodic modification of long syllables is more natural than the modification of short syllables; therefore, it is more natural to give segmental specifications of a syllable before the prosodic analysis, not vice versa.

The following is an attempt to estimate some schemes of prosodic analysis of Estonian from these points of view.

HINT /2/



Types of syllables:

(1) short stressed syllables (Q1);
(2) long stressed syllables (Q2);
(3) long stressed syllables with an extra quantity (tense pronunciation, Q3);
(4) short unstressed syllables;
(5) long unstressed syllables.

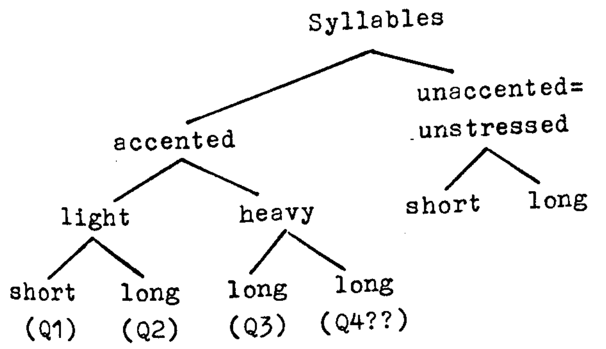
Examples:

1 4 4 2 5 4 3 5
kalale minnakse 'õhtul
'to fishing one goes in the evening'

2 5 3 5 2 5
karjus 'karjus metsas
'shepard shouted in the wood'

In this scheme the phonological stress and quantity are treated as two separate prosodic phenomena, the phonemic stress being a precondition for quantity distinction in long syllables /2/; +stress may be either a main or a secondary stress (this additional branching does not affect the system of quantity contrasts).

VIITSO /4/



Syllable types:

(1) short light-accented (Q1) syllables;
(2) long light-accented (Q2) syllables;
(3) long heavy-accented (Q3) syllables;
(4) long extra heavy accented (Q4) syl.;
(5) short unaccented syllables;
(6) long unaccented syllables.

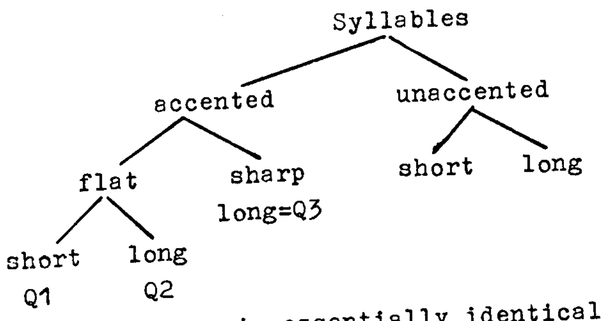
In this scheme stress and quantity are incorporated into a unique prosodic complex - accent. There appear to be some inherent difficulties in this scheme:

(1) all the stressed syllables are accented, but if there is a difference between the main and secondary stress, and if both the accents and stresses are unpredictable within the word, then it follows that the number of accents should be doubled according to the number of stress degrees (main and secondary stress);

(2) long syllables need for their three different accents tertiary branching - if the first differentiation in this scheme were between short and long syllables, then the long syllables would clearly need tertiary branching;

(3) Q4 has been suggested by Tiit-Rein Viitso for several years but it has not been proved experimentally (descriptive inadequacy); it seems that this doubtful quantity (accent) degree does not fit into an ordinary prosodic scheme, either; without Q4 the scheme would look much more plausible.

EEK & HELP /1/



This scheme is essentially identical with the Viitso's analysis, except the terminology, and Q4 which Eek and Help have abandoned as unsubstantiated. During many decades this scheme has been suggested by Valter Tauli (whose terms were light and heavy stress, cf. /3/).

The comparison of these conceptions underlines the following pecularities of Estonian prosodic system:

(1) there are both short and long syllables with light accent (lax pronunciation); this is the main point in the schemes by Valter Tauli, Tiit-Rein Viitso, and Arvo Eek & Toomas Help; in Hint's conception these syllables are considered to be unmarked in respect of syllabic quantity;

(2) short syllables do not participate in quantity contrasts; this is most distinctly revealed in Hint's scheme;

(3) it is possible to interpret the Estonian prosody as having only one accent or extra syllabic quantity (Q3); this is best revealed in the scheme by Arvo Eek and Toomas Help; in Hint's conception this is expressed by specially marked +extra quantity;

(4) extra syllabic quantity is possible only in long stressed syllables; this is clearly pronounced in Hint's conception.

PALATALIZATION

Palatalization in Estonian is a phonological correlation (in Trubetzkoy's terminology) of limited positional occurence. Its realization in different Estonian dialects brings forth the different aspects of its phonological nature.

The palatalization in Estonian is characterized by the following:

(1) the list of palatalizing consonants varies greately in different dialects: in South Estonian dialects /ṕ ḿ t′ ń ś l′ ŕ ḱ/ may be palatalized; in North Estonian dialects palatalization occurs only in dental consonants; Standard Estonian palatalizes /t′ ś l′ ń/; there is no palatalization in the Northern Costal dialect;

(2) the pattern of palatalization before /-i/ or /-j/ differs in various dialects: in the Mulgi dialect (South Estonia) and in the Islands' dialect there is no palatalization before an overt /-i/ or /-j/; in other dialects there is an automatic palatalization before /-i/ and /-j/.

These differences cause great variations in the functional load of palatalization in different dialects. At the same time, the palatalization or non-palatalization before /-i/ and /-j/ is an overt reflection of various phonemicizations of palatalization, that is, whether in a position before /-i/ or /-j/ the phonetic palatalization represents a palatalized or non-palatalized phoneme.

It is easy to see the morphophonemic consequences of one or another interpretation. Compare, for example, the pattern of palatalization in the word kast 'box'.

	In Standard Estonian		In Islands' dialect	
Nom. sg.	/'kaśt/	+pal	/'kaśt/	+pal
Gen. sg.	/kaśti/	+pal	/kasti/	-pal
Part. sg.	/'kaśti/	+pal	/'kasti/	-pal
Part. pl.	/'kaśte/	+pal	/'kaśte/	+pal

The palatalization in Estonian deserves attention for its low functional load. The following table illustrates the percentage of palatalized consonants in the only position where distinctive palatalization occurs - in the position after a nucleus of main-stressed (first) syllable (where both single consonants and the first components of consonant clusters may be palatalized: /'klaaśe/, /'lol′le/, /'kaśte/).

The data are based on a statistically reliable sample of literary texts (total of 14.563 words: Q1 - 4.249, Q2 - 2.898, Q3 - 7.416).

In the table +pal max stands for maximum count of palatalization, that is, palatalized segments are interpreted before /-i/ and /-j/ and elsewhere as realizations of palatalized consonants;

+pal min indicates minimum count of palatalization, that is, automatic palatalization before /-i/ and /-j/ is interpreted as realization of non-palatalized consonants;

-pal min presents percentage of non-palatalized counterparts of this phonological correlation.

In the table only +pal min represents distinctive palatalization; its rate in Q1 and Q2 words is practically zero.

Palatalization percentage

		/t/	/s/	/n/	/l/	Σ
Q1	+pal max	0.6	3.4	2.2	7.0	13.2
	-pal min	8.0	3.4	8.8	14.5	34.7
Q2	+pal max	1.9	2.1	2.1	5.5	11.6
	-pal min	11.3	6.2	13.2	16.6	47.3
Q3	+pal max	1.0	2.0	1.4	2.0	6.4
	+pal min	.28	.46	.14	.27	1.15
	-pal min	17.1	13.9	10.2	10.0	51.2

Both ways of counting may be of interest for the low reading of palatalization. In spite of this there is no tendency to eliminate the palatalized consonants from the phonemic inventory of Estonian. In the lexical system the palatalized consonants obviously have more pronounced role (contrasts such as tall 'lamb' and tal′l 'stable', kott 'large shoe' and kot′t 'sack').

REFERENCES

/1/ Arvo Eek, Toomas Help, "Rütminihked eesti keele kujunemisloos". Tallinn 1986.

/2/ Mati Hint, Viron prosodisen systeemin perusluonteesta. - "Virittäjä" 1986, p. 428-440.

/3/ Valter Tauli, "Standard Estonian Grammar". Part I. Uppsala 1973.

/4/ Tiit-Rein Viitso, "Läänemeresoome fonoloogia küsimusi". Tallinn 1981.

THE ORGANIZATION OF A PHONETIC WORD AND SENTENCE PROSODY IN BIBLICAL HEBREW

ALEXANDRA YU. AIKHENVALD

Institute of Oriental Studies,
Moscow, USSR

The main principles of a phonetic word organization in Biblical Hebrew are discussed, with rules for vowel changes as to the place of the stress formulated. A basic analogy in the structure of a phonetic word and that of a sentence is postulated, for specific accent properties alongside with a special vowel change paradygm dependent on different positions as to the word-stress vs sentence prosody are characteristic for both.

I.I. There exists more or less general agreement about the importance of analyzing the pronounciational structure of a text and its constituents, i.e. sentences, tacts, or syntagms, phonetic words etc, see for discussion /1/. Nevertheless , the rules operqting within such units as phonetic words and concerning their pronounciational structure have remained rather a terra incognita for an overwhelming majority of linguistic descriptions (thus, very few attempts,if any,have been so far made to propose a calculus of phonetic words possible in this or that language). The present paper is concerned with an attempt to formulate the rules of the organization of phonetic words (see 2.1.,2.2.), to propose the rules for constituting bigger pronounciational units out of phonetic words,with their specific prosody, i.e. accent characteristics,and to discuss specific prosodic patterns of a sentence, with the rules for distribution of phonetic words and/or syntagms according to their position within a sentence (see 3.), all on the material of Biblical Hebrew (hence BH). Discussing the organization of a phonetic word, we'll offer a classification of BH morphological units as to their "constructive class" (see 2.1., /2/). To describe the structure of a phonetic word and to classify specific positions within a sentence and/or a phonetic word as to its prosody and the type of vowel changes, it appears necessary to provide a fragment of BH morphonemics (or archiphonemics, see 2.2.,3). The main conclusions of our study which might be of a certain interest for future Hebrew studies as well as for studies in the field of typology of sentence prosody are presented in the last section (see 4).

I.2. The BH material is of a great interest for the analysis of sentence prosody and other pronounciational characteristics of the text, for the texts in BH are not only supplied with vocalization marks, which is a rare thing for a text in an ancient Semitic language, but with accent marks as well. The BH distinguishes two systems of accents - poetical ones (as in Books of Job, Psalms, Proverbs) and prosaic ones (as in other Books of the Bible) (see for details /3/, /4/). Our study will cover the system of prosaic accents only, basing on the text of the Tanah in Tiberian vocalization (early X cent. A.D.) without taking into consideration minor problems concerning specific vocalization marks (f.ex., dages, swa medium etc) and other slight inconsistencies within the text of the Bible, for on the whole it appears obvious that BH possesses a common system of rules for phonological, prosodic (see just above) and morphonological organization.

2.1. To study the structure of phonetic words, it is necessary to classify the units of the language into constructive classes, according to their behavior as as to accentual independence and to phonological processes operating on the inter-unit boundary (a common stress being held for a main parameter to distinguish a separate phonetic word). In BH, we distinguish three constructive classes of units and three types of inter - unit boundaries respectively: I. Bases, i.e. the only units capable of constituting an independent phonetic word all alone, which fall into accented ones (here belong a great many lexical units, such as dāḇā́r 'word', šāmár 'he kept')and unaccented ones (here belong some prepositions as ʿal 'upon', taḥat 'under', et alia, and some adverbs, as gam ' also'); this opposition is relevant but for a syntagmatic level (the interunit boundary for bases will be marked with ##). 2. Affixes, falling into declensional(or desinential, such as noun pl.masc. -im, fem.-ot) and word-formational (as verbal stems affixes(prefixes), noun suffixes -on,-an etc) The affix boundary is marked with + (for lack of space we'll not discuss here the opposition of prefixes, suffixes and transfixes in BH). 3. Clitics, falling into proclitics (as ha- (definite article ,wə- 'and', še -'that,which', prepositions- 'in', lə-'to', kə-'like',min 'from'(traditionally denominated: prefixed prepositions) and enclitics (here belong direct object pronouns used with the verb), the clitic boundary is marked with # .

Each type of boundary is characterized by specific processes in pperation. The peculiarity of a clitical boundary is that another phonetic word may be inserted into the phonetic word given only before it. Unlike bases, clitics and affixes can not alone form a phonetic word. Within a phonetic word, both affixes and clitics are ascribed a rang as to their place. Thus, wordformational affixes tend to be placed nearer to the base than desinences. In a set of clitic, each clitic has its specific rang, cp. an admissible sequence of proclitics, with the number of the rang as to the base given in brackets: wə(1)še(2)lā(3,4)ʿir from wə(1)-še(2) lə (3)ha(4)ʿir ' and-which-to-the-town...',as opposed to a wrong sequence *šewələhaʿir (the process ləha > lā is an example of a specific process operating on a clitic boundary #) (comp. the rangs for clitics in clitic complexes in Hittite and other Anatolian languages, as well as in Berber, Cushitic etc).The specific rules regulate the compatibility of different types of clitics, affixes and bases with one another; a more tailed discussion of the problem lies outside the frame of the present paper. An ideal phonetic word admissible would be: Cl(I)Cl(2)Cl(3)Cl(4)Aff(I)Aff(2)B-Aff(I)Aff(2)Cl, where Cl stands for a position reserved for clitics only, Aff- for that of an affix, B- for that of a Base, figures in brackets stand for rangs. Strangely enough, a normal phonetic word can consist also of CL # +Aff (a phonetic word of such a structure behaves as an unaccentod base), f.ex. bi(bə + -i)'in me'. So, we may use the formula for 'allocations with repetitions' A_2^{10} and obtain the number of phonetic words possible which equals 2^{10} . Naturally, not every sequence possible is allowed, because of the restrictions on compatibility between separate units (see above).Here are some examples of allowed phonetic words: 1/ B # CL: šəmārḥā́ 'he preserved her', 2/ B+Aff :šāmər+ā 'she preserved', (cp.different phonological processes of vowel change on the # and + boundaries!),3/ CL# CL# CL# B+Aff:wəšebəsifr+enu 'and-which -in-book+our'.

2.2. For the further study in the prosody of phonetic words,it would be necessary to study the main stress patterns proper to it. In general, the main stress would tend towards the end of a phonetic word.When the stress place is changed, there occur phonologically conditioned vowel changes. These changes can be given in a special paradygm, called an archiphonemic paradygm. Let's call ' an archiphoneme' an abstract unit to be interpreted with any phoneme (in the sense of Prague School 'soundtype') belonging to set of phonemes distributed according to purely phonological context. The units (i.e.archiphonemes) with a common set of phonological rules (or a common context) are united into one archiphonemic paradygm. So, for archiphonemic paradygm of BH vowels the position as to the stress (see below) together with openness of the syllable is the main 'context-forming'feature. One more context -forming feature is the type of inter-unit boundary (see 2.1.,I/,2/ for different vowel changes on + boundary and # boundary).(One should remark in brackets that distinguishing between phonologically conditioned vowel changes and morphologically conditioned ones provides us with a most powerful tool for the description of morphonology and morphology of BH, as well as that of Modern Hebrew, allowing,f.ex.,to reduce the number of noun declensional classes to 4 from about 340 and verbal ones to 2 from about 14). Most phonetic words consisting of a B only are stressed on the last syllable, except for words in(-)CVCeC,V being ē,e, o, obtained by phonological rules from (-)CVCC and similar, like sēp̄er <*sip̄r, meleḵ < *malk etc, some aramaic loans (as lāmmā 'why') and a few real exceptions (as láylā 'night', kodkōd 'skull'). If a phonetic word contains a clitical and/ or an affix boundary before the base, it does not affect the stress and thus the vowels (with an only exception being the verbal declensional prefix wa-'waw consecutivum'). If an affix boundary lies after the Base, following situations can occur: a) affix belongs to 'unstressed' ones, as,f.ex., noun locative -a, verbal I Sg Perf.-ti, 1 Pl Perf.-nu etc; then no changes occur; b) affix belongs to 'stressed'ones, then the stress is moved to the affix, as with affixes like masc. Pl.-īm, fem.Pl. -ṓt,3 Sg Fem Perf -ā́ etc.

(The information as to whether an affix is 'stressed' or 'unstressed' is due to a special morphological dictionary.) When the stressed is transferred to the suffix, vowel changes according to an Apchiphonemic paradygm take place (let's call it AP I).For lack of space, the whole AP I will not be presented here. We'll but bring some examples: $\bar{A}_I$ (an archiphoneme): $A_I \to \bar{a}$ in an open syllable immediately preceding the stressed one, $A_I \to \emptyset$ in an open syllable not immediately before the stress, applying the rules to an AP version DA_IBA_IR of dābār 'word' + pl.-īm, we obtain a coprect form dəbārīm, ə being an automatic vowel; $A_o \to \emptyset$ in an open syllable immediately before the stressed one, $\bar{A}_I \to \bar{a}$ in a closed syllable not immediately before the stressed one, so from $\check{S}A_IMA_oR$ + -ā́ we get a correct form šāmərā́ (see 2/), with an automatic ə .

(The AP rules are to be applied beginning from the stressed syllable).
If a clitical boundary lies after a B, or B+Aff, the stress is removed to the clitic , and the vowel change operating come from a different AP paradygm, thus(let's call it an AP 2), $\bar{A}_I \to \emptyset$ in a syllable not immediately before the stressed one, and $A_o \to \bar{a}$ in a syllable immediately before the stressed one, and according to these rules we get a correct form out of $\check{S}\bar{A}_IMA_cR$ # -ā: šəmār#ā́ (see above, ex. I/). Therefore, it is absolutely necessary to distinguish between two APs for a phonetic word.
3. Now let's pass to the analysis of the structure of bigger units, i.e. syntagms, or tacts, and the sentence prosody properties. A syntagm may be equal to a phonetic word, or excede it. Within the frame of a sentence (in BH, the end of a sentence is usually marked with :) several positions can be identified,and if we regard a sentence as an accentual or prosodical unit, these positions can be treated as analogous to the positions as to the stress place within a phonetic word. These positions in a sentence are: an unaccented position, a strongly stressed (or a 'pausal') position and a normally stressed (or a 'non-pausal')one. The opposition of s.c. pausal and non-pausal forms in BH has been known since the earliest descriptions of BH (see /3/,/4/), but no evaluation of a pausal position (i.e., a position, for which a pausal form is required) has ever been proposed. Every position in a sentence is characterized by a specific set of accent marks used to identify it. So, a phonetic word within a syntagm occupying a normally stressed positions has no special accents, but for a secondary stress (the accent mark meteg) on every closed syllable with a long vowel, as in battim 'houses', the secondary stress not affecting any AP. The phonetic word within a syntagm in the position under consideration may combine with so called 'weak disjunctive accents' marking the role of this or that constituent in the logical organization of the sentence (here belong accents as zakkef, gereš and some other). If a phonetic word occurs in an unstressed position, it is automatically united with another phonetic word or a syntagm into a new syntagm. Moreover, an unaccented base (see above) may not constitute a separate syntagm. The unstressed position is marked by the s.c. ' conjunctive accents', lying on the second(i.e. stressed) constituent of a syntagm, the graphic marker of the unstressed position being also a horisontal line -linea makkef between the constituents. A secondary stress may appear on the constituent in an unstressed position, unless it is an unaccented base. F.ex., 3/ Gen.I,5 wayəhī--ʿereb wayəhī-boker ' evening came and morning came' (where - stands for linea makkef, $\underset{,}{x}$ is meteg on a phonetic word in an unstressed position, x - a conjunctive accent merha marking an unstressed position for yəhī 'be, was').

A phonetic word and/or a syntagm stands in a strongly stressed (pausal) position before a pause, i.e. in the very end of a sentence and/or in the end of a logically complete passage. S.c. 'strong disjunctive accents' (atnāḥ and sillūk) are used to identify the position in question. The most interesting property of BH from the point of view of archiphonemics consists in the fact that there are independent APs with specific vowel change paradygm for each position.. Thus, in a strongly stressed position no vowel changes occur and the stress is never removed, whatever structure a phonetic word may possess, and another vocalism is characteristic of it in comparison with other positions. Cp. following examples of syntagms in strongly-stressed positions:4/ Kings II, 11, 14: wattikraʿ ʿătāləyā 'et-deḥā wattikrā' kešer kašer 'Athalia tore her garments and shouted:" Treason,treason!' (X -sillūk, kaser-a specific form used in a strongly stressed position of the word keser); 5/ Jer. 22,29: 'erec 'erec 'ārec šimʿi dəbar-yhw ' O land, land,land! Listen to the word of Yahwe!' ($\hat{X}$ - atnāḥ, 'ārec- a strongly stressed position form of erec); compare the different vowel patterns in : pausal:'āmártī, nonpausal;'āmártī 'I said','āmārū vs'āmərū ' they said'.

The unstressed position also possesses a specific vowel pattern; its AP is close to the AP I of a normally stressed position. The only complication about the vowel patterns occurring in unstressed positions is that they partly coincide with the morphenemes set of noun declension, for there are specific morphological forms of nouns(those of construct state) which are found only in this position.
Thus, the position of a phonetic word in a sentence (and/or that of a syntagm) appears to constitute one more 'context-forming ' (in the above sense) parameter for an AP. In this respect the prosodic organization of a sentence in BH is analogous to that of a phonetic word, compare also the analogy of the operation of a secondary stress within a syntagm and/or a phonetic word and that of the s.c. 'weak disjunctive accents'in a sentence. A specific archiphonemic paradygm is characteristic for different positions in a phonetic word as to the stress, as well as for different positions of a phonetic word in a sentence; in both cases the 'strongly stressed' is the word vs sentence final position.
4. By the way of analyzing the principles of a phonetic word organization and its functions within syntagms and sentences we have arrived at a conclusion of a basic analogy between the structure of a phonetic word and that of a sentence at least on the archiphonemic level of presentation. This analogy may be of some interest not only for Hebrew studies but perhaps for historical and typological studies of sentence prosody as well. The analogy in question reminds us of an analogy in the structure of hierarchically regulated units of different kind -i.e., phonemes, morphemes, words etc - drawn by some linguists. Can a structural analogy between the hierarchically regulated units of a different type - those connected with a'linear', or pronunciational organization of the language, and not with the 'paradygmatic' organization of the language in the sense of /I/ - be maintained ,too ? (Angway, it might be interesting to analyze from this point of view other laws and patterns of pronounciation organization of a sentence, as well as the rules for placing phonetic words of different types within them; consider, f.cx., the rule for placing clitics in a specially reserved position, usually a second position in the sentence in many Indo-european and Afro-asiatic languages; the ban for unstressed bases to occupy a sentence-final position in Modern Hebrew and so on).

References.

/I/. I.Boduen de Courtenay. Selected Works, v. 2, Moscow; p.249, 1963.
/2/. A.Barulin. Theoretical problems of the Turkish nominal word-form. PhD, Moscow, 1984.
/3/ G.Bergsträsser. Hebräische Grammatik. Leipzig, 1918.
/4/ J.Kurylowicz. Studies in Semitic Grammar and Metrics. Wroclaw, 1972.

All Bible translations cited from:
The Jerusalem Bible. Jerusalem, 1974.

ДИНАМИКА АТТРАКЦИИ УПРАВЛЕНИЯ

(на материале литовских диалектов)

ПАБРЕЖА ЮОЗАС

Шяуляйский пединститут им К. Прейкшаса
Кафедра литовского языка
235400 Лит. ССР, Шяуляй

Резюме

На значительной территории жятайтского диалекта аттракция ударения представляет собой новое явление. Однако зародыши этого процесса могут быть очень древними и восходить к периоду интенсивных куршско-жямайтских контактов. Куршкий язык мог дать первычный импулс для аттракции ударения, однако, если и принять "куршскую" гипотезу, то все же следует признать, что аттракция ударения в северожемайтском наречии более просто и убедительно объясняется не языковыми контактами, а внутриязыковыми мотивами, среди которых главными являются редукция безударных гласных, определенные морфологические факторы, а также характер фразовой интонации.

Одной из важнейших черт просодии северожемайтского наречия является аттракция ударения - его оттяжка с конечного и долгого циркумфлектированного слога, напр.: gèrà ← gerà "хорошая", vàkã ← vaikaĩ "дети". Большинство литовских диалектологов считает, что аттракция представляет собой совершенно регулярное и законченное явление, не допускающее почти никаких исключений. Однако при более близком рассмотрении оказывается, что данная закономерность на уровне спонтанной, живой речи не является абсолютной, т.е. аттракция ударения в современном северожемайтском наречии представляет собой не статическое (законченное), а динамическое явление. В настоящем исследовании рассматриваются именно такие случаи колебания аттракции ударения, возможные причины этого явления. Диахронические гипотезы проверяются точными статистическими методами на основе систематического анализа магнитофонных записей спонтанной связной речи. Все статистические данные обработаны с помощью электронно-вычислительной машины.

Полученные результаты показывают, что аттракция ударения и её отсутствие в северожемайтском наречии подчиняется довольно четким внутренним закономерностям. Прежде всего, она зависит от определенных фонетических факторов: с долгих окончаний ударение не оттягивается значительно чаще, чем с кратких. Во вторых, сильное влияние на аттракцию ударения или её отсутствие оказывает морфологический фактор, который можно определить как тенденцию к колумнальной акцентуации, напр.: результаты аттракции ударения в местоимениях существенно отличаются от его оттяжки в других именных словах, поскольку многие местоимения являются целиком окситоническими по своей природе; глагольные формы 3-го лица будущего времени исключительно часто сохраняют конечное ударение, несомненно, под воздействием I-го и 2-го лица (ср.: ušpìksọ, ušpìksi и ušpìks "рассержусь, рассердишься, рассердится"); неизменяемые слова более часто сохраняют конечное ударение, чем изменяемые и т.д.

Большое (и даже, может быть, решающее) значение для генезиса и развития аттракции имели интонационные факторы. Ударение чаще всего не оттягивается в том случае, когда слово имеет логический акцент, и особенно, если оно произносится с восходящей интонацией, напр.: ka‿jau‿mẹrẹ marẽle, e‿papũsk i‿rurẽlę "когда уж умерла Маряле, и ничего тут не поделаешь". По - видимому, прежде всего аттракция ударения началась в слабых позициях фразы, т.е. в словах, лишенных логического акцента; затем этот процесс охватил и окончания акцентированных окситонических словоформ, если они не имели определенного дополнительного стилистического оттенка (например, сильной эмфазы). Наконец, окситоническое ударение с течением времени постепенно теряет даже стилистическую функцию и вытесняется из всех позиций. Следы этого процесса видны и в современных северожемайтских говорах: в южной части сохраненное окситоническое ударение является сигналом простого логического акцента, в северожемайтских говорах средней полосы - сигналом экспрессивного (эмфатического) логического акцента, а в северной части всякая закономерность практически теряется - окситоническое ударение появляется лишь как крайне редкое исключение. Связь между колебаниями ударения (аттракцией, переносом) и интонацией отмечена в латышском /I/, французском /2/, английском /3/, польском /4/, исландском /5/, корейском /6/ и во многих других языках мира.

В проведенном исследовании выявлена закономерная связь между аттракцией ударения и редукцией окончаний: по направлению с юга на север редукция постепенно усиливается, а степень аттракции также постепенно возрастает. Следовательно, можно утверждать, что аттракция ударения является своеобразным проявлением редукции, причем более архаичными являются те варианты словоформ, которые сохраняют конечное ударение и имеют отчетливо произносимое окончание (промежуточное положение занимают варианты с колумнальной неконечной акцентуацией). С усилением редукции неокситонические варианты закрепляются и становятся абсолютно преобладающими. Аттракцию ударения с редукцией связывают также интонационные факторы более общего характера. В одном и том же отрезке текста ударение весьма часто не оттягивается в том случае, когда слово находится под логическим акцентом, а отчетливо произносимое (т.е. нередуцированное) окончание тоже связано с более экспрессивным интонированием фразы, напр.: tik‿žaltẽ↑ mọn‿niẽrvu↓ negadinkẹt "только, мерзавцы, мне на нервах не играйте", dabãr↑ jau‿vẹrš‿šim̃ta↑ mẽtu↓ ir "теперь уже сто с лишним лет". Как видно, в одном и том же высказывании одни эмоционально окрашенные слова часто сохраняют окситоническое ударение (žaltẽ, dabãr), другие же сохраняют нередуцированное окончание (niẽrvu, šim̃ta, mẽtu).

Сам процесс аттракции ударения, очевидно, начался в северной (точнее - в северо - западной) части диалекта и постепенно распространялся в южном направлении. Волна аттракции продвигается к югу и в настоящее время. Во времена К. Яуноса (1848-1908) в его родном кведарнском городе аттракции ударения не было (за исключением так называемой димоции /7/); теперь же этот говор уже имеет полную (правда, сильно колеблющуюся) аттракцию. Следовательно, на самом деле аттракция ударения во многих местах представляет собою новое явление.

Аттракция ударения не только распространяется географически, но и становится все более последовательной. Любопытно, что почти во всех обследованных нами населенных пунктах северожемайтского наречия ударение наиболее регулярно оттягивают не самые старые, а, наоборот, наиболее молодые информанты. Например, в речи представителя жемальского говора, родившегося в 1887 г., ударение на конце слова сохраняется в 10,3% всех окситонических форм, а в речи более молодого информанта (1906 г. рожд.)- 5,7%. В речи одного из старейших жителей Вежайчай (1890 г. рожд.) словоформ с сохраненным конечным ударением встретилось 24,9%, а в речи представительницы этого же говора следующего поколения (1915 г. рожд.)- - только 12,4%. Замечено также, что и редукция окончаний в речи более молодых северных жемайтов выражена значительно сильнее, чем у престарелых информантов, а это, в свою очередь, может способствовать более сильной аттракции ударения. По-видимому, усиление аттракции в настоящее время стимулируется гиперкоррекцией "снизу", возникшей под влиянием литературного языка. Следовательно, происходит своеобразная поляризация /8/ диалекта по отношению к литературному языку.

Определить момент начала процесса аттракции ударения по имеющимся данным трудно. Мнение лингвистов по этому вопросу сильно расходятся: одни аттракцию ударения считают древним явлением /9/, другие - совершенно новым /10/. Как показывает наше исследование, на значительной территории жемайтского диалекта аттракция ударения, вне всякого сомнения, представляет собой новое явление. Однако зародыши этого процесса в самом центре его возникновения могут быть очень древними и восходить к периоду интенсивных куршско-жемайтских контактов. Куршский язык мог дать первичный импульс для аттракции ударения, однако его влияние вряд ли было таким прямолинейным, каким его обычно представляют. Здесь важное значение могли иметь и различные косвенные влияния. Например, вполне обоснованно считается, что курши (как пруссы, западные литовцы и латыши) сокращали конец слова /II/. Поэтому представляется совсем реальным, что у них могла быть позаимствована тенденция к редукции конца слова, которая, в свою очередь, могла вызвать развитие аттракции ударения. Также возможно, что у куршей были позаимствованы определенные интонационные модели: яркий контраст акцентированных-неакцентированных слов мог оказать влияние как на редукцию, так и на аттракцию ударения. С куршским влиянием могла быть связана и резкая каденция конца фразы, которая способствовала редукции окончаний и аттракции ударения в конечных словах фразы. Как показало наше исследование в северожемайтских говорах слова с окситонической акцентуацией крайне редко встречаются в абсолютном конце предложения. Например, в Барстичяй (один из населенных пунктов территории северожемайтского наречия) из 107 слов с сохраненным конечным ударением только 8 (или 7,6%) находились в конце предложения (напр.: mõna vàks nebûs mergáutûo pavardiẽ "мой ребёнок не будет носить девичью фамилию").

Внешние языковые контакты усматриваются и в тенденциях упрощения, "нормализации" морфонологии, в некоторых выравниваниях по аналогии, характерных для западных говоров северожемайтского наречия. Предполагается, что обобщение перехода конечных i, u в e, o (т.е. упразднение ассимиляции гласных) в этих говорах как раз и обусловили интенсивные языковые контакты с такими диалектами, для которых явление ассимиляции должно быть чуждым. Наконец, куршский субстрат мог вызвать стремление к упразднению аффрикат, т.е. в функциональном отношении нецелесообразного чередования согласных t':č'- d':dž' в западных говорах. Любопытно, что исчезновение этого чередования, как и аттракция ударения, всё более распространяется /12/. Имея в виду эти факты вполне обоснованно можно предполагать, что и "агрессивность" морфологических факторов, тенденция к колумнальной акцентуации могут объясняться теми же причинами – упрощением морфонологической системы, связанным с языковыми контактами.

Следовательно, прямолинейно отбросить влияние куршского субстрата на аттракцию ударения никоим образом нельзя, однако, если и принять "куршскую" гипотезу, то всё же следует признать, что аттракция ударения в северожемайтском наречии, особенно сам ее процесс и географическое распространение, более просто и убедительно объясняются не языковыми контактами, а внутриязыковыми мотивами, среди которых главными являются редукция безударных гласных, определенные морфологические факторы (тенденция к колумнальной акцентуации и т.п.), а также характер фразовой интонации/13/.

/1/ Rudzīte M. Latviešu dialektologija, Rīga, 1964.
/2/ Leon P. Patrons expressifs de l'intonation. - Acta Univ. Carolinae. Philologica I. Phonetica Pragensia III, 1972.
/3/ Bolinger D. L. Accent Predictable.- Language, 1972, vol. 48, N 3.
/4/ Dogil G. Autosegmental Phonology in Contrastive Linguistics. - In: Theoretical Issues in Contrastive Linguistics, Amsterdam, 1980.
/5/ Benediktsson H. The Non-uniqueness of Phonemic Solutions. - Phonetica, 1963, vol. 10, IV 3-4.
/6/ Поливанов Е. Д. Статьи по общему языкознанию, Москва, 1969.
/7/ Jaunius K. Dialektologiniai darbai, Kaunas, 1891-1900.
/8/ Girdenis A. Žemaičių dzūkavimas: dabartinė padėtis ir istorija. - Baltistica, 1980, t. 16(1).
/9/ Grinaveckis V. Dėl lietuvių kalbos tarmių kirčio-atitraukimo kilmės.-LTSR MA darbai. Ser. A, 1977, t. 4(61).
/10/ Girdenis A., Rosinas A. Keletas samprotavimų dialektologinės fonetikos klausimais. - Baltistica, 1976, t. 12(2).
/11/ Girdenis A. Kuršių substrato problema šiaurės žemaičių teritorijoje.- Kn.: Iš lietuvių etnogenezės. Vilnius, 1981.
/12/ Girdenis A. Baltiškųjų tj, dj refleksai 1759 m. "Žyvate". - Baltistica, 1972, t. 8 (2).
/13/ Pabrėža J. Diachroninės pastabos apie kirčio atitraukimą šiaurės žemaičių tarmėje. - Baltistica, 1986, t. 22(2).

О НЕКОТОРЫХ ФУНКЦИЯХ РУССКОГО СЛОВЕСНОГО УДАРЕНИЯ

ЕЛЕНА ЯСОВА

Маитолантие 8
54800 Савитаипале
Финляндия

ВВЕДЕНИЕ

Целью этой статьи является рассматривание функции словесного ударения в омографах и в акцентологических вариантах, являющихся неотъемлемой частью системы современного русского языка.Ударение, как просодический признак русского слова, по существу реализуется в звучащей речи и, как правило, в письменных текстах не обозначается. Впрочем, при звуковом оформлении русских письменных текстов нерусские в первую очередь сталкиваются с трудностями, связанными некоторым образом с постановкой ударения, от которой зависит весь звуковой образ слова, в том числе и его ритмика.

Известно, что ритмическая организация русского слова связана с сильноцентрализующим типом квантитативно-динамического разноместного и подвижного ударения и с редукцией гласных по силе и длительности двух степеней в безударных слогах. Итак, в основе ритмики русского слова лежит контраст ударности и безударности слогов и ритмическая структура в общем определяется количеством слогов и местом ударения. Вследствие разноместности и подвижности словесного ударения существует в русском языке и значительное количество омографов. И с развитием языка тесно связано и развитие нормы акцентуации, вызывающее сосуществование акцентологических вариантов в пределах норм современного литературного языка. Так например, развитие акцентуационной нормы иллюстративно можно показать у существительного "душа́". В Грамматике русского языка 1952 г. /1/ нормативная рекомендация ударения в дат., пред. и твор. пад. мн. ч. на окончании, т.е. "душа́м - душа́х - душа́ми". В примечании допускаются и "более новые" формы - "ду́шам - ду́шах - ду́шами" с ударением на основе. В словаре-справочнике Аванесов и Ожегов 1959 г. /2/ уже кодифицирована акцентуация существительного "душа́" на основе. И более старая форма акцентуации указана только в фразеологизме "говорить по душа́м". Акцентуация на основе существительного "душа́" рекомендуется и в новой Академической грамматике 1970 г. /3/.

ФУНКЦИИ УДАРЕНИЯ

Русское словесное ударение, как суперсегментное свойство слова, в силу своей разноместности и подвижности выполняет и некоторые функции. Так например, с точки зрения нерусского чтеца русских художественных текстов, как в плане выражения, так и в плане содержания, являются важными следующие функции словесного ударения:

1/ функция организации или построения ритмических структур /моделей или просодем/. Ясно, что степени редукции безударных гласных зависят от места ударения в слове, это значит, что ударение организует весь его звуковой образ. Ср., напр., золото = – ◡ ◡ [зОлътъ], болото = ◡ – ◡ [бʌлОтъ], полотно = ◡ ◡ – [пълʌтнО] и т. п.

2/ функция идентификации или определения, распознавания слова. Явно, что каждое слово отличается определенной ритмической структурой и ее нарушение приводит к неправильной ритмической реализации, что во всяком случае затрудняет идентификацию слова. Ср., напр., золото ≠ ◡ – ◡ [зʌлОтъ], болото ≠ – ◡ ◡ [бОлътъ], полотно ≠ ◡ – ◡ [пʌлОтнъ] и т.п.

3/ функция дифференциации или различения, проявляющаяся в разных планах языка, указывается в омографах и акцентологических вариантах. Таким образом, ударение может выполнять функцию дифференциации одновременно в разных планах языка: а/ план ритмико-звуковой и семантический. Напр., атлас = атлас – фонемная идентификация, но: – ◡ ≠ ◡ – [Атлъс] ≠ [ʌтлАс] и: сборник географических карт ≠ сорт ткани – ритмико-звуковая и семантическая дифференциация или среду = среду, но: – ◡ ≠ ◡ – [ср'Эду] ≠ [ср'идУ] и: третий день недели ≠ окружение, обстановка и т.п.

б/ план ритмико-звуковой и лексико-стилистический. Напр., молодец = молодец – фонемная идентификация, но: ◡ ◡ – ≠ – ◡ ◡ [мълʌд'Эц] ≠ [мОлъд'иц] и: нейтральный стиль ≠ народно-поэтическая стилистическая окраска или компас = компас, но: – ◡ ≠ ◡ – [кОмпъс] ≠ [кʌмпАс] и: нейтральный стиль ≠ профессиональный стиль и т. п. Выбор того или иного акцентологического варианта зависит от характера текста. Так например, при чтении русских былин, чтобы создать народно-поэтическую стилистическую окраску, надо выбрать и реализовать акцентологические варианты с ударением на первом слоге: "мо́лодец, де́вица".

б/ план ритмико-звуковой и семантико-грамматический. Напр., страны = страны – фонемная идентификация, но: ◡ – ≠ – ◡ [стрʌнЫ] ≠ [стрАны] и: род. пад. ед. ч. ≠ им. пад. мн. ч. , т. е. ритмико-звуковая и грамматическая дифференциация или насыпать = насыпать, но: ◡ – ◡ ≠ ◡ ◡ – [нʌсЫпът'] ≠ [нъсыпАт'] и: совершенный вид ≠ несовершенный вид и т. п.

г/ план звуково-ритмический. Напр., гналось = гналось, но: ◡ – ≠ – ◡ [гнʌлОс'] ≠ [гнАлъс'] или творог = творог – фонемная идентификация, но : ◡ – ≠ – ◡ [твʌрОк] ≠ [твОрък] – звуково-ритмическая дифференциация. И в таких случаях существует идентификация не только в фонемном составе слова,но также в семантике, стилистике и грамматике.

Интересно, что в русском современном языке находится значительное количество равноценных в отношении к норме акцентологических вариантов, т. наз. дублетов. На основе словаря-справочника /4/ мы составили список акцентологических вариантов и обнаружили, что из общего числа 1 200 акцентологических вариантов 580 дублетов.

Дублеты чаще всего встречаются, напр. :

а/ в именах существительных в им. и род. пад. мн. ч. : шри́фты = шрифты́ – шри́фтов = шрифто́в; фло́ты = флоты́ – фло́тов = флото́в; фро́нты = фронты́ – фро́нтов = фронто́в и т. д.

б/ в именах прилагательных чаще всего в краткой форме ср. род. и мн. ч. : бе́ло = бело́ – бе́лы = белы́ и в прилагательных, выражающих пространственные отношения: дли́нно = длинно́ – дли́ны = длины́; высо́ко= высоко́ – высо́ки = высоки́ и т. д.

в/ в глаголах 2-го лица ед. ч.: кружи́шь = кру́жишь; горо́дишь = городи́шь; запру́дишь = запруди́шь. И больше всего дублетов встречается в возвратных глаголах в форме прошедшего времени ср. род. мн. ч. : назвало́сь = назва́лось – назва́лись = назвали́сь; пробрало́сь = пробра́лось – пробрали́сь = пробра́лись и т. п.

г/ в наречиях: высо́ко = высоко́; на́бело = набело́; а́ло = ало́ и т. д.

Если, например, в стихотворении встречаются дублеты, мы должны выбрать и произнести тот вариант, который требует заранее автором заданная метрическая схема. Иными словами /5/ метр стихотворения решает о выборе того или иного акцентологического варианта. Итак, метр является в некоторой степени практической опорой чтеца при соблюдении правильного ритма. Например, в стихотворении В. Солоухина "Яблоко" в отрывке во втором стихе встречается дублет: роди́лось = родило́сь. И в таком случае ямбическая метрическая схема данного стихотворения решает о выборе варианта: роди́лось. Ср., напр. :

То я́блоко – дитя́ Земли́ и Со́лнца
◡ – ◡ ◡ ◡ – ◡ – ◡ – ◡

Роди́лось,
◡ – ◡

Вы́росло из за́вязи
– ◡ ◡ ◡ – ◡ ◡

Созре́ло...
◡ – ◡

Таким образом, ритм является конструктивным фактором стихотворения и каждая неправильно построенная ритмическая модель повлечет за собой деформацию, искажение звукового образа слова и ритма вообще. И в смографах эта деформация может быть связана в определенной степени и с семантикой, стилистикой или грамматикой, несмотря на то, что контекст или ситуация содействуют правильному определению значения и понятия высказывания.

ЗАКЛЮЧЕНИЕ

Очевидно, что в памяти человека хранятся некоторые ритмические структуры и всвязи с этим надо уделять больше внимания в процессе обучения русскому языку, как иностранному, выработке правильной ритмики русского слова. И заучивание ритмических моделей, по нашему практическому опыту, полезно и на основе смены, контраста ритмических структур в смографах и акцентологических вариантах.

ЛИТЕРАТУРА

/1/ Грамматика русского языка, Москва, 1952.

/2/ Русское литературное произношение и ударение, словарь-справочник, под ред. Р. И. Аванесова и С. И. Ожегова, Москва, 1959.

/3/ Грамматика современного русского литературного языка, под ред. Н. Ю. Шведовой, Москва, 1970.

/4/ Трудности словоупотребления и варианты норм русского литературного языка, словарь-справочник, под ред. К. С. Горбачевича, Ленинград, 1973.

/5/ J. Rybák, Recitujeme po rusky, Bratislava, 1977.

RELATIVE IMPORTANCE OF ACOUSTIC FEATURES FOR PERCEPTION OF LITHUANIAN STRESS

ANTANAS PAKERYS

Dept. of Lithuanian Language
Vilnius State Pedagogical Institute
Vilnius, Lithuania, USSR, 232034

ABSTRACT

The relative importance of acoustic features for perception of word stress in Standard Lithuanian has been studied by method of artificial substitution using a computer.

Like in most languages, in Standard Lithuanian word stress is based on a number of acoustic features of sounds, i. e. duration (T), fundamental frequency (Fo), intensity (I), spectrum (S). Direct instrumental analysis of speech, however, is unable to reveal relative importance of the above features. In our opinion, the relative importance of acoustic features for perception of word stress may be effectively studied by method of artificial substitution. The method suggested may be defined as intersubstitution of acoustic features between members of an accentual opposition. Such modification of words makes it possible to reveal a certain competition of features.
To this aim the words kìtas ("other" nom. sing. masc.) and kitàs ("others", acc.pl. fem.) pronounced as statements by male speaker were fed into a BESM-6 computer via a digital converter (sampling frequency-50,000 cps). The prosodic features of both vowels in the word kìtas were modified according to the model of the word kitàs and vice versa. The features were substituted one by one, in pairs and all the three together. In addition, the vowels of kìtas were transferred to the word kitàs and vice versa. The variants of natural and modified words were recorded in random order and presented to 45 listeners. These were asked to find which of the two words (kìtas or kitàs) is heard and which of two intonations (statement or question) was used.

Table 1

Perception of stress and intonation (%, N=90). Variants of stimuli: 1 - (T,Fo,I, S/-), 2 - (Fo,I,S/T), 3 - (T,I,S/Fo), 4 - (T,Fo,S/I), 5 - (I,S/T,Fo), 6 - (Fo,S/T, I), 7 - (T,S/Fo,I), 8 - (S/T,Fo,I), 9 - (-/T,Fo,I,S), where in brackets unchanged features are presented before a slash (/) while the modified features are given after it. Statement is marked by a point (.) and question - (?).

Stimuli	Perception				Adequate stress perception
	kìtas		kitàs		
	.	?	.	?	
kìtas 1	97.8	2.2	-	-	100.0
kitàs 1	-	-	93.3	6.7	
kìtas 2	97.8	1.1	1.1	-	97.8
kitàs 2	1.1	2.2	92.2	4.4	
kìtas 3	6.7	78.9	10.0	4.4	86.7
kitàs 3	7.8	4.4	71.1	16.7	
kìtas 4	46.7	5.6	44.4	3.3	63.3
kitàs 4	6.7	18.9	50.0	24.4	
kìtas 5	11.1	44.4	36.7	7.8	44.4
kitàs 5	58.9	7.8	27.8	5.6	
kìtas 6	25.6	7.8	61.1	5.6	37.2
kitàs 6	1.1	57.8	27.8	13.3	
kìtas 7	-	18.9	57.8	23.3	11.1
kitàs 7	92.2	4.4	1.1	2.2	
kìtas 8	1.1	3.3	88.9	6.7	2.8
kitàs 8	91.1	7.8	1.1	-	
kìtas 9	1.1	1.1	93.3	4.4	1.1
kitàs 9	80.0	20.0	-	-	

The results of auditory experiments are presented in Table 1. The data obtained show that the feeding of the words into the computer followed by a reproduction do not distort the word stress: non-modified words (kìtas 1, kitàs 1) were perceived adequately. When the stressed and unstressed natural syllable nuclei of the quasi-homonyms were replaced paradigmatically, the perception of stress underwent a radical change:the listeners heard kitàs instead of kìtas and vice versa. Consequently, we believe the main carrier of information on word stress to be the syllable nucleus.
Having changed the acoustic features of vowels, the perception of stress varies to some degree.The percentage of auditory responses (i.e.in how many cases one or another feature is helpful in perceiving word stress) can be considered an indi-

cator of relative importance of those features. Thus, the decreasing sequence of separate features was found to be:

I	>	Fo	>	S	>	T
36.7%		13.3%		2.8%		2.2%

As in changing any separate feature the words were preceived adequately in more than 50% of cases, no individual feature can be considered as a relevant one since it is unable to rival the complex of other features. Much more effective are combinations of two features whose relative importance for perception of stress is as follows:

(Fo,I)>	(T,I)>	(T,Fo)>	(I,S)>	(Fo,S)>	(T,S)
88.9%	72.8%	55.6%	44.4%	37.2%	11.1%

Especially effective are complexes of three features:

(T,Fo,I)=	(Fo,I,S)>	(T,I,S)>	(T,Fo,S)
97.8%	97.8%	86.7%	63.3%

The data presented in Table 1 also show that stress perception is related to perception of intonation. When listening to the stimulus kìtas 3, for instance, most subjects basing on the non-modified features (T,I,S) heard the first syllable stressed while the contour of the fundamental frequency(Fo) transferred from the word kitàs was evaluated as the indicator of interrogative (rising) intonation.

The present experiment shows that it is combination of acoustic features with a different degree of relevance that makes a phonetic basis of Standard Lithuanian word stress. It has been also found that in the process of perception a distribution of features between word stress and intonation does take place.

THE TURKIC WORD PROSODY PROBLEM

A.DZHUNISBEKOV

The Institute of Linguistics of the Academy of Sciences
Alma-Ata, Kazakhstan, USSR, 480021

ABSTRACT

The hypothesis of a word-stress lack in the Turkic languages phonetic system is being put forward. The constitutive, culminative and word-distinctive functions of synharmonism as well as the role of synharmonic co-articulation both in Turkic syllable formation and in syllabation are being elicited. Synharmontsm predetermines the linear size of a morpheme, the latter being less than a syllable does not exist in the Turkic languages.

The Turkic word prosody has been completely reduced to the Indo-European word-stress, vowel-harmony being neglected as the result of thereof. Besides, vowel-harmony is a phonologically unjustified and phonetically inexact term. However, the idea of vowel harmony which presupposes the presence of at least 2 syllables in a word, has played its misleading role as far as research is concerned, monosyllabic words being excluded from the field of study.

It is inferred from the present situation in the Turkic prosody that synharmonism has been denied a proper place in the succession of known in general linguistics prosodic units. Being theoretically the most thoroughly elaborated ones, stress and tone remain nowadays in general linguistics as the only generally recognized prosodic units.

"Europocentrism" has not solved and is not in a position to, cardinal problems of the Turkic phonetics, the reason thereof being the transference of the accentual (non-synharmonic) languages phonological analysis principles and means to the non-accentual (synharmonic) ones.

Accumulated experimental data have failed to lay down the basis for creating the Turkic phonological theory as it is, because of their "europocentriot" interpretation, the latter, in the final run, confirming the result obtained by traditional acoustic methods.

Attempts to produce evidence for the existence or lack of Turkic stress and its place only by means of experimental phonetics' methods are bound to fail.

To our mind, the reasons thereof are as follows: all the linguistic functions of synharmonism are, this way and that, attributed to word-stress, the latter acquiring the status of an important linguistic unit in the eyes of researchers. The identity of both word-stress functions and synharmonism as the same level prosodic units presents illusive logic of such a substitution, and the hypnotic influence of word-stress ideas still remains an insuperable obstacle.

The problem seems to envisage a Turkic word either having accentual nature (and this means segment analysis being carried on phonemic level) or synharmonic one (thus, obliging us to find the predetermined by it, principles of division into functional synharmosegments and synharmosegments proper).

This problem still remains obscure as researchers fail to understand the fact that phrasal words and words in a phrase are intonationally alike as far as sentence prosody is concerned. However, reseachers differentiate phrasal words as isolated words proper, in contrast to the same words in a phrase. As a result, various manifestations of phrasal word intonation are interpreted as acoustic stress correlatives. Taking into consideration that both phrasal words and expanded phrases can be pronounced with various logical, emotional, expressive accessory intonation in dependence of the phrasal word contextual semantics, the difficulty of word-stress unambiguous interpretation is quite understandable. In effect, researchers are oblivious of the fact that isolatedly pronounced words, allegedly proving word-stress presence in the Turkic languages, are, in fact, contained in a syntactical unit with a more-than-word volume, and, thus, they bear partly this units' intonation.

Hence, the Turkic "word-stress" does

not posses the Indo-European word prosody's main characteristics of being the word acoustic image's obligatory element.

The analysis of results in various investigations into the Turkic languages "word-stress" nature shows that one should speak of phrasal, rhythmic-syntagmatic,logical-expressive prominence of this or that syllable in the word rather than of word-stress, and this makes quite a difference. Phonemic stress significance in the Indo-European languages and tone significance in the syllabic ones is known to win recognition from all researchers. As far as synharmonism in the Turkic languages was concerned, phoneticians did not pay special attention to it from the functional analysis point of view despite synharmonism being recognised as a really existing acoustic phenomenon.

Partially this occured due to the fact that turkologists-phoneticians as we have already told cherished their pettheory of word-stress.

Thus, the leading phonological function of synharmonism pertains to keeping homogenious synharmotimbre of the Turkic word's whole image, this being the obligatory element of its phonetic image. Violation of homogenious character of timbre disrupts the word, grates upon the ears, impedes perception or makes it absolutely unintelligible.

Thus, the constitutive function of synharmonism provides proper recognition of a word. For example, such Turkic words as [bɑs], [b'es'], [b°os°], [b'°ös'°] etc. are characterised not only by a certain linear combination of sound but also by unique quality of each word's synharmotibre, both vowels and consonants alike being synharmonic timbre bearers. The importance of synharmonism constitutive function is also proven by the fact that any synharmonically properly organised word is easily and correctly pronounced by Turkic languages native speakers. A word sounds familiar though its meaning may not be clear, such as dialectal or professional vocabulary.

Another function of similar importance is the culminative one, i.e. unification of word image forming sounds. Provided the word is polysyllabic, all syllables are organised according to one of the synharmonic timbres. This function plays an important role in the Turkic word general phonetic image formation and it proves synharmonism being characteristic not only of polysyllabic words but monosyllabic ones as well. It means that the terms "synharmonism" and "vowel harmony" are not synonyms, the latter being inexact both as a term and a phenomenon. Vowels play but syllable forming role in the Turkic languages without being "harmonisers" and, moreover, without performing word-distinctive function.

From the phonetic point of view, the set of synharmonic allophones (synharmosounds) proper is of great importance. Correct recognition of a Turkic word under unfavourable phonetic conditions of communication, as in case of vowel devoicing, depends on the audibility of the whole syllable, i.e. on consonant synharmonism.

Word-distinctive function of synharmonism is significant as well. One can prove it by taking minimal (and polysyllabic) synharmonic pairs or quartettes of words widely used in the Turkic languages. (One can say that Turkic vocabulary contains systems of synharmonic pairs or quartettes of words). For instance, the words [tys], [t'ıs'], [t°us°], [t'°üs'°] are distinguish not only through vowel synharmonism, but also by consonant one. Participation of all sounds comprising the word in word distinction (contrast)is strictly obligatory. It is impossible for any synharmonic variant of one consonant to be replaced by another one. In other words, the above words should not sound as [tys'] or [t'ıs] or [t°us] etc. Such a violation brings about unnatural sounding of profoundly Turkic words which become inconvenient to be pronounced by native speakers of the Turkic languages.

Since spectral characteristics of sounds comprising a syllable (a word) is acoustic correlative of synharmonic timbres, and its general spectral picture is, in a certain way, retentionary and constant, one can speak of register character of synharmonic timbres. These timbres are distinguished from one another by this or that order of placing vowel and consonant formants. A certain type of synharmonic timbre (its characteristic acoustic contour) begins with a consonant preceding a vowel (if a syllable begins with a consonant) and is expanded over to a consonant concluding a syllable (if the syllable ends up with a consonant). Thus, synharmonic timbre is a property of the whole syllable, both vowel and consonant included.

Existence of synharmonic timbres is proven by their functioning as word-distinctors, word formers and word-dividers, thus the difference between them being of phonological significance. Since synharmonic phonology allows to distinguish 4 synharmonic timbres (hard, soft, labial, nonlabial), Turkic languages can be called polytimbral ones.

Thus, the language functions inherent in stress of accentual languages and in tone of syllabic ones are found on the Turkic languages in synharmonism. This shows their functional identity in general linguistics plane and seems to represent important leading typological features pointing at principal differences rather than at similarrities of these language groups.

In connection with these typological differences one should distinguish, in consecutive order, phonetic (universal) co-articulation, characteristic of all languages as a result of mutual influence of ajacent sounds, and phonological (particular) one, where it is preconditioned by the Turkic languages synharmonism. Violation or some inconsistency of co-articulation in the first case is quite possible, while in the second case it should be strictly observed, for each synharmosegment, synharmosyllable included, is a phonological unit. That is why, one should look for syllable division types, syllable boundary features in synharmonism. A Turkic syllable is the smallest pronounced language unit, acoustically strictly limited by one of synharmonic timbres. Such limitation is sostable that synharmonic co-articulation violation within a syllable is absolutely impossible. The feature of syllabic synharmonic co-articulation is that very linguistic signal showing syllable-boundary.

To our mind, a morpheme less than a syllable does not exist in the Turkic languages. The morpheme linear size is equal, at least, to a syllable. It is predetermined by the very nature of synharmonism, for timbral characteristics of synharmonism can be realised only in a syllable. It excludes the existence of consonant morphemes, while vowels comprise morphemes because they can form syllables independently.

Traditional concepts of general linguistics were unable to explain the fact that the word's first syllable predetermined the synharmonic accesory of the antecedent syllables, their stable phonetical homogenity, i.e. strong position of the first syllable and recognised by all researchers fixed stress at its end which means another strong position at the opposite end of the word. This led to a compromise: the existence of 2 opposite strong positions in the Turkic structure is recognised, i.e. word-stress and synharmonism which allegedly complement each other.

Such a paradoxical compromise would not have existed, if the Turkic word prosodic feature were scientifically justified and were not attributed to disagreeable with it accentual prosody.

We have formulated the following main principles:

1. Word-stress, word-tone and word-synharmonism as the same level units, implement analogical functions, i.e. they unite acoustic segments of words into the wholw. While word-stress is prosodical means of a word unity on a phonetic level in the Indo-European languages, word tone is the same means for the syllabic languages, and word synharmonism - for the Turkic (and, possibly, for all the Ural-Altaic) languages. These means are equal in carrying out constitutive and word-distinctive functions. Therefore, each of these means contains prosodic feature, characteristic for a certain language type.
2. All the 3 means, being prosodical features of the word, regulate phonetic gradation of syllable, i.e. word-stress-accentual (stressed, pretonic, counter-pretonic, posttonic, etc. syllables), word tone-tonal (low, medium, high, rising, falling and the like registers), word synharmonism - timbral (hard non-labial, hard labial, soft non-labial, soft labial timbres).
3. Each of the 3 means originally regulates articulation-acoustic interaction (co-articulation) of sounds in a syllable.
4. Each of 3 means accomplishes specific word division into minimal (in functional plane) sound segments, i.e. word-stress-into phonemes, word-tone-into tonemes, word synharmonism - into synharmosegments (synemes).
5. The common basic phonetic unit for all 3 means is a syllable, but their phonetic realisation is different.
6. In our opinion, the existence of all or 2 identical in function but different in realisation types of word prosody in one language or a related languages group phonetic system is impossible. Therefore, the word-stress existence in the Turkic languages should be considered false.

ACOUSTIC VS. LEXICAL JUDGEMENTS IN THE PERCEPTION OF FALLING ACCENTS IN SERBO-CROATIAN: A PRELIMINARY STUDY

VESNA MILDNER

University Computing Center
University of Zagreb
41000 Zagreb, Yugoslavia

LEIGH LISKER

Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104, USA

ABSTRACT

This paper concentrates on the difference in duration between the long falling and short falling accent in Serbo-Croatian. Another aim of the study was to determine whether listeners who do not speak the language would be able to make acoustic distinction between the two, and if they would, whether they would shift their judgements from long to short at the same point or along the same lines as the speakers of Serbo-Croatian.

INTRODUCTION

In Serbo-Croatian (SC) the word accent consists of three elements: stress, length and pitch. The combination of these elements gives four accent types: short falling (\\); long falling (∩); short rising (\) and long rising (/).

There are some restrictions with regard to the distribution of the four accent types: in monosyllabic words only the falling accent can occur; the last syllable is never accented; polysyllabic words can carry the falling accent only on the first syllable. While tonal patterns (pitch) are associated with stressed syllables, "the quantity system is relatively more independent, since quantity contrasts also occur in unstressed syllables" /1/. Apart from numerous dialectal variations, two variations are acceptable in standard SC: optional short tonal distinction and optional or non-existent posttonal vowel length.

The fact that only the falling accent can occur in monosyllabic words has probably lead J. Gvozdanović /2/ to an imprecise conclusion that "monosyllabic prosodic words have no tone and accent, but can only have prominence in a phrase or a sentence". The very fact that a word has prominence (stress) requires that word to have one of the accents (hence, tone). A more precise statement pertaining to tone in monosyllabic words is given by Lehiste and Ivić /1/: "Monosyllabic words do not show tonal contrasts." This statement strikes closer to home, since in monosyllabic words, bearing only falling accents, the contrast can be primarily found in their duration and, possibly, fundamental frequency peak location /3,4,5,6/ and final fundamental frequency value of the accented vowel /7,8,9/. Lehiste and Ivić /1/ state the following: "From the point of view of the system the short and long rising accents differ from each other in terms of duration; in the same way, the two falling accents differ in terms of duration. The cue value of the difference in the placement of fundamental frequency peak on the stressed syllable thus seems dependent on length...".

Most authors studying SC accents have dealt mainly with disyllabic or polysyllabic words and concentrated on the distinctions between long falling and long rising or short falling and short rising accents, probably due to the fact that these distinctions are richer, and depend on a number of variables.

The main aim of this study was to concentrate on the difference in duration between the long falling (LF) and short falling (SF) accent, keeping all other parameters constant. That the question of duration in these two types of accent is not trivial was shown by the studies of Magner and Matejka /10/ who, among other distinctions, tested the perception of native speakers of SC in an attempt to determine how much of the accentual system developed by V. Karadžić in the early 19th century and adopted as standard for SC, is in actual use and whether native speakers of the language who may not use all the distinctions in their own speech can still detect those differences and make lexical judgements based on them. They found that not all of their listeners could identify the distinction between the short and the long falling accent in the word *pas* when presented with the natural production of these words in sentences *Čiji je to pȁs tamo (Whose dog is that there)* and *Čiji je to pâs tamo (Whose belt is that there)*. Unfortunately, the authors do not provide any acoustic measurements, so it is not known from their reports what the duration of the accented vowel in the target words was. However, their results show that speakers of SC in most of the major cities can identify the difference between the long and the short accented vowel and conclude "that in their speech accentual quantity is meaningfully utilized and appears as a functional prosodic system". The authors have also found that even speakers who do not distinguish these two accents in their own speech (big cities), "are capable of identifying distinctions which they themselves do not implement...".

MATERIALS AND PROCEDURE

Preparation of test material: Two native speakers of SC from the city of Zagreb, who both utilize the long-short distinction in their speech, recorded several tokens of the word *pâs (belt)* and *pȁs (dog)* in medial and final sentence positions and in isolation, via a Crown 700 series taperecorder, using an Electro-Voice microphone (model 635A Dynamic Omnidirectional).

The tokens were then sampled via an analog-to-digital converter with a rate of 12,500 samples per second. The samples were stored in a PDP-11 digital computer. A low-pass filter with the cut-off frequency of 5000 Hz and a slope of -48 dB per octave was used to filter out the 12,500 sampling frequency. Using the ILS package for acoustic analysis the tokens were displayed and the duration and fundamental frequency contour calculated and displayed. Table 1 shows vowel durations for different tokens of *pȁs* and *pâs*.

Table 1. Duration of the five tokens of *pȁs* and five tokens of *pâs* (in msec) in ascending order

Accent type	SF (\\)	LF (∩)
(msec)	90	170
	120	210
	130	220
	140	240
	140	250

As it can be seen from Table 1., the longest vowel bearing the SF accent was 140 msec and the shortest vowel bearing the LF accent was 170 msec long. These data are in agreement with those of Lehiste and Ivić /11/ for di- and polysyllabic words bearing short and long falling accents. It should also be reported here that, although most authors have found a slight rise, peak and then the fall of the fundamental frequency contour in samples of falling accents, no such movement of fundamental frequency was found in any of the tokens here. This can be explained by the fact that the consonant preceding the examined vowel was a voiceless stop (/p/) and it has been found (/12/ and an earlier study of this author) that in that case the peak occurs immediately after the onset of voicing.

One of the originally recorded sentences, *Ovo je krasan pâs (This is a beautiful belt)*, was chosen as the starting point for all the other test sentences. In that particular sentence the vowel in the word *pâs* had a duration of 185 msec. Of these 185 msec 138 msec was the duration of the voiced interval and 47 msec was the duration of the whisper-like portion which could still be identified as vowel (/a/). In order to keep the relationship between the initial and final fundamental frequency value constant, the vowel was then shortened in such a way that individual complete pitch periods were removed from the stimulus. These pitch periods were extracted at regular intervals, using WENDY, the in-house program for acoustic analysis, on a VAX computer, at Haskins Laboratory, New Haven, CT. For each magnitude of reduction the periods were chosen so as to be equally distributed over the voiced period. By this method all the parameters except duration were kept constant. In this way 8 different tokens of *pas* were obtained, all incorporated into the same carrier sentence *Ovo je krasan ___ (This is a beautiful ___)*. Each of these 8 sentences was then recorded 4 more times, which yielded 40 test sentences. The sentences were randomized, with a silent interval of 3 seconds between subsequent sentences. Table 2. shows the 8 durations of the vowel in the word *pas* (including the whisper-like portion).

Table 2. Durations of the vowel /a/ in the word *pas*

Stimulus	Duration (in msec)
1	185
2	174
3	163
4	158
5	147
6	137
7	131
8	119

The experiment: There were two groups of subjects. One group consisted of 6 native speakers of SC. All the subjects in this group were born and raised in the city of Zagreb, and all of them utilize the long-short distinction in their own speech. The second group consisted of 8 Americans - seven students and one professor of linguistics. None of them speak SC.

Native speakers of SC were asked to make lexical judgements. Each subject was provided with answer sheet consisting of 40 pairs of words *životinja/pojas (animal/*synonim for *belt)* and was asked to underline or circle the one which, in their judgement corresponded to the stimulus used in the sentence. The American subjects were asked to make acoustic judgements. It was explained to them before the test that all the sentences would be the same except for the last word, in which the duration of the vowel would vary. It was pointed out that they should only pay attention to the duration of that vowel. Each subject was provided with an answer sheet consisting of 40 blank lines on which he/she was asked to write L (for long) or S (for short), depending on their judgement of the duration of the vowel in the last word of the sentence.

Before the test both groups were presented with two sentences containing the longest vowel in the word *pâs* followed by two sentences containing the shortest vowel in the word *pȁs*. These four sentences served as a training session for the American subjects and as control for the group of native speakers of SC. Those native speakers who could not make the distinction between the two extremes were not tested.

The sentences were presented to the listeners in a free space room via the Crown 700 series taperecorder, connected to a Z-400 Jans Zen electrostatic loudspeaker through the Crown D60 Model amplifier, at a comfortable listening level, approximately 2 meters from the listeners.

Figure 1. shows pooled responses of native speakers of SC in terms of percentage of long *(pâs)* and short *(pȁs)* responses to a particular vowel duration. As it can be seen from the Figure, the perception of the long-short distinction is very nearly categorical for native speakers of SC. The cross-over point is at stimulus 5, in which the duration of the vowel was 147 msec (43.33% *pȁs* and 56.67% *pâs* responses). Stimulus 4 (vowel duration of 158 msec) elicited 80% *pâs* and 20% *pȁs* responses while stimulus 6 (vowel duration of 137 msec) elicited 86.67% *pâs* responses and 13.33% *pȁs* responses.

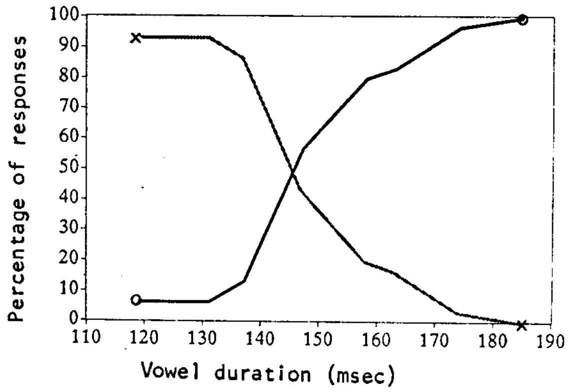


Figure 1. Pooled responses of native speakers of SC to eight different vowel durations (o - *pâs*; x - *pȁs*)

With respect to their responses American subjects can be divided into two groups. Five out of 8 Americans made obviously random judgements of the vowel duration. No pattern was found that might indicate at least a tendency to label the stimuli with some consistency in accordance with their duration. Three out of 8 American subjects were non-random in their responses. Figure 2. shows pooled responses of these three listeners in terms of percentage of long and short responses to a particular vowel duration.

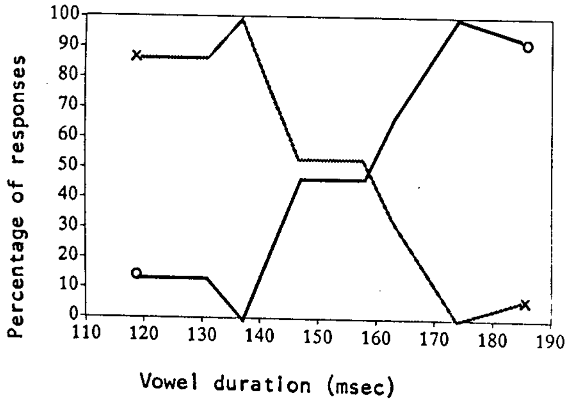


Figure 2. Pooled responses of 3 American subjects who had non-random responses to eight different vowel durations (o - long; x - short)

As it can be seen from the Figure, these 3 listeners exhibit a near categoricity of perception. Two things distinguish these listeners from the native speakers of SC. First of all, their responses are random for two stimuli, rather than one, which shows that their perception is not as categorical as that of native speakers of SC. The vowel duration of these two stimuli (4 and 5) was 158 and 147 msec, respectively, with 53.33% "short" and 46.67% "long" responses in each. Obviously, the significant shift in judgements from "long" to "short" occurs at the same point as for native speakers of SC while the point at which their responses become random occurs earlier on the duration scale, than for the native speakers of SC.

The other interesting detail that can be observed in Figure 2. is that these 3 American subjects are not entirely consistently shifting their judgements. Several unexpected peaks and valleys can be seen in Fig. 2. - 100% "short" or "long" judgements do not occur in responses to the stimuli of shortest or longest duration, respectively. Stimulus 1 elicited 93.33% "long" responses. Similarly, of the 3 stimuli predominantly labeled as "short", the longest one, stimulus 6, elicited 100% "short" responses, while the actually shorter stimuli 7 and 8 elicited 86.67% "short" responses each. Closer examination of the responses of these 3 listeners and the order of stimuli presentation shows that all tokens of stimulus 2 (100% "long" responses) and stimulus 6 (100% "short" responses") occur after the 12th position on the test tape. It appears that these listeners were actually in the process of establishing some sort of a reference scale in the first quarter of the test and all the inconsistencies are found in responses to stimuli presented as the first 12 test stimuli. This indicates that the more categorical perception of native speakers of SC is a result of their being more attentive to phonemic length which they use and hear in everyday communication. On the basis of these results and observations it can be expected that re-testing of the same 3 American subjects or providing them with a short pre-test session, which would include all durations, rather than just the extremes, would yield results closer to those obtained for native speakers of SC.

It should also be noted that 3 out of 5 tokens of stimulus 5 (vowel duration of 147 msec), to which random responses were given both by native speakers of SC and the 3 Americans, occurred very early in the test (positions 3, 8 and 9). Stimulus 5 was only slightly (7 msec) longer than the longest vowel bearing the SF accent, found in acoustical measurements preceding the experiment and in literature. The fact that the stimulus of such "borderline" duration was presented so early in the test might have contributed to the randomness of responses of the above mentioned subjects. It remains to be determined whether a pre-test session provided for the native speakers of SC would result in a clearer switch from *pâs* to *pȁs* judgements, without randomness of responses in between. The acoustic measurements of natural productions of the words *pâs* and *pȁs* carried out during the preparation for the experiment, as well as the data found in literature, show that the vowels bearing the LF accent are not shorter than 170 msec and that the vowels bearing the SF accent are not longer than 140 msec. The results of this study indicate that the native speakers of SC are more apt to label shorter-than-natural durations of vowels under LF accent as long than the longer-than-natural durations of vowels under SF accent as short. Even the fact that the word *pȁs* is more common than the word *pâs* (which has become to be regarded as slightly archaic and is not frequently used in modern SC, did not cause a bias toward it in the judgments of native speakers of the language.

CONCLUSION

On the basis of the results of this pilot study the following conclusions can be drawn:

- Native speakers of SC, who utilize the long-short distinction (between the LF and the SF accent) in their own speech, exhibit categorical perception of this distinction when presented with words (in carrier sentence) which differ only in the duration of the vowel and when asked to make lexical judgements.
- The cross-over point, at which the judgements of native speakers of SC shift from long (*pâs*) to short (*pȁs*) occurs at the stimulus with the vowel duration of 147 msec, which is slightly longer than the longest duration of the naturally produced vowel bearing the SF, found in literature and in preliminary acoustic measurements.
- American subjects, who do not speak SC, exhibit two types of perceptual behavior in their acoustic judgements of the duration of the target vowel - their responses are either entirely random or show a pattern similar to that found in the responses of native speakers of SC.
- American subjects whose responses are not random start to shift their judgements from "long" to "short" earlier than the native speakers of SC, i.e. at a longer stimulus (158 msec) but the significant switch occurs at the same point as for native speakers of SC (137 msec).
- There is evidence that native speakers of SC are more attentive to the long-short distinction than the American subjects, which can be attributed to the fact that vowel duration is phonemic in SC and native speakers of this language utilize it in their own speech and hear it in everyday communication.
- Testing of larger groups of subjects is necessary to determine which type of perceptual behavior is more characteristic for the Americans who do not speak SC.

REFERENCES

/ 1/ Lehiste, I. and P. Ivić, "Interrelationship between word tone and sentence intonation in Serbo-Croatian", In: Napoli, D.J. (ed.) *Elements of Tone, Stress and Intonation*, 100-129, Washington, 1978.

/ 2/ Gvozdanović, J., *Tone and Accent in Standard Serbo-Croatian*, Wien: Osterreichischen Akademie der Wissenschaften, 1980.

/ 3/ Purcell, E.T., *The Realization of Serbo-Croatian Accents in Sentence Environments*, Hamburger Phonetische Beitrage, Hamburg, 1973.

/ 4/ Purcell, E.T., "A model of word-tone sentence intonation and segmental duration in Serbo-Croatian: A preliminary report", In: Kourbourlis, D. (ed.) *Topics in Slavic Phonology*, Slavica, 178-202, Cambridge, 1974.

/ 5/ Purcell, E.T., "Pitch peak location and the perception of Serbo-Croatian word tone", *Journal of Phonetics 4*, 265-270, 1976.

/ 6/ Purcell, E.T., "Two parameters in the perception of Serbo-Croatian tone", *Journal of Phonetics 9*, 189-196, 1981.

/ 7/ Lehiste, I., "Some acoustic correlates of accent in Serbo-Croatian", *Phonetica 7*, No. 2-3, 114-147, 1961.

/ 8/ Lehiste, I., "Influence of fundamental frequency pattern on the perception of duration", *Journal of Phonetics 4*, 113-117, 1976.

/ 9/ Lehiste, I. and P. Ivić, "Experiments with synthesized Serbo-Croatian tones", *Phonetica 26*, 1-15, 1972.

/10/ Magner, T.F. and L. Matejka, *Word Accent in Modern Serbo-Croatian*, The Pennsylvania State University Press, University Park and London, 1971.

/11/ Lehiste, I. and P. Ivić, "Accent in Serbo-Croatian: An experimental study", *Michigan Slavic Materials 4*, University of Michigan, Ann Arbor, 1963.

/12/ Ivić, P. and I. Lehiste, "Prilozi ispitivanju fonetske i fonološke prirode akcenata u suvremenom srpsko-hrvatskom književnom jeziku", *Zbornik za filologiju i lingvistiku 6*, 31-37, 1963.

AN ACOUSTIC STUDY ON MURMURED AND "TIGHT" PHONATION IN GUJARATI DIALECTS - A PRELIMINARY REPORT

CH. LANGMEIER U. LÜDERS L. SCHIEFER

Institut für Phonetik und Sprachliche
Kommunikation der Ludwig Maximilians-
Universität München, FRG

BH. MODI

Dept. of Linguistics
Faculty of Arts,
M.S. University of
Baroda, India

ABSTRACT

The purpose of our study was twofold: (i) to define "tight" phonation in acoustic terms and (ii) to examine the acoustic differences between murmured and "tight" phonation in Gujarati. The analysis was based on the parameters: Fo contour, overall intensity, amplitude of the 1st and 2nd harmonic, the frequency of F1 and F2, and the bandwidth of F1 and F2. The amplitude of the first two harmonics as well as the bandwidth of F1 and F2 turned out to serve best in distinguishing murmured from tight phonation.

INTRODUCTION

Gujarati -an Indo-Aryan language- is usually treated as a member of that group of languages which contrast murmur phonation and normal voicing. Both phonation types are used on the one side to separate murmured from clear vowels, on the other side they serve to distinguish murmured stops from voiceless, voiceless aspirated, and voiced ones. Acoustical analyses of murmur which have been carried out since the late fifties revealed several acoustic parameters by which murmur may be distinguished from normal voicing. Murmur is characterized by the following features: a lowering of fundamental frequency (Fischer-Jørgensen [2], Ohala [6], Schiefer [8]), an increase in the amplitude of the first harmonic in relation to the second one (Bickley [1], Ladefoged [4], Huffman [3]), broader formants [2], a later onset of higher formants [2], a lowering of the second formant (Pongweni, 7), an irregular intensity course [2], and a lowering of the overall intensity [8]. One of the most extensive acoustic studies on Gujarati, and a quite early one, is that of E. Fischer-Jørgensen [2], who examined the differences between murmured and clear vowels. It is apparent that the seven subjects used in her investigation showed great variability in producing murmured vowels. As Fischer-Jørgensen points out "all informants have murmured vowels in their natural speech, and this pronunciation seemed to be very constant for PvB, SK, and GU. In RD's and PBP's speech murmur is optional" [2, p.74].

The differences between the subjects seem to reflect different dialects, as RD and PBP were born in Saurashtra (western part of Gujarat), whereas PvB (Baroda), SK (Surat), and GU (northern Gujarat, Ahmedabad) originate from the northern and eastern part of Gujarat. The dialectal differences of Gujarati have been subjected to an extensive study by one of us (Modi, 5), who employed the method of tomography in her analysis. It appeared that two dialect groups have to been treated separately according to the phonation types used. One group, which she calls "murmur", shows a low larynx position, whereas the other group ("tight") has a high larynx position in order to avoid murmur phonation. As the term "tight" for the non-murmur dialects was introduced impressionistically by Modi [5] it still lacks definition in terms of acoustic features. The aim of our present study was therefore to examine the influence of several acoustic parameters in murmur and tight phonation. The following parameters have been examined: (i) the course of the fundamental frequency (Fo), (ii) the overall intensity, (iii) the amount of energy in the first (H1) and second (H2) harmonic, and (iv) the frequency of F1, F2 as well as (v) their corresponding bandwidths B1 and B2.

MATERIAL AND INFORMANTS

Our analysis was based on a rather limited material, and the results should be taken as a preliminary report on the selectivity of the acoustic parameters for the separation between murmur and tight phonation. We based our analysis on murmured stops rather than vowels as we felt that the stops would provide the most stringest test for the saliency of the single acoustic parameters. Murmured stops occur in both dialects and are contrasted from the other stops by a distinctive release of the stop, which is characterized by an incomplete closure between the vocal folds during the phonatory cycle.
The material consisted of isolated words containing the murmured stops in five places of articulation (labial, dental, retroflex, palatal, and velar) followed by the vowel /a/ in word-initial position. Each CV syllable occurred five to 15 times in the material. The material was recorded on tape in Baroda, India. One speaker (male) from Rajkot and one from Ahmedabad served as informants for tight and murmured phonation.

PROCEDURE

The acoustic analysis of the data was run in Munich, where the words were digitized (using a sample rate of 20 kHz) filtered with a cut off frequency of 8 kHz and stored on a PDP11/50. The periodic portions of the initial CV syllables of all words were segmented into single pitch periods by the help of a segmentation routine (for further detail cf. [8]) and stored for further analysis. The fundamental frequency was calculated from the segmented material and measured for the first 14 pitch periods after the burst of the stop. The intensity was measured for the same vowel portion. The same (segmented) material was used to calculate the contribution of H1 and H2 to the overall intensity of all pitch periods of the vowel. A second analysis was run on the unsegmented data in order to gain F1, F2 data and their corresponding bandwidths by the use of a LPC procedure. The following adjustments were made: frame size = 512 samples (this is equivalent to a segment duration of 25.6 ms), window shift size = 128 samples, filter degree = 22, Hamming window size = 512 samples, preemphasis factor = 0.7. There was a limitation for bandwidth of the formants, which could not exceed 2/3 of the formant's value. Greater bandwidth led to a rejection of the formant proposed by the routine. As great problems were involved in the calculation of F1 in murmur (for detail see below) this preliminary analysis was run on the velar stops only. Separate multivariate two factorial analysis of variance were run for (i) F0, (ii) intensity, (iii) H1 and H2, (iv) F1, (v) F2, (vi) B1, and (vii) B2.

RESULTS

Fundamental frequency. The results are given in Figs. 1 to 3 and in Table 1. The differences in Fo between both speakers are small. Fo at vowel onset is low in both speakers and increases towards P14. In the murmured dialect a Fo fall from P1 to P3 can be observed, which is obviously not produced by the other speaker. Concerning the influence of the stop's place of articulation great differences between the dialects can be observed. The murmured speaker shows a quite regular pattern as for all stops a fall from P1 to P2/P3 can be found and a quasi-linear rising towards P14. The Fo differences at vowel onset are smaller than at the end of the contour. At the end of the (measured) vowel portion higher Fo values are assigned to the [+ant] (/dh bh ḍh/), lower values to the [-ant] stops (/ǰh gh/). The tight-phonation speaker shows somewhat greater Fo differences at vowel onset, a rising Fo after /bh jh/ and a falling-rising pattern (from P1 to P4) after /gh dh ḍh/. The difference between the stops at P14 is greater and Fo seems to depend on the apicality of the stop rather than on its position: [-apic] stops show slightly higher, [+apic] stops lower values.

Table 1: Statistical results for Fo, intensity, and harmonics H1 and H2 D=dialect, P=place of articulation, H=harmonics.

	Fo	intensity	H1/H2
INTERACTIONS			
D-P-H	---	---	n.s.
H-P	---	---	n.s.
D-H	---	---	< .001
D-P	<.001	<.001	n.s.
H1/H2	---	---	<.001
DIALECT	<.01	<.001	<.001
PLACE-OF-ARTIC	n.s.	<.001	<.001

Intensity. Fig. 4 displays the results for the intensity averaged over all places of articulation for both speakers, whereas the influences of the place of articulation are plotted separately for murmur and tight in Figs. 5 and 6, respectively. The statistical results are given in Table 1. The intensity is lower in tight than in murmur phonation. In both dialects the intensity is lowest at vowel onset, increases rapidly towards P3/P4, and increases slowly towards the end of the contour in murmur, whereas in tight phonation the amount of increase is greater from P8 to P14, which indicates a change in the underlying phonation process. In murmur the influence of the place of articulation on the intensity is small, smallest at vowel onset and increases slightly towards the end of the contour. The increase in intensity over the contour is nearly the same for all stops. At P14 [+ant] stops show a somewhat greater intensity than do [-ant] stops. In tight phonation the influence of the stop's place of articulation is greater at vowel onset as well as at the end of the contour. The intensity is greater in [+ant] stops and less in [-ant] ones. The intensity course after /gh/ differs significantly from the other ones as there is an abrupt increase in intensity after P9. This again can be explained by a change in the underlying phonation type, as we believe that murmur can be sustained after /gh/ only if it is accompanied by a low larynx position.

Amplitude of H1 and H2. Figs. 7 and 8 display the results for H1 and H2 for both dialects, whereas the statistical results are again given in Table 1. We have measured the amount by which the single harmonics contribute to the overall intensity of the single pitch periods. In tight phonation the amount of energy is slightly higher in H1 than in H2. This feature is associated, as mentioned above, with murmur phonation. The difference remains relatively constant throughout the vowel. In murmur on the other hand the difference between H1 and H2 is much greater. Whereas the course of H1 and H2 is nearly level in tight phonation, the amount of energy in H1 increases in murmur from P1 to P14. H2, on

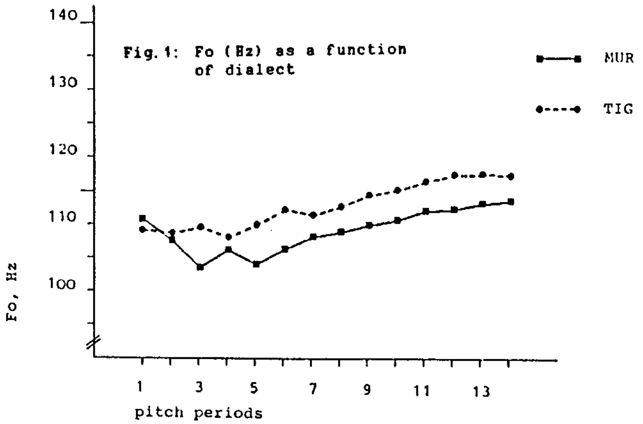


Fig. 1: Fo (Hz) as a function of dialect

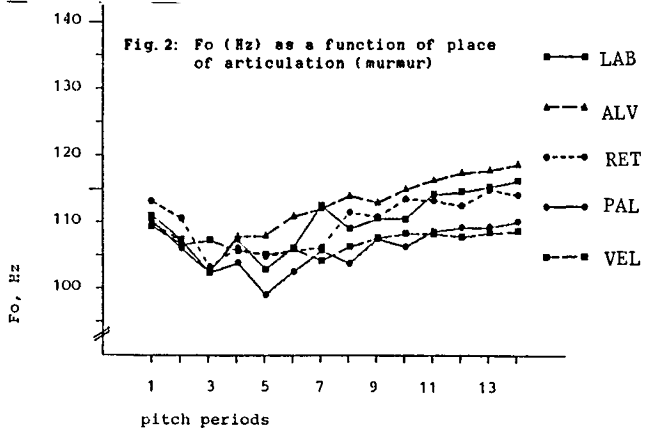


Fig. 2: Fo (Hz) as a function of place of articulation (murmur)

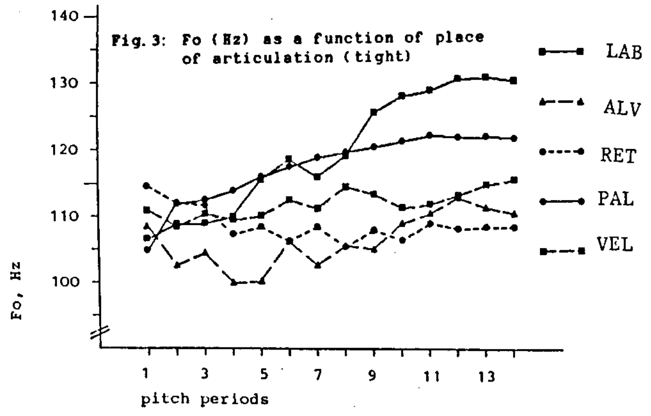


Fig. 3: Fo (Hz) as a function of place of articulation (tight)

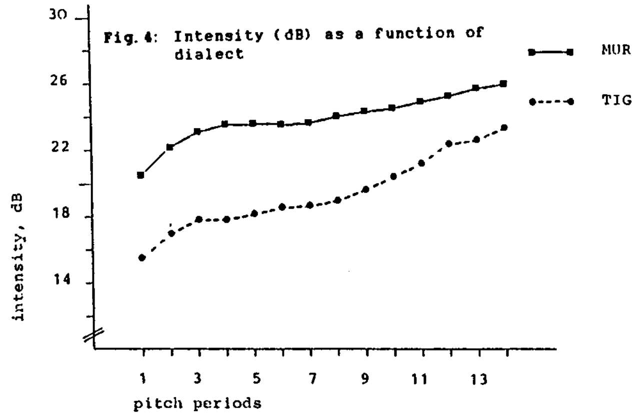


Fig. 4: Intensity (dB) as a function of dialect

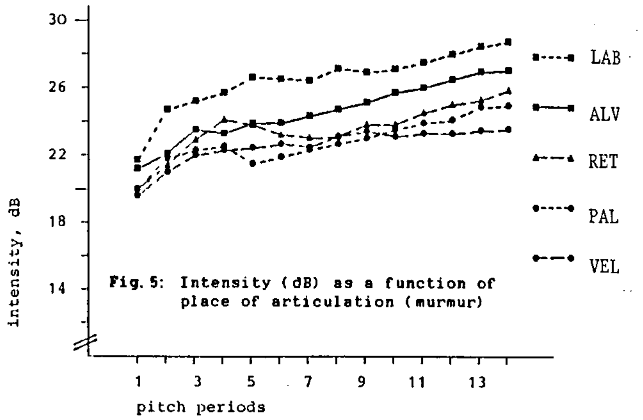


Fig. 5: Intensity (dB) as a function of place of articulation (murmur)

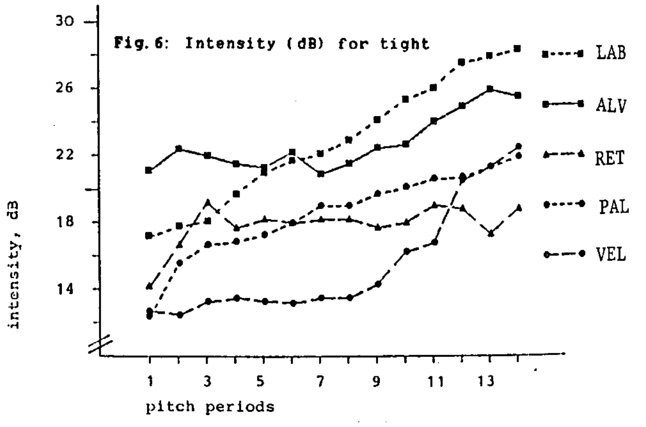


Fig. 6: Intensity (dB) for tight

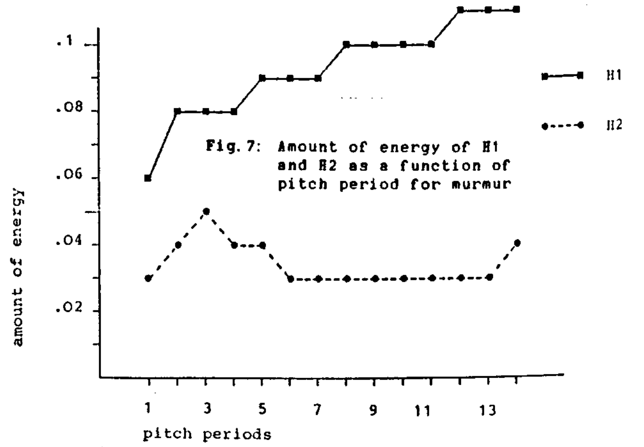


Fig. 7: Amount of energy of H1 and H2 as a function of pitch period for murmur

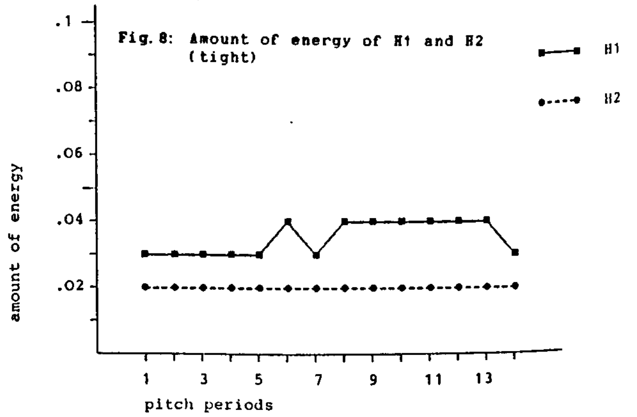


Fig. 8: Amount of energy of H1 and H2 (tight)

the other hand, shows a rising-falling-level pattern. The influence of the stop's place of articulation is significant in both dialects, where [+ant] stops again have somewhat higher values than [-ant] stops.

F1, F2, B1 and B2. As the LPC failed to calculate F1 precisely for about 250 ms of the vowel after the stop's release, F1 and B1 are measured for the steady vowel portion only. The results for F1 differ extremely between the murmured and tight speaker: F1, averaged over 368 ms is 660.0 Hz in murmur and 906.5 Hz in tight phonation (for details cf. Table 2). The corresponding bandwidth is 370.9 Hz in murmur and 203.3 in tight phonation. The bandwidth decreases slowly in murmur (427.4 Hz at the beginning and 347.0 Hz at the end of the contour) and in tight phonation, where B1 is 257.8 Hz at the beginning and 149.8 Hz at the end of the contour. The frequencies of F2 are rather comparable: F2 = 1373.7 Hz in murmur and 1383.9 Hz in tight phonation, whereas the mean of B2 of tight phonation (152.0 Hz) is higher than that of murmur (125.1 Hz).

Table 2: Averaged formant- and bandwidth values and standard deviations for the murmured and tight dialect in Hz; minimum and maximum values of the formants and bandwidth; level of significance from the analysis of variance for F1, F2, B1, and B2.

	F1	F2	B1	B2
murmur				
x	660.0	1373.7	370.9	125.1
sd	86.1	60.5	130.5	76.4
MIN	622.0	1331.3	294.0	91.3
MAX	708.0	1450.3	495.2	183.5
tight				
x	906.5	1383.9	203.3	152.0
sd	57.3	111.9	76.1	84.4
MIN	872.0	1293.2	150.1	80.4
MAX	1002.0	1485.0	265.0	254.8
p	< .001	< .001	< .001	< .01

DISCUSSION

The acoustic parameters involved in this study contribute in different degree to the separation between murmur and tight phonation. The overall Fo cannot be used to distinguish between murmur and tight as it is rather a feature of the speakers voice than of the underlying phonation type. On the other hand, there are great differences in respect to the influence of the stop's place of articulation, which is small in murmur, great in tight phonation. The same is true for the overall intensity, which first of all reflects differences in the recording level, more than differences due to the underlying phonation type. But again, the place of articulation of the stop influences the intensity course more in tight than in murmur phonation. Taking both parameters together, we argue that they reflect different degrees of variability in the phonation, showing greater variability in tight phonation (with a high larynx position) and less variability in murmur, where the larynx position is low.

The results of the analysis of H1 and H2 show that in both dialects "murmur" occurs. Whereas the degree of murmur is high in murmured it is low in tight dialects. This difference in the degree of murmur is reflected by the results from bandwidths B1 and B2. In both dialects the bandwidth of F1 is much more greater than found in other languages, a fact that accounts for less sharp boundaries in the spectrum. On the other hand B1 remains great throughout the contour in murmur, but decreases in tight phonation. The results from B2 again reflect a higher degree of murmur in the murmured speaker, as the bandwidth is smaller compared to the 'tight' speaker.

In summary, the murmured stops are produced with a murmur release in both dialects. But there are differences in the degree and duration of murmur between the speakers. The amplitude of the first and second harmonics, as well as the bandwidths of F1 and F2 are the most efficient acoustic parameters to distinguish between tight and murmur phonation in Gujarati.

REFERENCES

[1] Bickley, C.: Acoustic analysis and perception of breathy vowels. Working Papers, MIT Speech Communication, Vol. 1 (1980)

[2] Fischer-Jørgensen, E.: Phonetic analysis of breathy (murmured) vowels. Indian Linguistics 28: 71-139 (1967)

[3] Huffman, M.K.: Measures of phonation types in Hmong. University of California Working Papers in Phonetics 51: 1-25 (1985)

[4] Ladefoged, P.: The linguistic use of different phonation types. University of California Working Papers in Phonetics 54: 28-39 (1982)

[5] Modi, Bh.: The laryngeal dimension in Gujarati phonology. Fifth Intern. Phonology Meeting, Eisenstadt, Austria. Wiener Linguistische Gazette, Suppl. 3: 167-171 (1984)

[6] Ohala, M.: Phonological features of Hindi stops. S. Asian Lang. Analysis 1: 79-88 (1979)

[7] Pongweni, A.J.C.: An acoustic study of the qualitative and pitch effect of breathy-voice on Shona vowels. J. Phonetics 11: 129-138 (1983)

[8] Schiefer, L.: Fo in the production and perception of breathy stops: evidence from Hindi. Phonatica 43: 43-69 (1986)

GROUPES D'OCCLUSIVES ET CLICS

ALAIN MARCHAL

Institut de phonétique , U.A. 261 CNRS
29 avenue Robert Schuman
13621 Aix-en-Provence , FRANCE

INTRODUCTION

L'étude EPG de l'enchainement des gestes articulatoires lors de la production de groupes d'occlusives fait souvent apparaître des phases de double occlusion (1). Nous nous attacherons dans cette communication a examiner les mouvements de la langue associés au relâchement de C1 .
Nos observations portent environ sur 200cas de double occlusion relevés dans la prononciation de phrases naturelles françaises répétées par trois locuteurs . Au delà de la grande variabilité des données articulatoires , il est possible d'identifier d'après les principes généraux d'aérodynamiques (2) quatre types d'événements . Ceux-ci sont illustrés par les exemples suivants .

1- Stabilité (?) de la double tenue .

Le barrage médian caractéristique du /k / est établi depuis 110 ms lorsque se produit l'implosion de /t/ dans la séquence /aktu/ (Fig . 1) . Les images 139 et 140 montrent que l'occlusion antérieure et l'occlusion postérieure sont tenues simultanément pendant 20 ms . A en juger par la stabilite des appuis linguo-palatins observee par la technique de l'électropalatographie , il n'est pas possible d'emettre d'hypothèse sur la nature du relâchement de C1 . En effet le corps de la langue a pu se creuser, se bomber ou demeurer immobile , mais nous n'avons pas de moyen direct de le vérifier

2- Concaténation des tenues

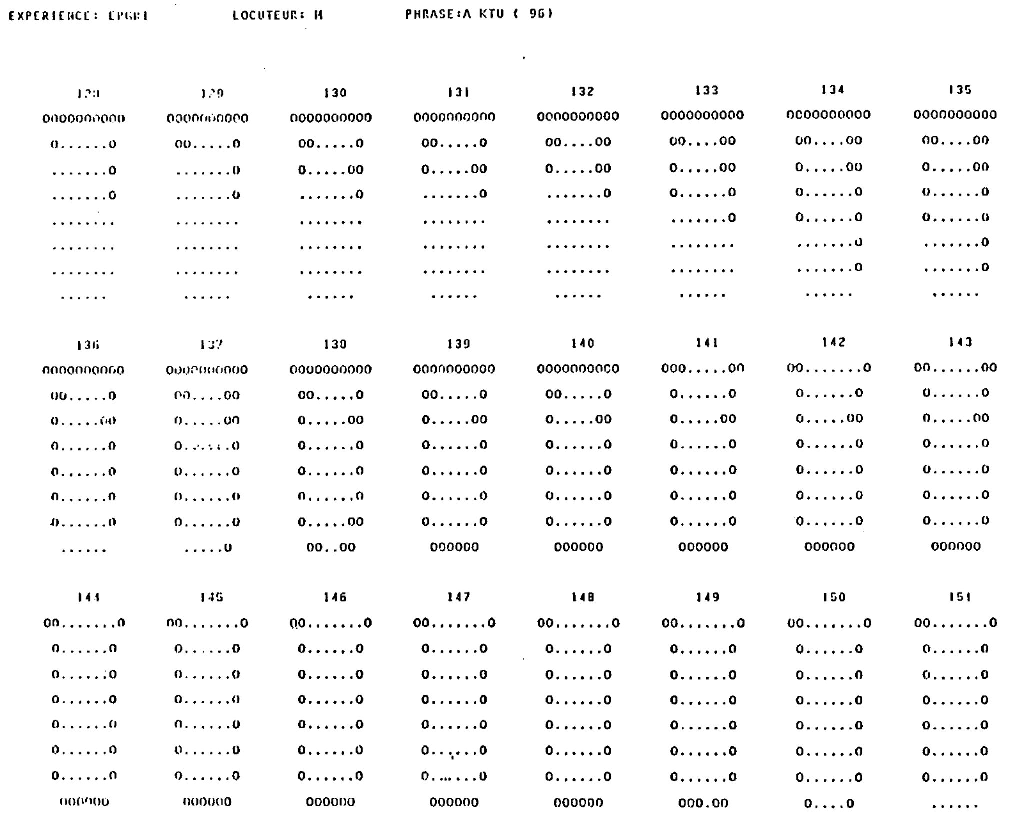
La préparation de C2 : /g/ consistant dans l'élargissement des appuis dans la zone postpalatale amorcé à l'image 101 se poursuit après l'implosion de C1 : /d/ (images 102 à 108 , Fig .2). Il s'agit de la manifestation du principe de coproduction (3) . Il n'y aura pas à proprement parler de phase de double occlusion dans cet enchainement consonantique car lorsque le barrage du /g/ s'établit , on observe en même temps le relâchement de l'occlusion dans la zone alvéolaire . La simultanéité de ces deux événements peut être liée à la fréquence d'échantillonnage des données EPG (100 Hz) . L'analyse acoustique montre effectivement que le bruit d'explosion du /d/ ne diffère pas du "burst" caractéristique de cette consonne dans un contexte comparable .

3- Diminution de volume .

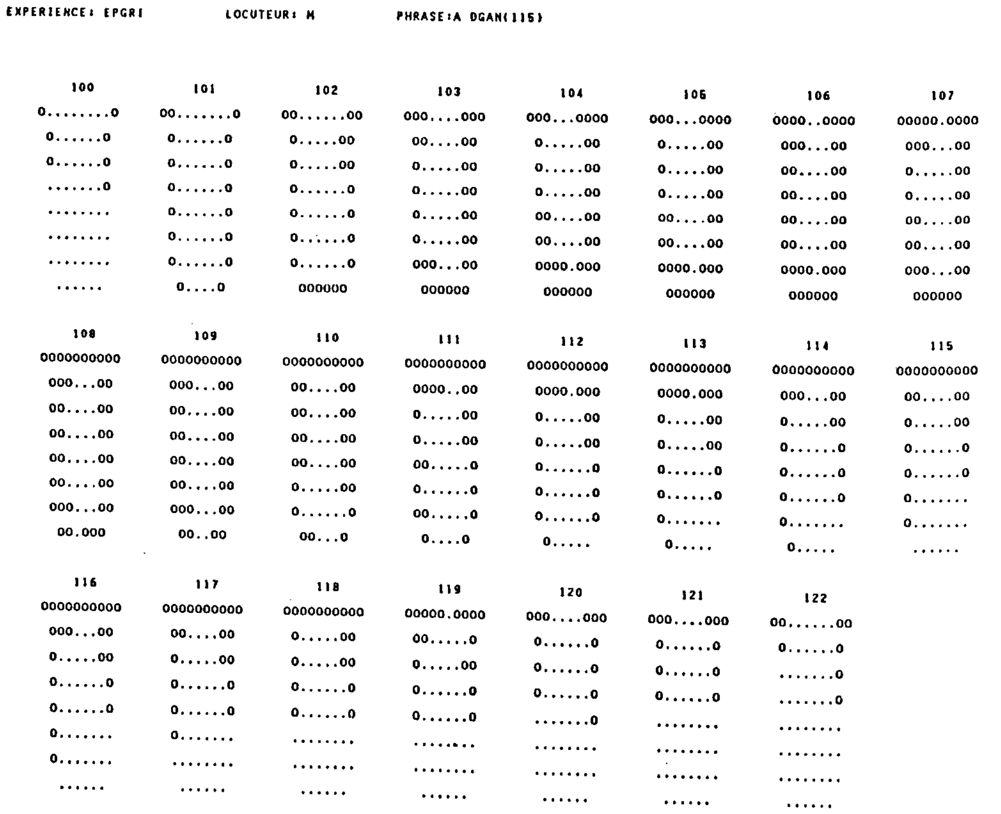
On relève pour l'enchainement de /d/ à /k/ dans /sedki/ une phase de double occlusion d'une durée de 70 ms (images 235 à 241 , Fig . 3) . Les deux occlusions délimitent une cavité dont le volume va varier . A l'implosion de C2 , 47 électrodes sont activées contre 50 au relâchement de C1 . Le renforcement de l'appui de la langue au palais se produit essentiellement dans la zone palatale et entraîne a partir des mesures prises sur un modele en plâtre une diminution de volume supérieure à 3 cm3 . L'élévation de la masse linguale provoque ainsi une augmentation de

```
EXPERIENCE: EPGB1          LOCUTEUR: M          PHRASE:EEDKI ( 88)

    231          232          233          234          235          236          237          238
000....000   000....000   000...0000   00000.0000   0000000000   0000000000   0000000000   0000000000
 00....00     00....00     00...000     000..000     000..000     000..000     000..000     000.0000
 00....00     00....00     00....00     0.....00     000...00     000...00     000...00     000...00
 0.....00     0.....00     00....00     0.....00     00...000     000..000     000..000     000..000
 0.....00     0.....00     00....00     00....00     00....00     00....00     00....00     00....00
 0.....00     0.....00     00....00     00...000     00...000     00...000     00...000     00...000
 0......0     0.....00     00.....0     000...00     0000.000     0000.000     0000.000     000..000
 .......       0....0       00..00       00.000       000000       000000       000000       000000

    239          240          241          242          243          244          245          246
0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000
 000.0000     000.0000     000.0000     00000000     00000000     00000000     00000000     00000000
 000...00     000...00     000...00     000...00     000...00     000...00     000...00     000...00
 000..000     000..000     000..000     000..000     000..000     000..000     000..000     000..000
 00....00     00....00     000...00     000...00     000...00     000...00     00....00     00....00
 00...000     00...000     00...000     00...000     00...000     00....00     00....00     00....00
 000..000     000..000     000..000     000...00     000...00     000...00     0.....00     0.....00
  000000       000000       000000       00..00       00..00       00..00       0...00       0....0

    247          248          249          250          251          252          253          254
0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000
 00000000     00000000     00000000     00000000     00000000     000.0000     000.0000     000..000
 000...00     000...00     000...00     000...00     000...00     000...00     0.....00     00....00
 000..000     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00
 00....00     00....00     00....00     00....00     00....00     0.....00     0.....00     0.....00
 00....00     00....00     0.....00     0.....00     0......0     0......0     0......0     0......0
 0.....00     0......0     0......0     0......0     0......0     0......0     0......0     0......0
  0....0       0....0       0....0       0....0       0....0       0.....       0.....       0.....
```

```
EXPERIENCE: EPGB1          LOCUTEUR: M          PHRASE:A TKU ( 79)

    236          237          238          239          240          241          242          243
00.....000   00.....000   000....000   000....000   000....000   000...0000   0000..0000   00000.0000
 0......0     0......0     0......0     0.....00     00....00     00....00     00....00     00...000
 0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     00....00
 0......0     0......0     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00
 .......0     0......0     0......0     0.....00     0.....00     0.....00     0.....00     0.....00
 ........     .......0     0......0     0......0     0.....00     0.....00     0....000     00...000
 ........     ........     0......0     0......0     0.....00     000..000     000..000     0000.000
  ......       ......       0....0       000000       000000       000000       000000       000000

    244          245          246          247          248          249          250          251
0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000
 00...000     00...000     00...000     00...000     000..000     000.0000     000.0000     000.0000
 00....00     00....00     00....00     00....00     00....00     00....00     00....00     00....00
 0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00
 0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00
 00...000     0....000     0....000     0.....00     0.....00     0.....00     0.....00     0.....00
 000..000     000..000     000..000     00....00     00....00     00....00     0.....00     0......0
  000000       000000       000000       000000       000.00       00..00       00...0       0.....

    252          253          254          255          256          257          258          259
0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000
 000.0000     000.0000     000.0000     000.0000     000..000     000..000     00....00     00....00
 00....00     00....00     00....00     00....00     00....00     0.....00     0.....00     0.....00
 0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0......0     0......0
 0.....00     0.....00     0......0     0......0     0......0     0......0     .......0     .......0
 0......0     0......0     0......0     0.......     ........     ........     ........     ........
 0......0     0.......     ........     ........     ........     ........     ........     ........
  0.....       ......       ......       ......       ......       ......       ......       ......
```



```
EXPERIENCE: EPGRI          LOCUTEUR: M          PHRASE:A KTU ( 96)

    128          129          130          131          132          133          134          135
0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000
 0......0     00.....0     00.....0     00.....0     00....00     00....00     00....00     00....00
 .......0     .......0     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00
 .......0     .......0     .......0     .......0     .......0     0......0     0......0     0......0
 ........     ........     ........     ........     ........     .......0     0......0     0......0
 ........     ........     ........     ........     ........     ........     .......0     .......0
 ........     ........     ........     ........     ........     ........     .......0     .......0
  ......       ......       ......       ......       ......       ......       ......       ......

    136          137          138          139          140          141          142          143
0000000000   0000000000   0000000000   0000000000   0000000000   000.....00   00.......0   00......00
 00.....0     00....00     00.....0     00.....0     00.....0     0......0     0......0     0......0
 0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00     0.....00
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
 0......0     0......0     0.....00     0......0     0......0     0......0     0......0     0......0
  ......       .....0       00..00       000000       000000       000000       000000       000000

    144          145          146          147          148          149          150          151
00.......0   00.......0   00.......0   00.......0   00.......0   00.......0   00.......0   00.......0
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
 0......0     0......0     0......0     0......0     0......0     0......0     0......0     0......0
  000000       000000       000000       000000       000000       000.00       0....0       ......
```



```
EXPERIENCE: EPGRI          LOCUTEUR: M          PHRASE:A DGAN(115)

    100          101          102          103          104          106          106          107
0........0   00.......0   00......00   000....000   000...0000   000...0000   0000..0000   00000.0000
 0......0     0......0     0.....00     00....00     0.....00     0.....00     000...00     000...00
 0......0     0......0     0.....00     0.....00     0.....00     0.....00     00....00     0.....00
 .......0     0......0     0......0     0.....00     0.....00     0.....00     00....00     0.....00
 ........     0......0     0......0     0.....00     00....00     00....00     00....00     00....00
 ........     0......0     0......0     0.....00     00....00     00....00     00....00     00....00
 ........     0......0     0......0     000...00     0000.000     0000.000     0000.000     000...00
  ......       0....0       000000       000000       000000       000000       000000       000000

    108          109          110          111          112          113          114          115
0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000   0000000000
 000...00     000...00     00....00     0000..00     0000.000     0000.000     000...00     00....00
 00....00     00....00     00....00     0.....00     0.....00     0.....00     0.....00     0.....00
 00....00     00....00     00....00     0.....00     0.....00     0.....00     0......0     0......0
 00....00     00....00     00....00     00.....0     0......0     0......0     0......0     0......0
 00....00     00....00     0.....00     0......0     0......0     0......0     0......0     0.......
 000...00     000...00     0......0     00.....0     0......0     0.......     0.......     0.......
  00.000       00..00       00...0       0....0       0.....       0.....       0.....       ......

    116          117          118          119          120          121          122
0000000000   0000000000   0000000000   00000.0000   000....000   000....000   00......00
 000...00     00....00     0.....00     00.....0     0......0     0......0     0......0
 0.....00     0.....00     0.....00     0.....00     0......0     0......0     .......0
 0......0     0......0     0......0     0......0     0......0     0......0     .......0
 0......0     0......0     0......0     0......0     .......0     ........     ........
 0.......     0.......     ........     ........     ........     ........     ........
 0.......     ........     ........     ........     ........     ........     ........
  ......       ......       ......       ......       ......       ......       ......
```

la pression de l'air contenu dans cette cavité et au relâchement de C1 , l'air va se diriger vers l'extérieur de la bouche . Ce courant d'air égressif ne provient pas des poumons puisque l'occlusion vélaire est complète ; il est donc initié par le mécanisme vélique . Le relâchement de C1 s'apparente donc à la production d'un clic inverse (3,4) .

4- Augmentation du volume

La première image de la double tenue (image 244 , Fig . 4) de /t/ et /k/ est caractérisée par l'activation de 42 électrodes . A l'image 247 précédant immédiatement le relâchement de /t/ , on constate que la langue s'est abaissé puisque le nombre de contacts touchés est passé à 38 . Ce mouvement entraine une augmentation de volume d'environ 4 cm3 , et par conséquent une diminution de la pression de l'air contenu dans la cavité délimitée par la double occlusion . Nous avons affaire à la production d'une occlusive à dépression : soit un clic .

CONCLUSION

Dans le cas de l'enchainement d'une occlusive antérieure suivie d'une occlusive postérieure caractérisé par une phase de double occlusion , le relâchement de C1 s'apparente dans de nombreux cas à la production d'un clic ou d'un clic inverse selon l'évolution de l'appui lingual et les variations concomittantes de pression d'air intra-buccale . Ce phénomène a un statut purement articulatoire en Français ; mais il serait intéressant de vérifier dans les langues qui connaissent les clics comme phonèmes si ceux-ci ne sont pas issus de groupes d'occlusives .

BIBLIOGRAPHIE

(1) **Hardcastle W.J.** ; **Roach P.J.** (1977) . : " An instrumental investigation of coarticulation in stop consonant sequences " *Work Prog. Phon. Lab. Univ. Reading* 1 : 27-44 .

(2) **Catford** J.C. (1977) *Fundamental problems in phonetics* . Edinburgh university press Edinburgh .

(3) **Marchal** A. (1985) *L'électropalatographie : contribution à l'etude de la coarticulation dans les groupes d'occlusives* . These doct. d'état , Nancy .

(4) **Ladefoged P.** (1975) . : *A course in phonetics* . Harcourt , Brace Jovanovich , New-york .

THE INFLUENCE OF ASPIRATION ON VOWEL DURATION

SUNIL KUMAR JHA

English Instruction Committee
SMBM Campus, Rajbiraj
Tribhuvan University, Nepal

ABSTRACT

In the present paper an attempt is made to put forward the results of an electro-glottographic study on the influence of aspiration on vowel duration in Maithili — a modern Indo-Aryan language spoken by a total of about 21 million people both in Nepal and India. The main aim of our study was to investigate whether phonation types other than voicelessness and voicing also affect the length of vowels preceding a consonant. Our results clearly show that the aspiration of the following consonant does affect vowel duration in Maithili. In fact, in Maithili the features of both voice and aspiration do independently lend increments of length to the preceding vowel.

INTRODUCTION

There have been in the past quite a few studies on vowel duration in various languages of the world. One of the major findings of most of these studies has been that, other things remaining the same, vowels are longer before voiced consonants than before voiceless ones. This phenomenon has usually been considered [e.g. 1; 2; 3] to be due to an inherent property of the speech production mechanism. And a number of different proposals have so far been made as to what precise mechanism is responsible for this lengthening of vowels. Some proposals [e. g. 4; 5] aim only to account for the lengthening of vowels before voiced and voiceless consonants, while others also aim to account for such factors as: the degree of opening of the vowel [e.g. 6;7; 8]; place, manner and force of articulation [e.g.9; 10; 11] of the following consonants; the structure of the syllable in which the vowel occurs [e.g. 12], the nature of the phonemic contrasts employed by the language in question [e.g. 13], and the degree of glottal opening [e. g. 5; 14] as well as the airflow rate [e. g. 15; 16] of the following consonants. It has to be admitted that comparatively little has so far been published on the effect of aspiration on vowel duration. Relatively recently, Maddieson and Gandour [17] studied the effect of aspiration on the duration of the Hindi vowel /a/ — as spoken in Delhi — and found some inter-action between aspiration and vowel duration. In a later study [18], Maddieson investigated five languages — i.e. Hindi, Bengali, Assamese, Marathi and Eastern American — and came to the conclusion that vowel lengthening before aspirated consonants is not universal.
In the present paper an attempt is made to put forward the results of an electro-glottographic study on the influence of aspiration on vowel duration in Maithili — the vowels studied here are those of a variety of the 'standard' dialect of this language. The main aim of the paper is to investigate whether phonation types other than voicelessness and voicing also affect the length of vowels preceding a consonant.

EXPERIMENTAL METHOD

Test Utterances

For the purpose of the present study, appropriate test utterances — as given in Table I — were prepared. This table lists 24 monosyllabic test utterances containing the following six Maithili oral vowels: /i e a ə o u/, each followed by phono-logically contrasting series of four stop or affricate consonants — i.e. voiceless unaspirated, voiceless aspirated, voiced unaspirated and voiced aspirated. Where complete minimal series of words containing all the stops with differing phonation types in a given place of articulation did not exist, nonsense utterances — i.e. utterances which are not available in the Maithili lexicon and which therefore do not mean anything in this language — were added to fill the gaps in distribution. Only three such nonsense items were required for the purpose of this study: i. e. *[gə:p^h], *[so:g^h] and *[ku:d^h], as given in Table I. It must also be pointed out that all the nonsense utterances thus added in this table are phonologically possible items in the Maithili language.

Table I: Test words containing six Maithili oral vowels followed by voiced and voiceless, aspirated and unaspirated consonants.

Vowels	Words with glosses
/i/	[bi:č] "centre"
	[bi:č^h] "pick up (imp.)"
	[bi:ǰ] "seed"
	[bi:ǰ^h] "rust"
/e/	[se:p] "saliva"
	[se:p^h] "safe (n)"
	[se:b] "to serve"
	[se:b^h] "shave"
/a/	[sa:t] "seven"
	[sa:t^h] "together"
	[sa:d] "longing"
	[sa:d^h] "capacity"
/ə/	[gə:p] "talk"
	*[gə:p^h] (a nonsense word)
	[gə:b] "seedlings made ready for transplantation"
	[gə:b^h] "pregnancy — a meta-phorical use"
/o/	[so:k] "sorrow"
	[so:k^h] "swallow"
	[so:g] "distress"
	*[so:g^h] (a nonsense word)
/u/	[ku:t] "amount of grain given by tenants to landlords"
	[ku:t^h] "push breath out of lungs"
	[ku:d] "jump"
	*[ku:d^h] (a nonsense word)

Apparatus Used

Each test utterance was afterwards put in a normal conversational sentence context, the frame of the sentence being ['p^{h}e:ro . . . 'uččа:rəṇ kə'ru:] "please . . . pronounce again". Each test utterance was said in the same frame so as to make sure that the differences are not due to variations in the rate of utterance. The sentences were first randomised and then spoken in a relaxed informal style at a normal conversational speed, without putting any contrastive stress on the test utterances. The pronunciation represented in this work is entirely the author's own. Sixteen tokens of each test utterance, each token embedded in the above sentence frame, were recorded in a soundproof studio of Essex University. All recordings were made on a Revox B 77 tape-recorder. The glottal signal was obtained using an Electroglottograph F-J Electronics Type EG 830. Oscillomink tracings of waveform and amplitude produced from the recorded readings were obtained using a Mingograf Type EM 34T. Calculations relating to the 'mean', the 'standard deviation'(SD) and the 'coefficient of variation'(v) of all tokens of each test utterance were made using a Tektronix 31 calculator.

Duration Measurements

Of the sixteen tokens of each test utterance, the first two as well as the last two tokens were ignored, and all the remaining twelve tokens of the middle were used to obtain the duration measurements of all the vowels investigated in this study. The first measurements of vowel duration were made from the start of the vowel in question to the closure of the following consonant. In the case of words beginning with voiced stops and even voiceless unaspirated stops and fricatives, the measurement was begun at the release of the concerned initial stop or fricative.
Afterwards, a simple arithmetic mean of the actual measured values of all the 12 tokens of each test utterance was worked out. In order to ascertain the reliability of the arithmetic mean as a quantified abstract value representing the realisation of the speaker's intention, the range of the variability occurring in all the 12 tokens of every test utterance was also taken into account. For this, the standard deviation of each test utterance was worked out. To relate the variation between the different sets of data presented in this paper, the duration values of all test utterances were normalised by obtaining a coefficient of variation of each test utterance, the equation used being: $v = \frac{SD}{mean} \times 100$.

RESULTS AND DISCUSSIONS

Since from a preliminary survey of some published sources [e.g. 19; 13; 10; 8; 6; 7] preceding consonants exhibit no readily discernible patterns of environmental influence on the duration of the following vowels, in the present study we have restricted ourselves to the influence of the following consonants on the duration of preceding vowels. Table II presents the results of this study [see 20, pp. 344-45, for more details]. It shows the mean duration values of the six oral vowels as obtained from the 12 tokens of each test utterance, the standard deviation of the 12 tokens of each test utterance, the coefficient of variation

Table II: Mean duration values, standard deviation, coefficient of variation, and the duration-ratio of the six Maithili oral vowels followed by voiced and voiceless, aspirated and unaspirated consonants.

Vowel	Word	Mean	SD	v	Ratio
/i/	[bi:č]	165	5	3	1.00
	[bi:č^h]	183	5	3	1.15
	[bi:ǰ]	210	5	3	1.27
	[bi:ǰ^h]	224	5	2	1.35
/e/	[se:p]	174	5	3	1.00
	[se:p^h]	195	5	3	1.12
	[se:b]	218	4	2	1.25
	[se:b^h]	240	5	2	1.37
/a/	[sa:t]	202	6	3	1.00
	[sa:t^h]	226	7	3	1.11
	[sa:d]	240	5	2	1.18
	[sa:d^h]	275	6	2	1.36
/ə/	[gə:ɔ]	104	4	4	1.00
	*[gə:p^h]	120	5	4	1.15
	[gə:b]	134	6	5	1.28
	[gə:b^h]	159	4	3	1.52
/o/	[so:k]	166	5	3	1.00
	[so:k^h]	180	8	4	1.08
	[so:g]	204	8	4	1.22
	*[sc:g^h]	239	6	3	1.43
/u/	[ku:t]	155	5	3	1.00
	[ku:t^h]	175	4	2	1.12
	[ku:d]	220	5	2	1.41
	*[ku:d^h]	240	5	2	1.54

of every utterance as well as the ratio of the duration of vowels preceding voiced and aspirated consonants to the duration of vowels preceding voiceless unaspirated consonants. A diagrammatic representation of the mean duration values of this table is given in Figure 1. The horizontal axis of this figure shows postvocalic stop and affricate consonants of various places of articulation, while its vertical axis shows the duration of the six oral vowels in milliseconds (ms).

Both Table II and Figure 1 clearly show that all the six vowels investigated in this study have longer mean durations before voiced and aspirated stop as well as affricate consonants than before their voiceless unaspirated counterparts. The mean duration-measurements and their diagrammatic illustrations given in Table II and Figure 1, respectively, amply show that the aspiration of the following consonant does affect vowel duration in Maithili. The present data sufficiently reveals that the overall pattern found between the relative durations of the six Maithili oral vowels — each vowel preceding consonants of four different phonation types — is very similar, and that in this language:

1. vowels are relatively longer in duration before voiceless aspirated consonants than before voiceless unaspirated consonants;
2. vowels preceding voiced unaspirated consonants are relatively longer in duration than those preceding either voiceless unaspirated or voiceless aspirated consonants;
3. vowels are relatively longer in duration before voiced aspirated consonants than before voiced unaspirated consonants; and
4. in general, other things being equal, open vowels are relatively longer than close vowels.

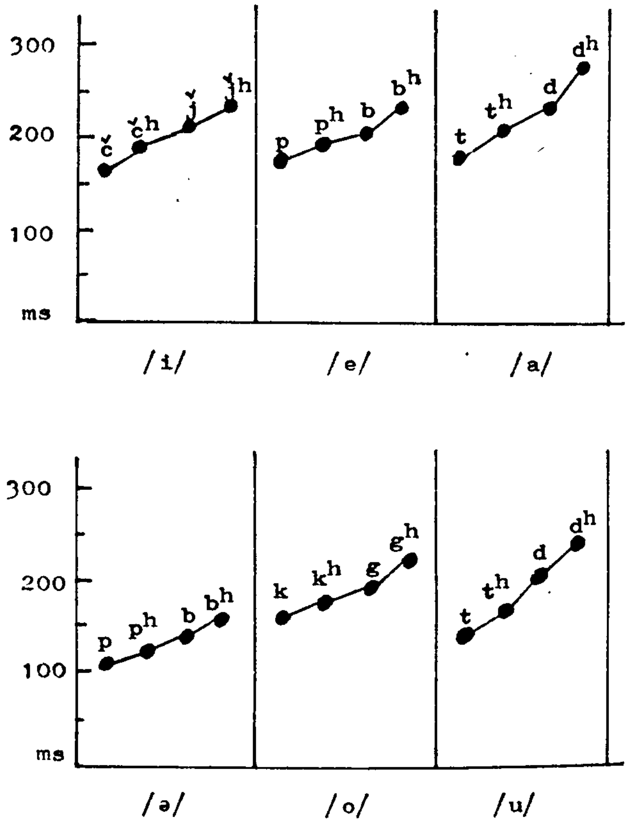


Figure 1: The mean duration of the Maithili oral vowels as spoken in monosyllabic minimal word pairs, each word of the pair differing only in the final consonant.

Our results of the present study suggest that two rules, perhaps 'low-level phonetic rules' operate in Maithili. These may be written as in (1) below:

(1) a. vowel adds 1 increment of length before aspiration; and
b. vowel adds 2 increments of length before voicing.

Applying these rules gives the results shown in (2) below:

(2)	
vowel before voiceless unaspirates:	0 increment
vowel before voiceless aspirates:	1 increment
vowel before voiced unaspirates:	2 increments
vowel before voiced aspirates:	3 increments

These findings offer support for the traditional grouping of the Maithili obstruents — like perhaps the obstruents of most other Indo-Aryan languages — not only into voiced and voiceless categories but also into aspirated and unaspirated. The most interesting aspect of our results is the challenge presented to the various proposed 'explanations' [e.g. 4; 5; 9; 11; 21; 22; 23; 19] of the cause of the intrinsic length of vowels before consonants of different phonation types [see 20, pp. 163-67, for more discussions in this respect].

CONCLUSION

To conclude, the present study clearly shows that phonation types other than voicelessness and voicing also affect the relative duration of the vowel preceding a consonant. We have found that the features of both voice and aspiration independently lend increments of length to the preceding vowels in Maithili. This clearly means that the 'explanations' proposed in the literature so far to account for vowel lengthening before voiced and voiceless obstruents cannot be extended to account also for vowel lengthening before both voiceless and voiced aspirated obstruents. We therefore hope that the results of our study will urge a rethinking of recent and current explanations of the interaction of phonation type and vowel duration, and will assist in the formulation of new theories predicting the influence of the following aspirated/unaspirated consonants on the relative duration of preceding vowels.

References

[1] H. Hollien, "On vowel registers", Journal of Phonetics, Vol.2 (1974), pp. 125-143.

[2] W.J.Hardcastle, The Physiology of Speech Production, Academic Press,1976.

[3] J.Laver, The Phonetic Description of Voice Quality, CUP, 1980.

[4] N.Chomsky & M.Halle,The Sound Pattern of English, Harper & Row, 1968.

[5] M.Chen, "Vowel length variation as a function of voicing of the consonant environment", Phonetica, Vol.22 (1970) pp. 129-159.

[6] A.S.House & G.Fairbanks, "The influence of consonant environment upon the secondary acoustical characteristics of vowels", Journal of the Acoustical Society of America, Vol. 25 (1953), pp. 105-113.

[7] G.E.Peterson & I.Lehiste, "Duration of syllable nuclei in English", Journal of the Acoustical Society of America, Vol.32 (1960), pp. 693-703.

[8] A.S.House, "On vowel duration in English", Journal of the Acoustical Society of America, Vol.33 (1961), pp. 1174-1177.

[9] S.Belasco, "The influence of force of articulation of consonants on vowel duration", Journal of the Acoustical Society of America, Vol. 25 (1953), pp. 1015-1016.

[10] E.Fischer-Jørgensen, "Sound duration and the place of articulation", Zeitschrift fur Phonetik Sprachwissenschaft und Kommunikationsforchung, Vol. 17 (1964), pp. 175-207.

[11] B.Mohr, "Intrinsic variations in the speech signal", Phonetica, Vol. 23 (1971), pp. 65-93.

[12] T.Balasubramanian, "Duration of vowels in Tamil", Journal of Phonetics, Vol. 9 (1981), pp.151-161.

[13] S.A.Zimmerman & S.M.Sapon, "Note on vowel duration seen cross-linguistically", Journal of the Acoustical Society of America, Vol. 30 (1958), pp. 152-153.

[14] F.Ingemann & R.Yadav, "Voiced aspirated consonants", In Papers from the 1977 Mid-American Linguistics Conference (D.M.Lance & D.E.Gulstad, editors), pp. 337-344. University of Missouri, 1978.

[15] C-W.Kim, "On the autonomy of the tensity feature in stop classification", Word, Vol.21 (1965), pp. 339-359.

[16] P.Nihalani, "Air flow rate in the production of stops in Sindhi", Phonetica, Vol.31(1975), pp. 198-205.

[17] I.Maddieson & J.Gandour, "Vowel length before aspirated consonants", UCLA Working Papers in Phonetics, Vol. 31 (1976), pp. 46-52.

[18] I.Maddieson, "Further studies on vowel length before aspirated consonants", UCLA Working Papers in Phonetics, Vol.38(1977), pp. 82-90.

[19] P.Delattre, "Some factors of vowel duration and their cross-linguistic validity", Journal of the Acoustical

Society of America, Vol.34 (1962), pp. 1141-1143.

[20] S.K.Jha, "A Study of Some Phonetic and Phonological Aspects of Maithili", Unpublished PhD thesis, University of Essex, U.K., 1984.

[21] M.Halle & K.N.Stevens, "On the mechanism of glottal vibration for vowels and consonants", Quarterly Progress Report, No. 101 (1967), pp. 198-213, Research Laboratory Electronics, MIT.

[22] I.H.Slis & A.Cohen, "On the complex regulating the voiced-voiceless distinction 1", Language and Speech, Vol.12 (1969), pp. 80-102.

[23] I.Lehiste, "Temporal organization of spoken language", Working Papers in Linguistics (Ohio State University), Vol.4 (1970), pp. 96-114.

THE MAIN CUES DIFFERENTIATING ASPIRATED AND UNASPIRATED STOPS AND AFFRICATES IN ARMENIAN

AMALIA KHACHATRYAN

Institute of Linguistics
Armenian Academy of Sciences
Yerevan, USSR, 375001

ALBERT AIRAPETYAN

Depart. of Radioengineering
Polytechnical Institute
Yerevan, USSR, 375009

The purpose of this paper is to further examine the nature of aspirated and unaspirated stops and affricates of Armenian presenting new data and re-evaluating the VOT as the only cue for differentiating these two categories of sounds. Generalized VOT + K I parameter is suggested for reliable distinction of both groups of sounds.

INTRODUCTION

During the last two decades aspiration has attracted the interest of phoneticians in many countries. This is partly due to the use new methods of articulatory investigation, such as electromyography, glottography, fiberscoping alongside with the more traditional acoustic ones, such as spectrography, oscillography.

This increase of interest is partly upheld by the cross-language study of aspiration in stops carried out by L.Lisker and A.Abramson, who have suggested a new cue - voice onset time (VOT) for discriminating the three categories of stops - voiced, voiceless and aspirated [I]. Aspiration has been studied from different aspects: its theory [2], mechanisms of production [3], its glottal and supraglottal articulation timing [4], its relationship to other phonetic features, such as fortis/lenis [5], acoustic expression and perception.

It has been mentioned that in such languages as English and Swedish, aspiration being the expression of lenis/ fortis feature is concomitant and differentiates voiced and voiceless stops. In Danish it is the only distinguishing feature between the sets ptk and bdg.

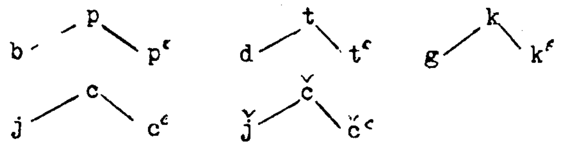
It is worth mentioning that the first experiments in voice timing in stops were carried out by H. Adjarian at the Rousselot laboratory as far back as I898 [7].His kymographic tracings were published in the journal "Revue internationale de Rhinologie, Otologie, Laryngologie et Phonétique expérimentale" in I899 and were furnished with Rousselot's commentaries. Yet the purpose pursued by the author was somewhat different from nowadays studies. Adjarian intended to show thegradual development of devoicing (lénition) which in the long run brought to the shift of voiced stops and affricates of Old Armenian into voiceless ones in many modern dialects and vice versa. Thus Adjarian paid attention to the fact, that in some dialects the voicing of b,d,g,j,ǰ may lag a little, in others-still more, whereas in some others - too much, which has brought to the shift of voiced consonants into voiceless aspirated ones. Actually, from the phonological point of view this difference is one of the maijor phonetic differences existing between Eastern and Western dialects of Armenian as well as between the two literary variants.

A new impulse to the study of this feature was given by Leigh Lisker and Arthur Abramson, who carried out cross-language experiments and a more detailed examination of this variable terming it VOT. This cue proved to be valid to distinguish the three main categories of stops: a)those with voice lead (fully voiced b d g), b) with short-lag voice (voiceless p t k) and c) with long-lag voicing (aspirated p t k). The authors showed the validity of this variable as compared to fortis/lenis or voiced/voiceless features used by most linguists. Of particular interest for us is that the author's investigation included the data concerning Eastern Armenian literary language.

Our first spectrographic experiments in consonants of Armenian were carried out in late 60-ies in Tallinn at the Institute of Language and Literature. Even the limited data we got brought us to the conclusion that the distinction between aspirated and unaspirated stops of Armenian was bound primarily to the duration of release, which later on came to be known as voice onset time - VOT. But since our primary pupose was to prove the monophonemic nature of aspirated stops as opposed to the traditional view that considered them compound ones consisting of one stop and the glottal fricative (h), we were interested in other features as well, particularly the timing relationship between the closure and release and the total durationof both series of consonants (6). A reverse relationship between closure duraticn and release time has been established.

LINGUISTIC MATERIAL

Unlike English and some WesternEuropean languages, in which aspiration is a redundant feature, in Eastern Literary Armenian it is an independent distinctive feature differentiating homorganic stops and affricates. Aspirated sounds form oppositional pairs with unaspirated voiceless cognates in all positions in monosyllabic and disyllabic words. Thus the Armenian stops and affricates can be presented in the following way:



In some dialects the fourth series has been claimed to exist by some linguists, but these sounds have been proved to have no phonological value, being only allophonic by nature and quite distinct from aspirated voiced sounds of Hindi.

In the first series of experiments monosyllabic or bisyllabic words with aspirated stops and affricates were used. They were presented to the speakers either in oppositional pairs or independently embedded in the carrier sentence "Sa ...e" 'This is ....'. The list of words included such words as payt - pʿayt 'horseshoe' - 'wood', taṙ - tʿaṙ 'letter'-'stack' , akama - akʿaɣaɣ 'unintentional'-'cock', cec - cʿecʿ 'beat' - 'moth', čanč - čʿanč 'fly' - 'palm', etc.

In the second series of experiments we chose words in which the phonemes under examination were in most unfavour - able conditions for the realization of aspiration, such as unstressed syllables, words, in which aspirated stops were preceded by fricatives (s) or (š), in words with two or three aspirated sounds. The stimuli were pronounced at a normal rate.

METHOD

A computor integrated distinctive feature analysing device was used with subsequent segmentation of speech signal into elementary segments corresponding to speech sounds. The accuracy of segmentation was being controlled visually on the screen of the display. The time quatization is equal to 10 milliseconds (ms). The release time (VOT) of stops was measured according to sampling intervals. The intensity of the noise above 2000 Hz was measured on logarythmic scale.

6 male and 4 female native untrained speakers of Armenian served as subjects for this experiment. The stimuli were read into microphone being directly put into the analyser.

RESULTS

We were mainly concerned with differences of VOT and intensity of noise in aspirated and unaspirated stops and affricates. Though we did not pay special attention to the duration of burst, but certain differences which relate to place of articulation will also be mentioned. Thus the duration of burst in labial plosives is 5 - 10 ms, in dentals - 15-20 ms, and in velars - 30-35 ms. In aspirated cognates this burst is followed by noise of considerable duration, which varies in different phonetic contexts. Thus it is the longest in the final position, it is very strong and long initially and short in the middle position.

In fig. I the scatter diagram of VOT values on a single timeline is presented for aspirated and unaspirated stops. It shows that VOT as a whole differentiates between the two categories of stops.There is some overlapping of ranges in velar k - kʿ stops, which hampers the reliable distinction of these stops - a matter of no less importance in automatic segmentation of speech sounds. From this point of view the overlap of the ranges of VOT in aspirated and unaspirated affricates (in Armenian linguistic tradition terms) is very typical. In fig 2 the scatter diagram of VOT in affricates shows the degree of overlap. It is quite evident that VOT alone is not sufficient for

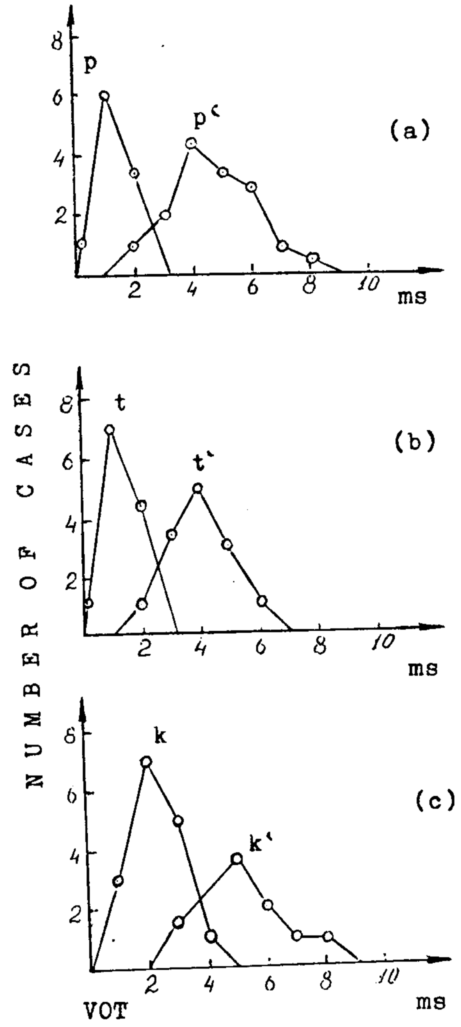


Fig. I The scatter diagram of VOT on a single timeline for unaspirated and aspirated stops (a) labials p - p ; (b) dentals t - t; (c) velars k - k.

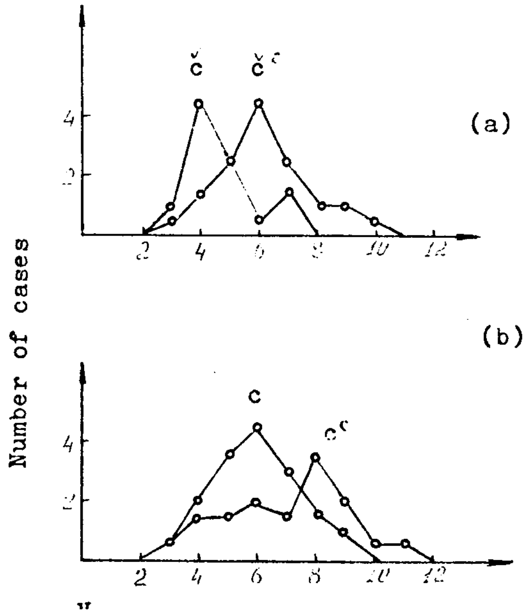


Fig.2. The scatter diagram of VOT on a single timeline for unaspirated and aspirated affricates: apical c - cʿ(a), palato-alveolar č - čʿ(b).

their distinction. For reliable discrimination of these consonant categories an additional parameter is necessary. The intensity of friction noise which characterizes the release of stops and affricates may serve as such a parameter. In fig. 3 a two-dimensional scatter diagram of VOT and Intensity (I) for the aspirated and unaspirated affricates is plotted. The dots present unaspirated stops and the circles - corresponding aspirated ones. Though VOT may serve as a cue for differentiation of these categories of stops,ptk and pʿtʿkʿ, which corresponds to the division of the planes by the line parallel to the abscissa axis, there is a noticeable correlation between VOT and intensity,and an oblique line separating them increases the reliablity of differentiation.

If in the case of aspirated and unaspirated stops the intensity cue is redundant or additional, in the case of affricates it is indispensable, as important

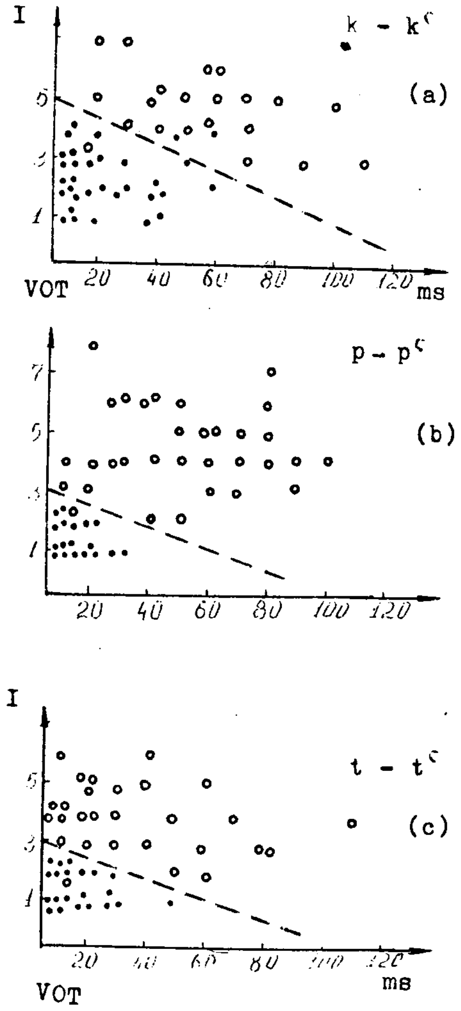


Fig. 3. A two-dimensional scatter diagram of VOT measurements and intensity values of stops : k - kʿ(a), p - pʿ (b) and t - tʿ(c) .The dots indicate unaspirated sounds, circles - aspirated ones.

as VOT. In fig. 4 a two-dimensional plot distribution of VOT and intensity values of aspirated and unaspirated affricates is given. There is a clear-cut correlation with linear regression of VOT and intensity with the corresponding line separating the two areas. The equation of this correlation is as follows:

$$VOT + K\ I = C$$

where K and C are scale coefficients defined by the sensitiveness and dynamic range of intensity measurement channel. In our experiments K= 0,05, and C= 6. The parameter VOT + KI may be considered as generalized cue characterizing the degree of aspiration in aspirated stops and affricates.

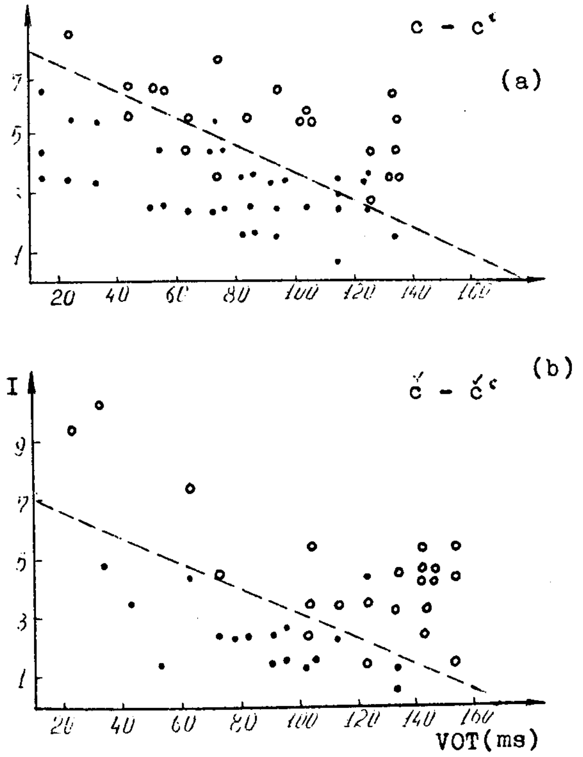


Fig. 4. Two-dimensional scatter-diagram of VOT and intensity values of affricates c - cʿ(a) and č - čʿ(b). The dots indicate unaspirated sounds, and circles - aspirated ones.

CONCIUSION

It is obvious that acoustic differences of aspirated and unaspirated stops are bound primarily with the parameter VOT. Among different local stops VOT is most valid for differentiating labials p -pʿ and dentals t - tʿ. It is somewhat less valid for k - kʿ pair. For the affricates the variable VOT is an ambiguous cue for discriminating the voiceless and aspirated categories. For differentiating them the parameter I (intensity of noise above2000Hz is as important as VOT. Moreover there is an essential correlation between both cues, and the use of only one of them is not sufficient for their differentiation. It has been shown that a reliable accuracy of discrimination between aspirated and unaspirated stops and affricates is ensured by the generalized VOT + KI cue. The voiced correlates of both stops and affricates are distinct, since in their production F_o is mostly present.

REFERENCES

I. L.Lisker and A.Abramson: A Cross-Language Study of Voicing in Initial Stops: Acoustic measurements, Word 20,384-422 (1964).

2. Chin-Wu)Kim: Theory of Aspiration. Phonetica 2I, I07-II6 (1970).

3. M. Halle and K. Stevens; A Note on Laryngeal Feature, MIT Quart.Prog.Rep. Res. IX. Speech Communication (I97I).

4.B.Hutters: Vocal Fold Adjustments in Danish Stops. Phonetica 42, I-24 (1985).

5. Р.Якобсон, Г.Фант и М.Халле. Введение в анализ речи. В кн.: Новое в лингвистике, вып. II, М., Изд. иностр. литературы, 1962, с. 207-208.

6.А.А.Хачатрян и В.Н.Айрапетян. Экспериментальное исследование согласных фонем современного армянского языка. Ереван, 197I.

7. H.Adjarian, Les explcsives de l'ancien Arménien étudiées dans les dialects modernes. Revue internationale de Rhinologie, Otologie, Laryngologie et Phonétique expérimentale, I899, Paris.

PALATALIZATION EFFECTS AND DEGREES OF ARTICULATORY CONSTRAINT IN TWO CATALAN DIALECTS

DANIEL RECASENS

Department of Catalan Philology
Universitat Autònoma de Barcelona, Bellaterra
SPAIN

ABSTRACT

Data reported in this paper suggest that the phonetic output of a phonological rule may depend on small but systematic articulatory differences for the same phoneme. Thus, the application of a progressive assimilation rule for the phonemic cluster /ʎs/ in Catalan is conditioned by the degree of palatal constriction for /ʎ/: the phonetic realization is [ʎʃ] in dialects showing a high degree of palatal constriction for /ʎ/ and [ʎs] in dialects showing a lower degree of palatal constriction for the same palatal consonant.

INTRODUCTION

It has been pointed out that the phonetic realization of a given phoneme may show systematic differences from one dialect to another. Thus, a higher F2 for [w] and a lower F2 for [j] in Zuni vs Amharic and Yoruba suggest that the two approximants ought to be produced with a less constrained gesture in Zuni than in Amharic and Yoruba (Maddieson and Emmorey [3]). Moreover, such articulatory differences may be related to contrasting degrees of coarticulatory resistance. Indeed, according to the data of Maddieson and Emmorey, Zuni semivowels appear to be less resistant than those of Amharic and Yoruba to coarticulatory effects from the adjacent vowels.

In the light of these observations, it is plausible to hypothesize that small articulatory differences for the same production gesture may have an effect on the phonetic output of given phonological processes. The validity of this claim will be tested with reference to the presence vs absence of a progressive assimilation rule changing /s/ into [ʃ] after alveolopalatal [ʎ] in Catalan. Catalan dialects A (spoken in the Girona region) and B (spoken in the Tarragona, Lleida and València regions) differ as to the availability of the phonological rule; thus, the rule applies in dialect B but not in dialect A, as indicated by the fact that the realization of /ʎ s/ is [ʎs] in dialect A and [ʎ(ƫ)ʃ] in dialect B. It can be suggested that the presence vs absence of progressive assimilation in Catalan dialects is related to two possible context-independent factors. A possible conditioning factor would be the palatalized nature of /s/ in dialect B (i.e., [ᶊ]) vs dialect A (i.e., [s]); in that case, an increase in the degree of palatal constriction for /s/ after a palatal consonant would result in alveolopalatal [ʃ] in dialect B and palatalized alveolar [ᶊ] in dialect A. An alternative factor may be that alveolopalatal /ʎ/ is produced with a higher degree of linguopalatal contact in dialect B than in dialect A; in that case, the change of /s/ into [ʃ] would be dependent on the degree of palatal constriction for the preceding [ʎ]. The purpose of the research reported in this paper is to find out whether dialects A and B differ with respect to the degree of palatality for /s/ and /ʎ/. If so, it follows that the progressive assimilation rule involving the feature palatal may be associated with small but systematic cross-dialect differences in the execution of the tongue-dorsum raising gesture towards the palatal region.

METHOD

Possible differences in the degree of palatal constriction for /s/ and /ʎ/ were inferred from acoustic measurements in VCV sequences. The two consonants /ʎ/ and /s/ were uttered in symmetrical and asymmetrical VCV sequences for V= /i/ and /a/. All sequences were preceded and followed by [t] in the Catalan carrier sentence Digues _____ sempre ("Say _____ always"). The recording material was repeated ten times by two speakers of dialect A (Pi, Ca) and two speakers of dialect B (Re, Ba) in a sound-proof room. Speecg data were digitized at a sampling rate of 10 kHz for acoustical analysis. Spectral analysis was performed with a Brüel and Kjaer 2033 spectrum analyzer.

Measurements for /s/ were based on frequency readings of the first spectral maximum at the midpoint of the fricative noise. Data were interpreted on the grounds that an increase in the degree of palatal constriction for the fricative causes a decrease in formant frequency values; according to acoustic theory of speech production, such a decrease is mainly due to an increase in front cavity size as the tongue dorsum is raised (Heinz and Stevens [2]).

Formant measurements for /ʎ/ were taken at the midpoint of the consonantal period. Data on F2 were collected on the grounds that, for palatal articulations, F2 frequency varies directly with the degree of palatal constriction (Fant [1]). A comparison of F2 frequency values across vowel contexts for each speaker should provide useful information about changes in palatal constriction and degree of coarticulatory resistance for the consonant. F3 readings are not given due to the fact that this formant was often cancelled or attenuated in the vicinity of a spectral zero.

Another measurement of the degree of palatal constriction for /ʎ/ was inferred from data on C-to-V coarticulation. Values for F2 and F3 of V1 and V2 were taken into consideration since, for /i/ and /a/, the two formants are inversely related to changes in front cavity size and directly related to changes in the degree of tongue-dorsum raising (Fant [1]). First, V1 and V2 formant frequencies were taken at the vowel midpoint, separately for the sequences /VsV/ and /VʎV/. Then, mean frequency values for a vowel adjacent to /ʎ/ were substracted from mean frequency values for the same vowel adjacent to /s/. Differences between vowel formant values in the contexts /VʎV/ and /VsV/ across speakers were considered to reflect cross-speaker differences in the degree of palatal constriction for /ʎ/, in line with the fact that, as shown in the Results section, the phonetic realization of /s/ was found to be highly analogous for dialects A and B. Thus, it was predicted that a higher degree of palatality for /ʎ/ ought to cause a larger departure from the F2 and F3 vowel frequencies in the context /VsV/.

RESULTS

Degree of palatality for /s/

Data on the frequencies for the /s/ spectral maximum are shown on Figure 1 for all speakers. They are highly consistent with data reported in Recasens [4] showing a first high amplitude spectral peak at about 4000 Hz. Cross-dialect differences are neglegible and inconsistent with the originary hypothesis that /s/ should be more palatal in dialect B than in dialect A. Were there a contrast, the /s/ peak in dialect B would presumably approach the first aplitude spectral peak for /ʃ/ which lies around 3000 Hz (Recasens [4]). Therefore, for the speakers chosen in this study, the claim that the presence vs absence of progressive assimilation in the /ʎ s/ sequence is associated with differences in the degree of palatality for /s/ must be rejected.

Degree of palatality for /ʎ/.

Figure 2 shows changes in F2 across VCV contexts for all speakers. According to the figure, F2 of /ʎ/ increases with adjacent /i/ vs /a/, more so for speakers of dialect B than for speakers of dialect A. Thus, it can be suggested that the palatal gesture for /ʎ/ is more constrained in dialect B than in dialect A when the consonant is adjacent to a high front vowel.

Figure 3 shows F2 and F3 frequency differences for /i/ in the symmetrical sequences /iʎi/ vs /isi/. Figure 4 shows F2 and F3 frequency differences for /a/ in the symmetrical sequences /aʎa/ vs /asa/. In both figures, data are plotted separately for each speaker, each dialect, and anticipatory (C-to-V1) vs carryover (C-to-V2) effects. Of all formant frequency differences plotted in the figure, those exceeding 50 Hz were found to be significant at the $p < 0.05$ or $p < 0.01$ levels. Overall, C-to-V coarticulatory effects in /iCi/ sequences are larger for speakers of dialect B than for speakers of dialect A. This is particularly the case for speaker Re who, contrary to the other three speakers, shows larger C-to-V effects when V= /i/ than when V= /a/. C-to-V data for V= /a/ in Figure 4 does not allow stating any contrasting coarticulatory trend between the two dialects.

CONCLUSIONS

Data on V-to-/ʎ/ and /ʎ/-to-V effects reported in this paper suggest that dialects A and B of Catalan differ as to the degree of palatal constriction during the production of the entire /iʎ i/ gesture. It may be that the same contrasting production strategy takes place for /ʎ/ in the vicinity of other high front articulations. Therefore, it is plausible to maintain the view that the presence vs absence of the progressive assimilation rule /s/→ [ʃ]/ [ʎ] _____ in Catalan is dependent on contrasting degrees of palatality for /ʎ/.

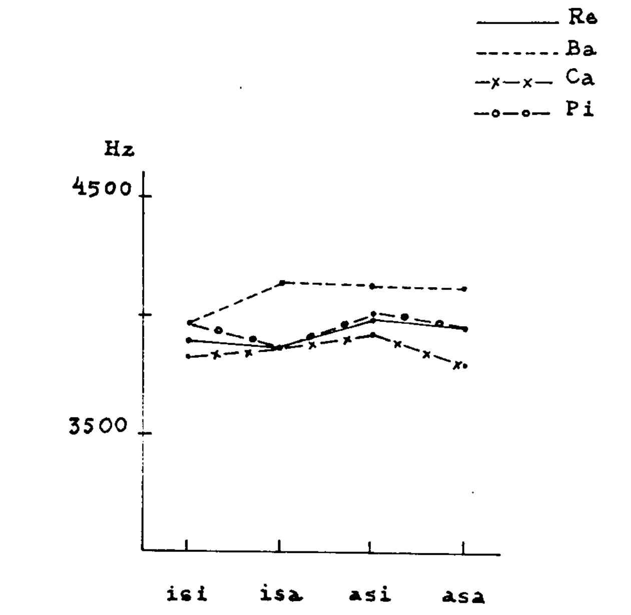


Figure 1. Frequency values for the first spectral maximum of /s/ as a function of vowel context. Data are plotted separately for speakers Re, Ba (dialect B), Pi and Ca (dialect A).

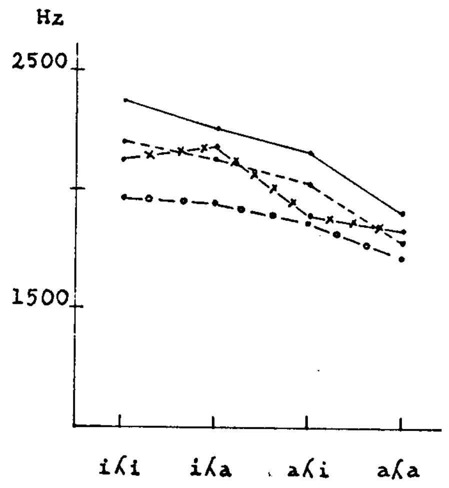


Figure 2. F2 frequency values for /ʎ/ as a function of vowel context. Data are plotted separately for speakers Re, Ba (dialect B), Pi and Ca (dialect A).

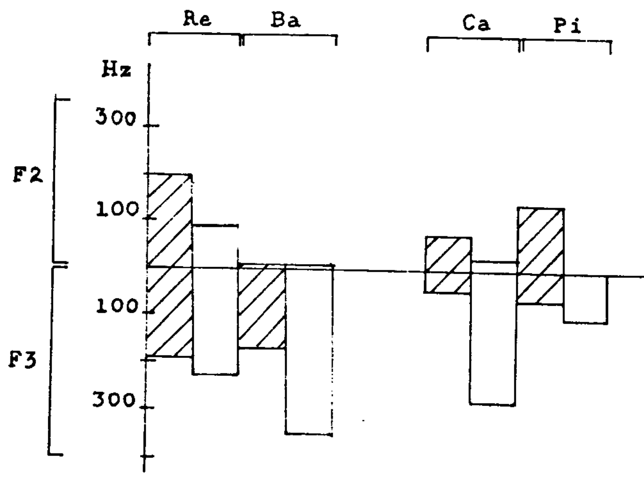


Figure 3. Differences in F2 and F3 frequency values at V1 (C-to-V anticipatory effects; solid bars) and V2 (C-to-V carryover effects; white bars) between /iʎi/ and /isi/. Data are plotted separately for speakers Re and Ba (dialect B), and Ca and Pi (dialect A).

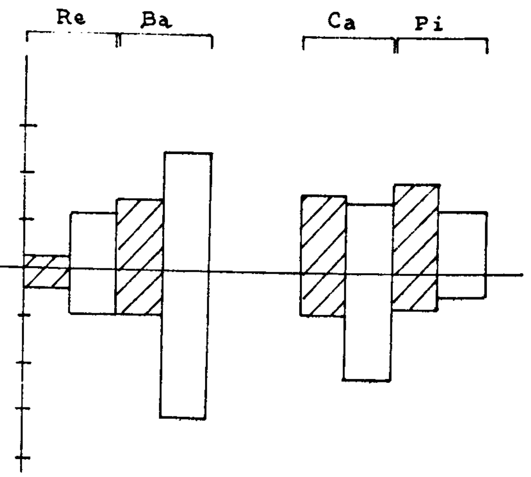


Figure 4. Differences in F2 and F3 frequency values at V1 (C-to-V anticipatory effects; solid bars) and V2 (C-to-V carryover effects; white bars) between /aʎa/ and /asa/. Data are plotted separately for speakers Re and Ba (dialect B), and Ca and Pi (dialect A).

It is believed that to account for these phonological facts, the phonemes /ʎ/ and /s/ should be specified for degrees of the feature palatal. Thus, /ʎ/ would be [1 palatal] in dialect A and [2 palatal] in dialect B, in line with contrasting degrees of the tongue-dorsum raising gesture. On the other hand, /s/ would be [- palatal] for speakers of dialects A and B in the present study, but possibly [1 palatal] for other speakers of dialect B. Progressive assimilation for Catalan /ʎs/ clusters would only apply in the following cases: (1) in dialect B, when C1 and C2 differ sufficiently in degree of palatality, as for /ʎ/ being [2 palatal] and /s/ being [- palatal] ; (2) possibly in dialect B as well, when C1 and C2 agree (entirely or partly) in degree of palatality, as for /ʎ/ being [1 palatal] or [2 palatal] and /s/ being [1 palatal] .

REFERENCES

[1] Fant, G. 1960 Acoustic Theory of Speech Production, The Hague.

[2] Heinz, J.M. and K.N. Stevens 1961 On the properties of voiceless fricative consonants, Journal of the Acoustical Society of America, 33, 5, 589-596.

[3] Maddieson, I. and K. Emmorey 1985 Relationship between semivowels and vowels: Cross-linguistic investigations of acoustic difference and coarticulation, Phonetica, 42, 4, 163-174.

[4] Recasens, D. 1968 Estudis de Fonètica Experimental del Català Oriental Central, Barcelona.

BALKAN-ROMANCE PARALLELS IN DISTRIBUTION OF PHONEMES

JACEK PERLIN

Dept. of Romance Philology
Warsaw University, POLAND

IRENA SAWICKA

Inst. of Slavic Studies
Polish Acad. of Sciences
Warsaw, POLAND

In our work concerning phonetic balkanisms we concentrated on the distribution of sounds. Here we present some conclusions resulting from a comparison of the distributional characteristics of segments which are not motivated by the direct context, but which are due to the position of segments in the syllable and in the word. Our investigation revealed the occurence of certain specific features in microregions extending beyond the territory of the Balkan Sprachbund. This caused the necessity of widening the scope of our study to include Romance material. Apart from Balkan and Romance languages, Serbo-Croatian and Turkish material has been taken into consideration.

Among Balkan languages, and generally in most European languages, certain similarities and common tendencies can be observed, while differences do not exceed certain limits. The similarities concern the phenomenon which could be called approximization to the symmetrical and sonorous syllable pattern. However, this should be treated neither as a Balkan feature nor as a universal tendency. By sonorous syllable pattern we understand here a pattern in which distribution of segments is based on the principle of increasing inherent loudness of sounds before the syllable peak, and falling loudness of segments after the peak. In languages in question this is reflected in the order of sonorants /S/ and obstruents /O/ in consonant clusters. In the sonorous syllable pattern the sonorant must stand neither between two obstruents, nor between an obstruent and a juncture. In such positions it has to undergo syllabification or the cluster is simplified.

Against the European background the Balkan languages are not distinguishable by anything special, except for one specific feature which consists in the presence of the NO- clusters /N - nasal sonorant/ in word initial position in some of them. On the contrary, as far as the syllable problem is concerned, we observed here some differentiation, while similarities concern trivial features.

With regard to syllable pattern, Balkan languages can be divided in two ways: /1/ into languages with sonorous syllable pattern and languages in which there are considerable deviations from the sonorous pattern, and /2/ into languages with relatively symmetrical syllable pattern /i.e. ones in which initial as well as final consonant clusters are allowed/ and languages with nonsymmetrical syllable pattern.

Among the Balkan languages we do not find two identical situations. In Bulgarian and Macedonian only the combinations of OS- at the beginning and -SO at the end of the word are allowed. In Macedonian, apart of this, fixed order of sonorants in multisonorant cluster is required, which is motivated by differences of loudness of subsequent segments and position in the syllable. These restrictions do not apply to Bulgarian. In Albanian and Roumanian, nasal sonorants partially belong to the distributional class of obstruents. In Albanian restrictions for nasal sonorants, as for other sonorants, remain at the end of the word, in Roumanian - at the beginning of the word. Thus, the NO- clusters are allowed in initial position in Albanian, and -ON clusters in final position are allowed in Roumanian. Greek has a sonorous syllable pattern, as has Macedonian, but it differs from Macedonian by relative asymmetry. Greek is the only Balkan language with nonsymmetrical syllable pattern, where word final consonants and final consonant clusters are considerably reduced.

The difference between languages with nonsymmetrical syllable pattern and the ones with symmetrical pattern slowly decays as a result of the introduction of symmetrical structures, mainly through borrowings. However, this fact does not seem to be connected with language contacts within the territory of the Balkans but mainly with invasion of Anglicisms which introduce consonants or consonant clusters in final position of the word. Thus, this dichotomy has a relative character - it results from comparison of the generalized situation, from the impression we get while ignoring structures of the lowest frequency - various "untypical" structures.

In Greek there are several loanwords with final consonants and final consonant clusters. Such foreign words still make up quite a small part of the Greek vocabulary - in texts words with final consonant clusters occur rarely, and some native speakers assimilate them according to the native pattern. If this language periphery is left aside, then for Greek we observe the word pattern with an open or relatively open last syllable. However, some groups of borrowings with final consonant clusters of -SO type do not undergo assimilation, which is an evident proof of changes in the standing syllable pattern.

Thus, taking into account the complete lexical material, the differentiation into symmetrical syllable pattern vs nonsymmetrical pattern has no justification, and Greek belongs to the same type as the South Slavonic languages. /Mutatis mutandis the same applies to the Turkish language in which consonant clusters appear in final but not in initial position/. However, differences in frequency of various syllable structures still remain, which creates some general view of the situation - impression of existence of restrictions which are already out of date.

All that has been said here about Greek also applies to several Romance languages in which, as in Greek, initial consonant clusters of sonorous structures occur, but, with the exception of several loanwords, final consonant clusters are not allowed. Words can end with vowels or single consonants, the inventory of which is very limited. Such situation is found in Spanish, Portuguese and Italian. In Portuguese domestic words /s/, /r/ and /l/ can stand at the end of the word; in the Andalusian dialect of Spanish - only /l/, /r/ and /n/; the same applies to Italian; in common Spanish also /s/ and /θ/; in Greek - only /n/ and /s/. The differences between these languages concern mainly the combinations of obstruents which are due to genetic difference. What is significant in these languages and especially in their colloquial realizations, various interventions occur adapting the foreign structures according to the domestic pattern, cf. Port. Nova Iorque, clube, dial.Ital. lapisse /stand. lapis/, Greek grup||grupa|| grupos, etc.

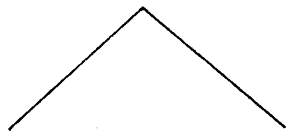
Final consonant clusters appear in Catalan and Occitan. They have simple sonorous structures and are less numerous than in French or Roumanian.

We are of the opinion that the stated similarity of syllable/word pattern is worthy of consideration as a typological feature. This feature consists in the nonsymmetrical syllable pattern with uncomplicated initial consonant clusters of sonorous structures and with open or relatively open syllable rhyme; the inventory of consonants which can stand at the end of the word in each of these languages is very limited. What is significant is that all these languages are concentrated in the basin of the Mediterranean and, with regard to syllable pattern, they stand in opposition to the Central and North European languages. The example of the Čakavian dialect of Serbo-Croat can also be instructive here. Compared with the standard Štokavian Serbo-Croat, Čakavian shows the tendency to simplify the structure of the syllable rhyme. Thus, with respect to the phonemic syllable pattern, one should speak of the Mediterranean community rather than of the Balkan Sprachbund.

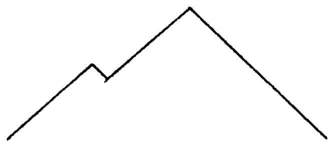
The only indubitable Balkanism partially connected with the syllable problem is the occurence of consonant clusters of the type ´nasal sonorant + occlusive`, which can occur in the same order in any position in a word on the limited territory of the Balkans. The situation is as follows: in contemporary Greek, in the colloquial variant of Demotic, there is a strong tendency for functional identification of the opposition: voiced vs voiceless occlusive with the opposition: occlusive with the nasal implosion vs occlusive without the nasal implosion, that is, cf. /p/ vs /b/ = /p/ vs /mb/, /t/ vs /d/ = /t/ vs /nd/, etc., with simultaneous reduction of the clusters with voiceless occlusives, which undergo voicing. Similar tendencies occur in Albanian, where additionally, unlike in Greek, the NO clusters occur also in word initial position. In standard Albanian the opposition /b/ vs /mb/ vs /p/ vs /mp/, etc. is phonologically relevant, but in dialects the situation is obviously differentiated. In dialects we find such phenomena as: voicing of obstruents after a nasal sonorant, prenasalization of voiced occlusives, occasionaly adding an occlusive after a nasal sonorant, etc. More detailed informations and exemplification can be found in our study: Bałkańsko-romańskie paralele w zakresie syntagmatyki fonologicznej, Językowe studia bałkanistyczne II, Wrocław 1987.

The schemes of syllable patterns can be featured using a line the level of which corresponds to the loudness of subseguent segments:

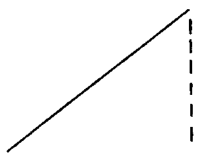
Macedonian
Bulgarian
Serbo-Croatian
/Catalan, Occitan, Čakavian/



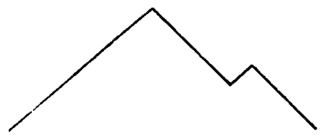
Albanian



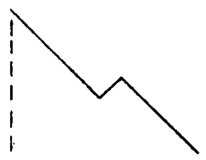
Greek
Italian
Portuguese
Spanish



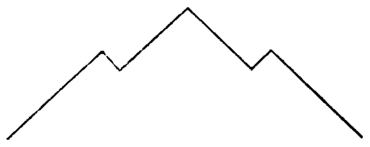
Roumanian



Turkish



Polish
Russian



PHONOSTATISTICAL CHARACTERISTICS OF THE ESTONIAN LANGUAGE

JUHAN TULDAVA

Dept. of Applied Linguistics
Tartu State University
Tartu, Estonia, USSR 202400

ABSTRACT

The paradigmatic as well as the syntagmatic (positional) relations between the phonemic units of the Estonian language are examined from the quantitative point of view. The results of the investigation are compared with analogical data from some other languages (particularly Finnish and Hungarian).

THE INVENTORY

The phoneme inventory of the Estonian language contains 9 vowels /a e i o u õ ä ö ü/ and 17 consonants /p t t́ k f h j l ĺ m n ń r s ś š v/ [1; 2]. All these phonemes may be short or long. The long monophthongs and long consonants are considered to be single phonemes. There are 36 diphthongs in Estonian [3] but phonologically they are treated as sequences of two vowels. All nine Estonian vowels contrast in stressed position but in unstressed position only four of them (/a e i u/) occur in the normal system (the literary language). The first component of an Estonian diphthong may be any of the nine vowels but the second component has to be chosen out of the first five vowels /a e i o u/, not all of these combinations being acceptable [3].
In orthography the long vowels are marked with two graphemes representing the same quality (e.g. maa /mā/ ´land, country´). The long consonants may be marked with two graphemes (e.g. linn /lin̄/ ´town´) or sometimes with one grapheme (linlane /liñlane/ ´town-dweller´).
All stops in Estonian are unvoiced, the distinction is made between short and long stops (lenis and fortis on the phonetical level). The short stops may be marked with the graphemes b, d, g or p, t, k (e.g. viga /vika/ ´mistake´ and kord /kort/ ´order´) The long stops are usually marked with two graphemes (pp, tt, kk) or in some positions with only one grapheme (pikk /pik̄/ ´long´ and piklik /pik̄lik̄/ ´oblong´). For more detailed analysis the quantity alternation of the Estonian language has to be considered (short, long, overlong).
The phonology of a language cannot be regarded as complete if it does not take into account some basic quantitative (statistical) features of the system and the functioning of its units in speech (text). For instance, the number of vowels in a phonemic system indicates the degree of "vocalism" (Vokalhaltigkeit) and may be regarded as a typological characteristic of a language [4; 5]. But even more important for the phonostatistical study of languages is the investigation of the frequency of occurrence of phonemic units in text.

TEXT FREQUENCIES

Our study is based on a corpus of texts of the contemporary Estonian language (55 % of fiction and 45 % of non-fiction) with a total of about 150,000 running phonemes. The results of the statistical investigation will be given in a simplified form: the frequencies of short and long phonemes (e.g. /a/ and /ā/) are counted summarily and so are the frequencies of the non-palatalized and palatalized forms of the consonants /t l n s/. In this case the total number of phonemes is 22.
If we group these phonemes according to their occurrence we can distinguish three main groups constituted by phonemes of relatively high frequency (p ⩾ 6 %), medium (6 < p < 2) and low frequency (p < 2):

a	12.2	n	4.6	j	1.9
t	11.9	m	4.0	h	1.7
e	11.0	o	3.1	ä	1.3
i	9.5	r	2.9	õ	1.3
s	9.0	p	2.6	ü	0.9
k	7.3	v	2.3	ö	0.2
l	6.2			f	0.05
u	6.0			š	0.05

In full accord with other linguistic levels the functioning of the phonemic system in text reveals the tendency of concentration and dispersion of its units: we can distinguish the "core" (nuclear part) of the system, the intermediate part, and the "periphery". The three most frequent phonemes /a t e/ cover 35.1 % of the Estonian text, the eight most frequent ones - 73.1 %, and the ten most frequent phonemes - 81.7 %.
The phenomenon of concentration and dispersion is well-known in lexical statistics where the statistical distribution of the units may be expressed analytically by the so-called Zipf's law in the form of a power function. Unlike the lexical level with a very large number of units the phonemic level with its limited inventory is not submitted to Zipf's law but to logarithmic or exponential law of growth (or decrease). This can be demonstrated on our experimental material (Fig. 1): there is evidently a linear relation between the logarithm of probability (relative frequency) of a phoneme and its place (rank) in the hierarchy of units. In other words, it means exponential dependence

$$p_i = ae^{-bi} \qquad (1)$$

where p_i is relative frequency, i - rank, a and b - constants, and e - the base of natural logarithms. In our example a ≈ 17 and b ≈ 0.15.

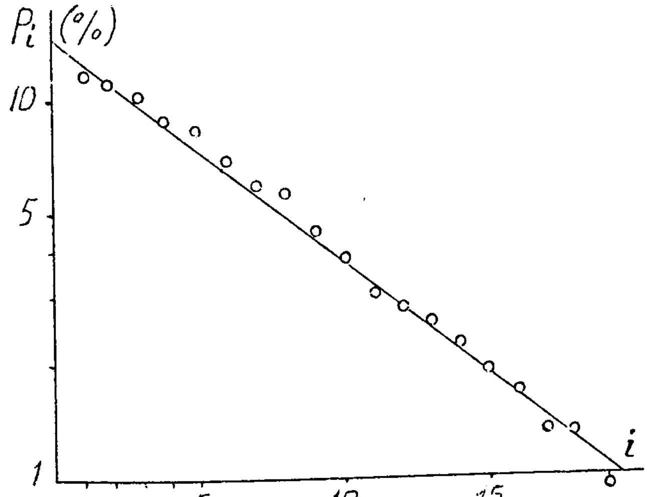


Fig. 1. Linear relation of rank (i) and the logarithm of occurrence probability of a phoneme (ln p_i).

The concrete form and the values of the constants in the formula approximating the empirical curve may serve as typological characteristics of a language. It may be added that the principle of concentration and dispersion of units in any concrete manifestation is considered to express a universal law which is peculiar to certain self-regulating systems in social life.
Another method of estimating the state of the functioning system as a whole is the measurement of the entropy of the system. The entropy of phoneme frequencies is defined as

$$H = -\sum_{i=1}^{K} p_i \log_2 p_i \qquad (2)$$

where H marks entropy, p_i is the probability (in the empirical case - the relative frequency) of the phoneme in a system of k phonemes; $\log_2$ means logarithm with base 2.
For the simplified Estonian system with 22 phonemes we get H = 3.9063. In terms of the theory of information, we can say that the entropy per phoneme occurrence is 3.9063 bits of information. Actually the entropy measures the degree of "equidistribution" of the phonemes in text. For comparison with other results we have to compute the relative entropy

$$H_{rel} = \frac{H}{H_o}$$

where $H_o = \log_2 k$. It is necessary in cases where the compared systems have different numbers of elements. For instance, we can compare our results with the results of other investigations [6] (Table 1).

Table 1
Entropy of phonemic systems

Language	k	H	H_o	H_{rel}
Estonian	22	3.9063	4.4594	0.8760
Hungarian	39	4.6028	5.2854	0.8709
German	33	4.4435	5.0444	0.8809
English	39	4.7098	5.2854	0.8911
Russian	41	4.8257	5.3576	0.9007

The smaller the value of H_{rel}, the more compressed is the series of phonemes against that of equidistribution. In this respect Estonian and Hungarian, having relatively low values of H_{rel}, differ discernibly from other languages under examination.
However, if we compare the statistical distribution of the frequencies of concrete phonemes e.g. in Estonian and Hungarian, we may find both resemblances and essential differences. In Hungarian the eight most frequent phonemes in texts of fiction are /e a t n l k m r/ [7]. Five of them coincide in Estonian and Hungarian and the three most frequent ones are the same (/a e t/). But there are differences in the distribution of medium and low frequency phonemes and in phoneme systems on the whole.
As to Finnish it must be noted that there are some pecularities in the distribution of phonemes in Finnish texts that make the difference between the two close cognates - Finnish and Estonian - remarkable enough. The most frequent phonemes in Finnish texts are /a n i e t s ä k o l/ [8]. The most striking difference lies in the frequency of occurrence of the phoneme /n/. In Finnish it occupies the second place with the relative frequency of about 10 %

(in Estonian /n/ is on the 9th place with the frequency of 4.6 %). This is to a great extent due to the high frequency of occurrence of final /-n/ in common words where Estonian has lost the final consonant in the course of historical development, e.g. Finnish niin - Estonian nii ´so´, kuin - kui ´when´, paljon - palju ´much´, or in genitive and illative forms, e.g. Finnish jalan, jalkaan - Estonian jala, jalga´foot. Here we can see real interdependence which exists among separate language levels where a quantitative change in the phonological system is parallel to or motivated by the structural needs and demands of some higher level of the same language, viz. of its morphological level (cf. [9]).

PHONEME CLASSES

At the first stage of classification the phonemes are divided into two large classes: vowels (V) and consonants (C). In phonostatistical works the ratio C:V is considered to be an important typological characteristic of languages [10]. In Estonian texts the ratio is 54.5:45.5 (%), or 1.20, i.e. the consonants exceed the vowels by 20 %. This value (1.20) can be compared with the corresponding values of the ratio in other languages:

	C		V		
Finnish	52.3	:	47.7	=	1.10
Italian	54	:	46	=	1.17
Lithuanian	56.3	:	43.7	=	1.29
Ukrainian	57.8	:	42.2	=	1.37
Russian	58	:	42	=	1.38
Hungarian	58.6	:	41.4	=	1.41
Czech	58.7	:	41.3	=	1.42
Polish	58.8	:	41.2	=	1.43
German	59.7	:	40.3	=	1.49
English	60	:	40	=	1.52

As to the vowel phonemes we can further divide them into several classes according to their phonetic properties. The frequencies of occurrence of vowel classes in Estonian texts are given in Table 2 (unr - unrounded, ro - rounded).

Table 2
The vowel system: frequencies in text

	Front		Back		Total
	unr	ro	unr	ro	(%)
High	i 20.8	ü 2.0	õ 2.9	u 13.2	38.9
Mid	e 24.2	ö 0.4	-	o 6.8	31.4
Low	ä 2.9	-	a 26.8	-	29.7
Total (%)	47.9	2.4	29.7	20.0	100.0
	50.3		49.7		

As can be seen, the front and back vowels in Estonian texts are equally distributed (about 50:50 %). In Finnish and Hungarian the front:back ratio is 52:48, the same in Italian, but for instance in Slovak it is 43:57, and in Sanskrit texts 20:80.

The relation of the frequency of short vowels to the frequency of long vowels in Estonian texts is 92 to 8 %. The same relation characterizes Finnish texts, whereas in Hungarian the long vowels occur more often and the ratio "short:long" is 80:20 (%).

The classification of consonants according to the manner of articulation and according to the place of articulation are brought together in the synoptic Table 3.

Table 3
The consonant system: frequencies in text

	Labial	Alveodent non-pal	Alveodent pal	Palatal	Velar	Total (%)
Stops	p 4.8	t 21.8	t´	-	k 13.4	40.0
Fricat	f 0.1	s 16.5	s´	š 0.1	h 3.1	19.8
Nasals	m 7.3	n 8.5	n´	-	-	15.8
Laterals	-	l 11.4	l´	-	-	11.4
Trills	-	r 5.3	-	-	-	5.3
Semi-vowels	v 4.2	-	-	j 3.5	-	7.7
Total (%)	16.4	63.5		3.6	16.5	100.0

Two parallel sets of alveodentals (except /r/) can be distinguished: non-palatalized and palatalized consonants. It has been ascertained that except in case of automatic palatalization before /i/ and /j/ the palatalized consonants /t´ s´ n´ l´/ cover only 0.15 % of all running phonemes in Estonian texts [11].

The identification of long consonant phonemes in a running text is problematic in some cases. We estimate that about 17 % of consonants are long and 83 % short.

As a whole, the quantitative distribution of phonemes in Estonian texts can be illustrated in the following manner:

	Vowels 45.5	"Resonants" (67.4)
"Consonants" (45.5)	Sonorants 21.9	
	Obstruents 32.6	
	100.0 (%)	

POSITION ANALYSIS

The phonemes occur with different frequencies in different positions of the word. In principle, initial, medial and final positions can be distinguished.

In the Orthological Dictionary of the Estonian language [12] in 115,000 entries the most frequent initial phonemes are (%): /k/ 17.4, /p/ 11.8, /t/ 10.3, /s/ 9.2, /v/ 6.3, /m/ 6.3, /l/ 6.1, /a/ 5.5, /r/ 5.4, /h/ 4.2. Among the ten most frequent phonemes there is only one vowel (/a/). On the whole, the vowels make up 15.5 and the consonants 84.5 per cent of all initial phonemes in the dictionary.

On the text level the most frequent initial phonemes are (%): /k/ 14.1, /t/ 9.9, /s/ 8.8, /m/ 8.3, /p/ 7.6, /o/ 7.0, /v/ 6.5, /e/ 6.4, /j/ 5.6, /a/ 5.3 followed by /n, l, h, r, i, ü, õ, u, ä, ö/ and the "foreign" phonemes /f/ and /š/. The five most frequent initial phonemes are all consonants and they cover about 50 % of all word initial phonemes in the text.

The over-all distribution of phoneme classes in initial positions is presented in Table 4.

Table 4
Distribution of initial phonemes

Phoneme class	Dictionary		Text	
Obstruents: stops	39.5	55.1	31.6	43.4
fricat.	15.6		11.8	
Sonorants: nasals	9.4	29.4	12.9	31.4
laterals	6.1		3.8	
trills	5.4		2.6	
semivowels	8.5		12.1	
Vowels	15.5		25.2	
Total (%)	100.0		100.0	

In Finnish the vowels cover 20 % and the consonants 80 % of all word initial positions in the text. The most frequent initial phonemes are /j s k h t m o v p e/. Compared with Estonian the phonemes /j/ and /h/ are of exceptionally frequent occurrence in initial positions.

As the structure of the stressed syllable is somewhat different from that of the unstressed syllables, it is expedient to examine the frequency distribution of vowels in the nuclei of stressed syllables separately (including the nuclei of monosyllabic words): single vowels 88.0 (76.5 % short and 11.5 % long) and diphthongs (i.e. 2-vowel sequences) 12.0 %. The frequencies of single vowels: /a/ 20.3, /e/ 19.0, /o/ 12.1, /i/ 11.2, /u/ 9.0, /ä/ 5.9, /õ/ 5.5, /ü/ 4.0, /ö/ 1.0. The most frequent diphthongs: /ei/ 2.7, /ea/ 1.7, /õi/ 1.6, /ui/ 1.3, /äi/ 0.9, /ai/ 0.7.

The distribution of word final phonemes reflects the morphological structure of the language and therefore the frequencies of final phonemes are considered to be specific for each language. In Estonian texts the most frequent final phonemes are: /a/ 21.1 % (of all final phonemes in the text), /t/ 20.5, /e/ 13.6, /s/ 13.4, /i/ 12.9. These five phonemes cover 81.5 % of all word endings in the text. They are followed by the less frequent phonemes: /l/ 4.7, /u/ 4.4, /n/ 3.1, /p/ 2.3, /k/ 1.7, /r/ 0.6, /v/ 0.4. Due to the restrictions in the distribution of vowels in unstressed syllables the phonemes /o õ ä ö ü/ are extremely rare in word endings (total 0.3 %) and so are the phonemes /h/ and /f š/ (the last two occur only in foreign or recent loan words); the three phonemes have a total frequency of 0.1 %. The distribution of phoneme classes in final position: obstruents 38.0 %, sonorants 9.7 %, and vowels 52.3 %.

In Finnish the most frequent phonemes in word final position are: /n a ä i t e s o u y/. The final /n/ covers almost 30 % of all word endings in the text.

On the basis of the frequencies of phonemes in initial and final positions their relative frequencies in medial positions can be calculated.

Some other traditional problems in phonostatistics, such as the valency fields of phonemes, phonotactic features and frequencies of phoneme sequences and syllables, word length, etc., as well as a more detailed quantitative analysis of phonological data - including stress and quantity - require special discussion.

REFERENCES

[1] T.-R. Viitso. Läänemeresoome fonoloogia küsimusi. Tallinn: KKI, 1981.

[2] A. Eek. Kvantiteet ja rõhk eesti keeles (II). Seisukohavõtt. - Keel ja Kirjandus 1987, nr. 3, 153-160.

[3] H. Piir. Acoustics of the Estonian diphthongs. - Estonian Papers in Phonetics 1982-1983. Tallinn: KKI, 1985, 5-96.

[4] G. Altmann, W. Lehfeldt. Allgemeine Sprachtypologie. Prinzipien und Meßverfahren. München: W. Fink Verlag, 1973.

[5] U. Strauß. Struktur und Leistung der Vokalsysteme./Quantitative Linguistics, vol. 4. Bochum: Brockmeyer, 1980.

[6] P. Zörnig, G. Altmann. The entropy of phoneme frequenciesand the Zipf-Mandelbrot law. - In: Glottometrika 6./Quantitative Linguistics, vol. 25. Bochum: Brockmeyer, 1984, 41-47.

[7] F. Papp. Lingvostatistika i vengerskij jazyk. - Acta et Commentationes Universitatis Tartuensis, vol. 518, Tartu, 1980, 15-37.

[8] V. Setälä. Suomen kielen dynamiikka I. Helsinki: SKS, 1972.

[9] J. Vachek. Prague phonological studies today. - Travaux linguistiques de Prague I. Prague: Academia, 1966, 7-20.

[10] Yu. Tambovcev. Konsonantnyj koefficient v jazykax raznyx semej. Leningrad, 1986.

[11] M. Hint. Häälikutest sônadeni. Tallinn: Valgus, 1978.

[12] Õigekeelsussõnaraamat./Toimet. R. Kull, E. Raiet. Tallinn: Valgus, 1976.

PHONOTAKTISCHE GESETZMÄSSIGKEITEN IM KONSONANTISMUS DES TREMJUGAN-OSTJAKISCHEN - EIN BEITRAG ZU PHONETISCHEN UNIVERSALIEN

ERHARD F. SCHIEFER
Congregatio Ob-Ugrica
München. FRG

LIESELOTTE SCHIEFER
Institut für Phonetik und
Sprachliche Kommunikation der
Ludwig-Maximilians Universität
München. FRG

RESÜMEE

Mit vorliegender Arbeit werden zwei Ziele verfolgt. Zum einen wird eine aus der 'combination analysis' weiterentwickelte phonotaktische Methode vorgestellt. die erlaubt. die Kombinationsfähigkeit jedes Konsonanten einer Sprache für jede Position innerhalb einer Konsonantenverbindung zu bestimmen. Die Kalkulation basiert auf der phonetischen Klasse des Konsonanten (z.B. Plosiv. Frikativ. labial. alveolar). der Klassengröße. der theoretischen und der tatsächlichen Kombinierbarkeit. Zum anderen wird diese Methode an Material aus dem Tremjugan-Dialekt des Ostjakischen demonstriert. einem Dialekt. der in medialer Wortposition nur zweigliedrige Konsonanten-Verbindungen duldet.

EINLEITUNG

Obwohl an der Notwendigkeit phonotaktischer Analysen zumindest seit Trubetzkoy [4] keine Zweifel bestehen. sind Arbeiten auf diesem Gebiet nach wie vor eher selten. Die verdienstvolle Arbeit von Ian Maddieson und Mitarbeitern (UPSID. [1]) beinhaltet zwar Phonemsysteme und die klassifizierende Auswertung derselben aus 317 Sprachen; die entsprechenden phonotaktischen Untersuchungen fehlen jedoch und werden sicher auch noch lange auf sich warten lassen: dabei sind Arbeiten auf phonotaktischem Gebiet durch die Möglichkeit des Einsatzes von Computern längst nicht mehr so zeitaufwendig wie zuvor. Eine der umfangreichsten phonotaktischen Arbeiten wurde von B. Sigurd [3] für das Schwedische vorgelegt, in der gleichzeitig die damals bekanntesten und erfolgversprechendsten phonotaktischen Methoden referiert wurden. Diese seien hier kurz charakterisiert. (1) In der 'position analysis' wird die Position einzelner Phoneme innerhalb bestimmter Grenzen (etwa der Silbe) thematisiert. (2) Die Ordnung der Phoneme in Gruppen steht im Vordergrund des Interesses bei der 'order analysis'. die als Hauptergebnis Klassen liefert. welche hierarchisch geordnet werden und deren Mitglieder nur mit Gliedern anderer Klassen. aber nicht mit Gliedern der eigenen Klasse zu Gruppen kombiniert werden können. (3) Die 'combination analysis'. deren Hauptinteresse der Kombinationsfähigkeit der Phoneme untereinander und damit der bestehenden Restriktionen. weniger der Ordnung innerhalb der Konsonantengruppen gilt. Die Anwendung dieser Methoden hängt zum einen wesentlich von der zu untersuchenden Sprache und deren Gesetzmäßigkeiten. zum anderen von der zugrunde gelegten Rahmeneinheit (Silbe, Morphem, Lexem; initiale. mediale, finale Konsonantengruppen etc.) ab. Die Anwendung aller bisher vorgeschlagenen Methoden erfordert zuviel Aufwand und ist daher meist nur schwer zu praktizieren. Andererseits sollten phonotaktische Ergebnisse mit denen anderer Sprachen vergleichbar sein. Und es sollten dabei sowohl phonostatistische wie phonotaktische Gesichtspunkte berücksichtigt werden.

Die hier angewendete Methode wurde an Vachostjakischem Material erprobt (Schiefer 1975, [2]) und stellt eine Weiterentwicklung der 'combination analysis' mit Übernahme von Gesichtspunkten aus der 'position analysis' dar. Das Hauptinteresse gilt der Kombinationsfähigkeit ('combinality') von Phonemen zu Gruppen und den dabei zu beobachtenden Restriktionen, die im Ostjakischen von großem Interesse sind. Die Methode berücksichtigt neben der Kombinationsfähigkeit die Positionsabhängigkeit, die phonetische Klasse des Phonems, die Klassengröße, die daraus berechnete theoretische und die tatsächlich gegebene Kombinationsfähigkeit. Die auf diese Weise erhaltenen, als Zahlenfolgen darstellbaren Ergebnisse werden mit denen aus anderen Sprachen direkt vergleichbar und können in Form von hierarchisch geordneten Kombinationsregeln formuliert werden. Im folgenden wird die Anwendung der Methode schrittweise an Material aus dem TrjO dargestellt.

PHONOTAKTISCHE ANALYSE DES TrjO

Der Konsonantismus des (TrjO) ist durch zwei generelle Restriktionen gekennzeichnet. (1) In initialer Position sind keine Konsonantenverbindungen (KV) zulässig. (2) Verbindungen von mehr als zwei Konsonanten werden nicht geduldet. Unsere Analyse beruht daher auf den zweigliedrigen medialen KV des TrjO. Als Rahmeneinheit wurde das Lexem gewählt, das die geringsten Restriktionen aufweist. Das TrjO besitzt 18 Konsonantenphoneme, die in Tab. 1 nach Artikulationsmodus (AM) und Artikulationsstelle (AS) geordnet aufgeführt sind.

Eine tabellarische Erfassung der KV basiert zweckmäßigerweise (in Abhängigkeit von den Gegebenheiten der zu analysierenden Sprache) auf den phonetischen Klassen des AM und der AS der Konsonanten. Sie wird hier (s. Tab. 2) für die AS dargestellt, da deren Einfluß auf die Kombinationsfähigkeit im TrjO größer ist als der anderer phonetischer Parameter. Sie liefert die Basis für alle weiteren Analysen.

Tab. 1: Konsonantensystem des TrjO

	labial	alveolar	retroflex	mouill.	palatal	velar
Plosive	p	t	č	ť		k
Nasale	m	n	n	ń		ŋ
Laterale		ʌ	l	ʎ		
Trill		r				
Frikative		s				ɣ
Glides	w				j	

Tab. 2: Mediale Konsonantenverbindungen im TrjO

	p	m	w	t	n	ʌ	r	s	ť	ń	ʎ	č	n	l	j	k	ŋ	ɣ
p				pt		pʌ		ps						pl			pŋ	pɣ
m	mp	mm		mt	mn	mʌ	mr	ms	mť			mč		ml				mɣ
w				wt	wn		wr											
t	tp	tm		tt	tn	tʌ		ts							tj	tk		tɣ
n	np	nm		nt	nn	nʌ		ns									nŋ	nɣ
ʌ	ʌp	ʌm		ʌt	ʌn	ʌʌ		ʌs								ʌk	ʌŋ	ʌɣ
r	rp	rm		rt	rn	rʌ		rs							rj	rk		rɣ
s	sp	sm		st	sn	sʌ		ss	sť							sk	sŋ	sɣ
ť								ťs		ťń						ťk	ťŋ	ťɣ
ń	ńp			ńt					ńť	ńń	ńʎ					ńk		ńɣ
ʎ				ʎt											ʎj			
č	čp									čń				čl				čɣ
n								ns				nč		nl		nk	nŋ	nɣ
l												lč		ll		lk		
j	jp	jm		jt	jn	jʌ						jč						jɣ
k				kt	kn	kʌ		ks						kl				
ŋ		ŋm		ŋt	ŋn	ŋʌ								ŋl		ŋk		
ɣ	ɣp	ɣm		ɣt	ɣn	ɣʌ	ɣr			ɣń			ɣn	ɣl				

Die Kombinationsfähigkeit (KF) der einzelnen Konsonanten ergibt sich aus der Häufigkeit ihres Auftretens (a) in den KV generell und (b) in 1. bzw. 2. Position einer KV. Daraus läßt sich die prozentuale Häufigkeit des Auftretens für jeden Konsonanten berechnen (s. Tab. 3). Diese Häufigkeitstabelle ist vor allem von phonostatistischem Interesse. Weiteren Aufschluß erhält man, wenn man die Differenzen zwischen der 1. und 2. Position berechnet, worauf hier aus Platzgründen verzichtet werden muß. Es sei jedoch festgehalten, daß die Unterschiede zwischen der 1. und 2. Position nicht für alle Konsonanten gleich sind: so treten die Plosive offensichtlich häufiger in der 2. Position auf (t 9:14, p 6:10. k 5:9) während der Trill (r 9:4) und die Glides (j 7:3, w 3:0) häufiger in der 1. Position zu finden sind. Mithin muß bei der Formulierung von Restriktionsregeln nicht von den Einzel-Konsonanten, sondern von phonetischen Klassen ausgegangen werden. Die Analyse beruht dann zunächst ebenfalls auf Häufigkeitsdaten, wie in Spalte 3 der Tab. 4 und 5 für die AM und AS dargestellt. Es ist offenkundig, daß bei den AM die Positionsabhängigkeit für die Nasale (39:40) und Frikative (19:21) gering ist, für die anderen Klassen jedoch groß (Plosive 28:40; Laterale 14:20; Glides 10:3; Trill 9:4), wobei Plosive und Laterale häufiger in der 2. Position, der Trill und Glides häufiger in der 1. Position auftreten. Ähnliche Positionsabhängigkeiten sind in Bezug auf die AS feststellbar: die Alveolare (45:49), Velare (21:27), Labiale (20:19) und Retroflexe (13:13) sind relativ unabhängig; die Mouillierten (13:8) und der Palatal (7:3) dagegen sind in der 1. Position häufiger als in der zweiten. Die so gewonnenen Ergebnisse vermitteln jedoch ein

Tab. 3: Anteil der einzelnen Konsonanten an den KV in 1. Position, 2. Position und Gesamtanteil

1. Position			2. Position			gesamt		
m	11	9.2	t	14	11.8	t	23	9.7
s	10	8.8	ɣ	12	10.1	ɣ	21	8.8
t	9	8.4	n	11	9.2	ʌ	20	8.4
ɣ	9	7.6	ʌ	11	9.2	m	20	8.4
ʌ	9	7.6	p	10	8.4	n	19	8.0
r	9	7.6	k	9	7.6	s	19	8.0
n	8	6.7	s	9	7.6	p	16	6.7
ń	7	5.9	m	9	7.6	k	14	5.9
ŋ	7	5.9	l	8	6.7	r	13	5.5
j	7	5.9	ŋ	6	5.0	ŋ	13	5.5
p	6	5.0	č	4	3.4	l	11	4.6
n	6	5.0	ń	4	3.4	ń	11	4.6
k	5	4.2	r	4	3.4	j	10	4.2
ť	4	3.4	ť	3	2.5	č	8	3.4
č	4	3.4	j	3	2.5	ť	7	2.9
w	3	2.5	n	1	0.8	n	7	2.9
l	3	2.5	ʌ́	1	0.8	ʌ́	3	1.3
ʌ́	2	1.7	w	-	-	w	3	1.3

Tab. 4: Kombinationsfähigkeit (KF) der Artikulationsmodi: Klassengröße, theoretische KV, existierende KV, KF in Prozent

1. Position				2. Position				gesamt			
FRIK (2)	36	19	52.8	FRIK (2)	36	21	58.3	FRIK (2)	72	40	55.6
TRIL (1)	18	9	50.0	PLOS (5)	90	40	44.4	NAS (5)	180	70	38.9
NAS (5)	90	39	43.3	LAT (3)	54	20	37.0	PLOS (5)	180	68	37.8
PLOS (5)	90	28	31.1	NAS (5)	90	31	34.4	TRIL (1)	36	13	36.1
GLID (2)	36	10	27.8	TRIL (1)	18	4	22.2	LAT (3)	108	34	31.5
LAT (3)	54	14	25.9	GLID (2)	36	3	8.3	GLID (2)	72	13	18.1

Tab. 5: Kombinationsfähigkeit (KF) der Artikulationsstellen: Klassengröße, theoretische KV, existierende KV, KF in Prozent

1. Position				2. Position				gesamt			
ALV (5)	90	45	50.0	ALV (5)	90	49	54.4	ALV (5)	180	94	52.2
VEL (3)	54	21	38.9	VEL (3)	54	27	50.0	VEL (3)	108	48	44.4
PAL (1)	18	7	38.9	LAB (3)	54	19	35.2	LAB (3)	108	39	36.1
LAB (3)	54	20	37.0	RET (3)	54	13	24.1	PAL (1)	36	10	27.8
MOU (3)	54	13	24.1	PAL (1)	18	3	16.7	RET (3)	108	26	24.1
RET (3)	54	13	24.1	MOU (3)	54	8	14.8	MOU (3)	108	21	19.4

verzerrtes Bild von der KF der einzelnen Klassen, da die Klassengrößen nicht berücksichtigt wurden: die KF müßte bei dieser Berechnung zwangsläufig mit der Klassengröße zunehmen. Berücksichtigt man diesen Gesichtpunkt, und setzt man die theoretische KF in Beziehung zu der tatsächlich gegebenen, so erhält man die KF für die einzelnen Klassen, wie in Spalte 4 der Tab. 4 und 5 zu sehen ist. Die Ausnutzung der theoretischen KF ist bei den Frikativen in beiden Positionen am größten (52.8:58.3). Die Plosive (52.8:58.3) und Laterale (25.9:37.0) zeigen erneut höhere KF in der 2. Position, während die Nasale (43.3:34.4), die Glides (27.8:8.3) und der Trill (50.0:22.2) in der 1. Position grössere KF aufweisen. Am geringsten ist jedoch die KF der Glides. Bei den AS gilt für die Positionsabhängigkeit: gering bei den Alveolaren (50.0:54.4) und den Labialen (37.0:35.2), groß bei den Velaren (38.9:50.0), den Mouillierten (24.1:14.8) und dem Palatal (38.9:16.7). Keine Positionsabhängigkeit liegt bei den Retroflexen (24.1:24.1) vor.

Die KF der einzelnen Konsonanten bzw. der einzelnen Klassen wurde bisher nur im Hinblick auf ihr Auftreten generell sowie auf die Positionsabhängigkeit betrachtet, unberücksichtigt blieb dagegen der 2. Konsonant der KV. Diese Abhängigkeit wird in den Tab. 6 (AM) und Tab. 7 (AS) aufgezeigt, wobei die Zahlen die Ausnutzung der theoretischen KF wiedergeben. In Bezug auf den AM sind drei generelle Restriktionen feststellbar: (1) Restriktionen bestehen nur in Bezug auf die 2. Position, (2) Restriktionen bestehen im Hinblick auf die Glides und den Trill und (3) Verbindungen innerhalb der einzelnen Klassen sind selten (ausgenommen Frikative). Bezüglich der AS lassen sich folgende generelle Restriktionen formulieren: (1) Restriktionen bestehen bezüglich der Klassen Mouilliert, Retroflex und Palatal und (2) Verbindungen innerhalb der labialen und velaren Klassen sind selten.

Die übrigen phonotaktischen Regularitäten lassen sich in Form des folgenden hierarchisch geordneten Regelsystems formulieren. (Ausnahmen von den Regeln stehen in Klammern). Die Regeln sind folgendermaßen zu lesen:

(a) C1 = LAB → C2= ALV bedeutet: wenn der 1. Konsonant ein Labial ist, muß der 2. Konsonant ein Alveolar sein.

(b) C2 ≠ w Der 2. Konsonant darf nicht /w/ sein.

(c) C1 = LAB, C2 = RET
C2 ≠ NAS

Wenn der 1. Konsonant ein Labial und der 2. Konsonant ein Retroflex ist, so darf dieser kein Nasal sein.

(d) C1 = ALV, C2 = α ALV

Wenn der 1. Konsonant ein Alveolar ist, so kann der 2. Konsonant ein beliebiger Alveolar sein.

(e) C1,2 = RET, C1,2 ≠ ALV

Tab. 6: Kombinationsfähigkeit der Klassen untereinander (Artikulationsmodi)

	PLOS	NAS	LAT	TRIL	FRIK	GLID
PLOS	28.0	28.0	40.0	--	70.0	10.0
NAS	56.0	36.0	46.7	40.0	70.0	--
LAT	40.0	20.0	22.2	--	33.3	16.7
TRIL	60.0	40.0	33.3	--	100.0	50.0
FRIK	60.0	70.0	50.0	50.0	50.0	--
GLID	40.0	30.0	16.7	50.0	25.0	--

Tab. 7: Kombinationsfähigkeit der Klassen untereinander (Artikulationsstellen)

	LAB	ALV	RET	PAL	MOU	VEL
LAB	22.2	73.3	33.3	--	11.1	33.3
ALV	66.7	80.0	----	40.0	6.7	80.0
RET	11.1	6.7	55.6	---	11.1	55.6
PAL	66.7	60.0	33.3	---	---	33.3
MOU	11.1	20.0	----	33.3	44.4	55.6
VEL	33.3	73.3	44.4	----	11.1	11.1

Wenn der 1. oder 2. Konsonant ein Retroflex ist, so darf der andere Konsonant kein Alveolar sein.

(f) C1 = MOU, C2 ≠ LAB/RET

Wenn der 1. Konsonant mouilliert ist, darf der 2. Konsonant weder labial noch retroflex sein.

REGELSYSTEM

Die Abkürzungen bedeuten: PLOS = Plosiv, FRIK=Frikativ, NAS=Nasal, LAT=Lateral, LAB =labial, ALV= alveolar, RET=retroflex, PAL=palatal, MOU=mouilliert, VEL=velar, C1=1. Konsonant, C2=2. Konsonant.

(1) C2 ≠ w
(2) C2 = TRILL → C1 = labNAS/labGLIDE (ɣr)
(3) C1 = PAL → C2 = LAB/ALV (jč, jɣ)
(4) C2 = PAL → C1 = ALV (ʌ́j)
(5) C2 = MOU → C1 = MOU (mť, sť, čń, ɣń)
(6) C1 = MOU, C2 ≠ LAB/RET (ńp)
(7) C1,2 = RET, C1,2 ≠ ALV (ns)
(8) C1 = RET, C2 ≠ LAB (čp)

C1 = LABIAL
(9) C1 = labGLIDE → C2 = alvPLOS/alvNAS
(10) C1 = LAB, C2 = LAB
C1 = NAS
(11) C1 = LAB, C2 = ALV
C1 = NAS, C2 = α ALV
(12) C1 = LAB, C2 = ALV
C1 = PLOS, C2 ≠ NAS
(13) C1 = LAB, C2 = RET
C2 ≠ NAS
(14) C1 = LAB, C2 = RET
C1 = NAS, C2 = α RET
(15) C1 = LAB, C2 = RET
C1 = PLOS, C2 ≠ PLOS
(16) C1 = LAB, C2 = VEL
C2 ≠ PLOS
(17) C1 = LAB, C2 = VEL
C1 = NAS, C2 ≠ NAS

C1 = ALVEOLAR
(18) C1 = α ALV, C2 = α LAB
(19) C1 = α ALV, C2 = α ALV
(20) C1 = ALV, C2 = PAL
C1 = PLOS/TRILL
(21) C1 = ALV, C2 = VEL
C1 = LAT/FRIK, C2 = α VEL
(22) C1 = ALV, C2 = VEL
C1 = PLOS/TRILL, C2 ≠ NAS
(23) C1 = ALV, C2 = VEL
C1 = NAS, C2 ≠ PLOS

C1 = MOUILLIERT
(24) C1 = MOU, C2 = ALV
C1 = PLOS, C2 = FRIK
(25) C1 = MOU, C2 = ALV
C2 = PLOS, C1 = NAS/LAT
(26) C1 = MOU, C2 = VEL
C1 ≠ LAT
(27) C1 = MOU, C2 = VEL
C1 = NAS, C2 ≠ NAS

C1 = RETROFLEX
(28) C1 = RET, C2 = RET
C2 ≠ NAS
(29) C1 = RET, C2 = RET
C1 = PLOS, C2 = LAT
(30) C1 = RET, C2 = VEL
C1 = PLOS, C2 = FRI
(31) C1 = RET, C2 = Vel
C1 = LAT, C2 = PLOS

C1 = PALATAL
(32) C1 = PAL, C2 = α LAB
(33) C1 = PAL, C2 = ALV
C2 ≠ FRIK

C1 = VELAR
(34) C1 = VEL, C2 = LAB
C1 ≠ PLOS
(35) C1 = VEL, C2 = LAB
C1 = NAS, C2 ≠ PLOS
(36) C1 = VEL, C2 = ALV
C2 ≠ FRIK (ks)
(37) C1 = VEL, C2 = RET
C2 = LAT (ɣn)
(38) C1 = VEL, C2 = VEL
C1 = NAS, C2 = PLOS

LITERATUR

[1] Maddieson, I.: UPSID: the UCLA Phonological Segment Inventory Database. UCLA Working Papers in Phonetics, Vol. 50: 4-120 (1980)

[2] Schiefer, L.: Phonematik und Phonotaktik des Vach-Ostjakischen. Veröffentlichungen des Finnisch-Ugrischen Seminars an der Universität München. Serie B: Beiträge zur Erforschung der obugrischen Sprachen, Bd. 1 (München 1975)

[3] Sigurd, B.: Phonotactic structures in Swedish. (Lund 1965)

[4] Trubetzkoy, N. S.: Grundzüge der Phonologie. Göttingen 1958, 4. Aufl.

THE TYPOLOGY OF VOCALIC STRUCTURES OF THE WORD IN CHUKCHI-KAMCHATKAN LANGUAGES

ALEXANDER S. ASINOVSKY, ALEXANDER P. VOLODIN

Institute for Linguistics,
Leningrad, USSR, I99053

ABSTRACT

The paper deals with phonetic mechanisms of Chukchi, Koryak and Itelmen vocalic word structure. It presents a new interpretation of Chukchi-Koryak vowel harmony. The paper also describes an original type of morpheme interaction in Itelmen.

The languages of Chukchi-Kamchatkan group (Chukchi, Koryak, Itelmen) possess a common vocalic system of five elements and manifest the rise gradation that is traditionally termed "the vowel harmony". Following W.G.Bogoraz, the vowels are usually classified into 3 groups: strong vowels /a/, /e/, /o/, weak vowels /i/, /e/, /u/, and the neutral vowel /ə/. The strong vowels can co-occur within the word with strong ones: if there is in the word a morph (a prefix, a suffix, or a stem) that contains a strong vowel, all the weak vowels alternate with the strong ones. The neutral /ə/ is indifferent to synharmonic alternations.

The phonetic mechanism of the vowel rise alternation in the Chukchi-Kamchatkan languages was specified by the authors of the present paper as a result of field work. Some acoustic analysis data was also made use of. It allowed us to interpret the processes that take place in derivation and inflexion of the agglutinating language in the following way.

The three vocalic sub-systems have a common phonetic base, namely, the range of the phonetic variativity of vowels. Strong vowels have minimal range of variativity. The degree of variativity of weak vowels is big enough for their synharmonic variants to approach or even coincide with the allophones of strong vowels. The neutral vowel /ə/ has maximum range of variativity; it is completely dissolved in the phonetic structure of the word, is dependent on its vocalic structure and on surrounding vowels. The neutral vowel can realise as allophones that are identical with allophones of any vowel of the systems.

The manifestation of the "vowel harmony" can be of two kinds: the synharmonic variants can be variants of one phoneme, or can belong to two different phonemes. For the neutral vowel the synharmonic variants are always its allophones. For the weak vowels in Chukchi their open allophones /i/∾/I/, /e/∾/ɛ/, /u/∾/ʊ/ are synharmonic variants. In Koryak and Itelmen the synharmonic variants represent corresponding strong vowels and are the alternants proper: /i/∾/e/; /e/∾/a/; /u/∾/o/.

The conditions for alternations can be of three types: phonetic context (where the morphemic structure of the word has no influence); morphonological context (where the phohetic structure of the morphemes that constitute the word is important); and morphological context (where the phonetic structure of the word and of the constituting morphemes loses its value). It is the second type, the morphonological context, that determines the synharmonic alternations: the rules of the vocalic word structure are deduceable from the phonetic structure of morphemes that trigger phonetic alternations but do not include strong vowels, cf. Chukchi muri 'we' - morə=kə 'us'. In Itelmen the synharmonic alternations ignore the phonetic structure of the morphological constituents. Alongside with cases like ŋič=enk 'in possession of the wife', wač=ank 'on the stone' (marker of localic case is represented by synharmonic variants =enk/=ank that depend on the vocalism of the stem) and cases like ŋeč=anke 'to the wife', wač=anke 'to the stone' (the vocalism of the stem depends on the vocalic type of a "strong" suffix of terminalis), there are cases like ŋič=kit 'because of the wife', wač=kit 'because of the stone' where the vocalism of causal case suffix seems to be independent of the vocalism of the stem, and cases like iw=lah 'long' ič'=al 'birch grove', where "strong" suffixes =lah 'adjective marker' and =al 'generic number' do not trigger the vocalic alternation in the stem. Finally, there are stems like i'naq 'ermine' and iyaq 'dreadful' where strong and weak vowels co-occur in onr word. Morphemes of this kind evidently contradicts the notion of existence of vowel synharmonism in Itelmen.

Table I

modifiers modifiables			strong		neutral		
			R	m	R	m	
			an'čp teach	anke terminalis	a'asx nest	al generic number	lah adjective
weak	R	ič' birch		eč'=anke to the birch		ič'=al birch grove	
	m	enk localis			a'asx=ank in the nest		
		'in III infinitive	k'an'čp='an he taught him				
neutral	R	iwl long					iw=lah long
		ŋič wife		ŋeč=anke to the wife			
	m	miŋ I p.sg. Ob.	an'čp=miŋ he taught me				
		kit Causal case			a'asx=kit because of the ňest		

An interpretation of phonetic inconsistency of Itelmen synharmonism is given in Table I. The columns of the table contain modifier: stems and affixes that contain strong vowels and can synharmonically modify other morphemes in the word. Modifiers can be divided into strong and neutral according to whether they trigger obligatory synharmonic change of other morphemes. Horisontal lines in the table contain modifiables: stems and affixes that can be modified to change weak vocalism into strong one. Modifiables are also divided into strong and neutral according to whether their synharmonic change is obligatory or not. Points to intersection show obligatory, optional, and non-obligatory synharmonic modification of morphemes.

The following information about of morphemes that constitute an Itelmen word is necessary to determine whether the synharmonic alignment will take place:

I) phonetic structure

2) class of the morpheme: stem or affix

3) morphonological class

The possibility of co-occurence within one word and even one morpheme of strong and weak vowels shows that for Itelmen it is more appropriate to speak morpheme harmony rather than of vowel harmony: the analysed alternation definitely take place on the morphological level.

An unusual phonetic mechanism was found out when we analysed the words that formerly were transcribed with o/a, i/u. In words ^{o}sis 'grass' [s^{o}Ys^o], ^{o}čeḽxčeḽx 'cowberry' [č^{o}œʎ̭^{o}x^{o}č^{o}œʎ̭^{o}x^o], ^{o}kic 'ox' [k^{o}Yc^o] etc. all vowels and consonants are labialized. These words have quasi-homonyms: sis 'needle' [sIs], čeḽxčeḽx 'fur'[čɛʎ̭xčɛʎ̭x], kic 'ladder' [kIc]. An interesting feature of the labialized words is that the marker of labialization can be placed "outside the bracket": all the sounds in the words are labialized, and labialization is their only distinction from the known Itelmen functional units.

Alongside with the cases when the word equals the stem, that were illustrated above, there are cases when labialized stems can modify the affixed part of the word: ^{o}sis=al 'thick grass' [s^{o}Ys^o=ɔl^o] - [^{o}sIs=al], ^{o}sis=kit 'because of the grass' [s^{o}Ys^o=k^{o}Yt^o] - [^{o}sIs=kIt], ^{o}ses=anke 'into the grass' [s^{o}œs^o=ɔn^{o}k^{o}œ] - [^{o}ses=ankɛ].

Affixes that are attached to a labialized stem become labialized too. No affixes were found that would show indifference to the influence of a labialized stem. On the other hand, there are two affixes that can labialize a non-labialized stem: =^{o}pk'ul - a derivational marker of singleness, and =^{o}lwin - suffix meaning 'himself etc.', cf. k'aač 'back' - ^{o}k'aa=pk'ul 'vertebra' [^{o}k'ɔɔ=pk'ul], kəmma 'I' - ^{o}kmi=lwin 'I myself' [^{o}kmI=lßIn].

In isolated pronunciation, especially with high vowels, the lips of speaker visually move forward and stay round through the whole word, or, more precisely,. they get round slightly before the beginning of the utterance and stay round a little bit after it has finished. This fact was noticed before but got no linguistic interpretation.

Tentative estimation gives about 20% of labialized stems of the total amount of Itelmen stems. No phonetic or lexical distribution was found.

The possibility of the Itelmen stems to labialize the affixal part of the word is, no doubt, unique for the Chukchi-Kamchatkan languages and distinguishes Itelmen sharply from the group.Alongside with other features, this fact prompts one to look for the genetic roots of Itelmen outside the Chukchi-Kamchatkan areal.

IMPROVING VOICE QUALITY OF HEARING-IMPAIRED BY USE OF ELECTRO-GLOTTOGRAPHIC DISPLAY

YUMIKO FUKUDA

Research Institute
National Rehabilitation Center for the Disabled
Tokorozawa-shi, Saitama-ken, 359 Japan

ABSTRACT

Using an electro-glottographic device, the electro-glottogram was displayed on a cathode ray tube along with the speech sound waveform, and the possibility of utilizing the display as visual cue for laryngeal adjustment of quality of voice in the speech training of hearing-impaired was investigated. As the results of a series of trials, it was ascertained that this visual cue was useful as a feedback for modifying the mode of vibration of vocal folds. By combining this method with various others for visual display of speech, an integrated program of speech training for hearing-impaired was proposed.

INSTRUMENTAL AID FOR TRAINING VOICE QUALITY

Nowadays, various instrumental aids for visual display of speech are widely used in speech training of hearing-impaired, but they are mostly designed for the training of articulatory gestures or control of pitch and loudness of voice [1]. As for improving voice quality, there has been no training aid effectively utilized for this purpose, although it is considered to be the most basic requirement for speech intelligibility of hearing-impaired to achieve natural quality of voice.

Since it has been reported by E. Abberton and others that electro-glottography served as a visual feedback for laryngeal control in voicing [2], the possibility of applying the method to improving voice quality of hearing-impaired should be investigated.

ELECTRO-GLOTTOGRAPHY

The device used in this study was "Portable Laryngograph" which was designed according to A.J. Fourcin's principle of electro-glottography [3] and manufactured by Laryngograph Ltd. in England.

In this device, a pair of electrodes (30 milli-meters in diameter, 9 milli-meters in thickness, and weight of about 7 grams) are attached to the outer skin surface of both lateral sides of the larynx, holding by an elastic band around the neck. By applying high frequency electric current (frequency: 3 MHz, and voltage: 10 volts), the change in the current (less than 10 milli-ampare) due to the change in electrical impedance across the larynx synchronized with vibration of the vocal folds, or opening and closing of the glottis, was detected (Figure 1). The electronic circuitries including the carrier signal generator and amplitude demodulator are battery operated, so that they are insulated from the displaying and recording units. The device is small and light weighted, and easy to be handled.

The waveform of the output signal (6 volts peak to peak), the electro-glottogram, was displayed on the cathode ray tube of a synchroscope, and recorded on a data recorder in order to minimize low-frequency phase distortion.

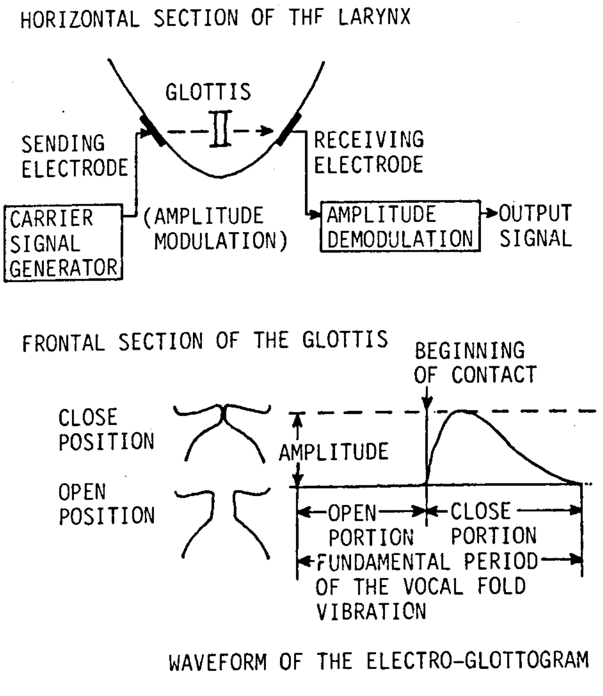


Figure 1. Description of basic components involved in the device of electro-glottography and the waveform of electro-glottogram.

For detailed inspection, the waveform of the electro-glottogram was printed out on a visicorder along with that of speech sound recorded simultaneously, then their frequency spectrum were analyzed using a sound spectrograph.

NATURE OF WAVEFORNM OF THE ELECTRO-GLOTTOGRAM

The relationship between the vibration of the vocal folds and nature of the electro-glottogram had been investigated by the researchers on the electro-glottography through simultaneous recordings of opening and closing of the glottis observed by the fiber scope and the optical glottography, and also through modelling of the vocal fold vibration [4, 5 and 6].

Referring to their discussions, it was examined that the higher and narrower peak (or lower flat valley) in each fundamental period of the waveform of the electro-glottogram which corresponds to the tighter and shorter contact of the glottis, and the steeper rise of the curve which correspond to the quicker increase of the contact, could be used as indications of richness of the higher harmonic components of the voice source in the training of voicing (Figure 2).

The lack of the higher harmonic components in the range of lower formant frequencies results in a significant defect in the speech sound. This is one of the most difficult aspects in the articulatory training of the hearing-impaired having defective voice quality.

PROCESS OF IMPROVEMENT OF THE VOICE QUALITY

In order to find a subject for the preliminary experiment of applying the electro-glottography to the speech training as a visual feedback, firstly, eight hearing-impaired among forty (aged 19 and 20 years) who were staying in the Department of Vocational Training, Training Center of the National Rehabilitation Center for the Disabled were selected. They met the condition of; having hearing level of over 100 dB, poor speech quality, and consequently being required of integrated speech training. After analyzing their speech, a female, aged 19, who had defective voice quality but rather good articulation was chosen as the subject.

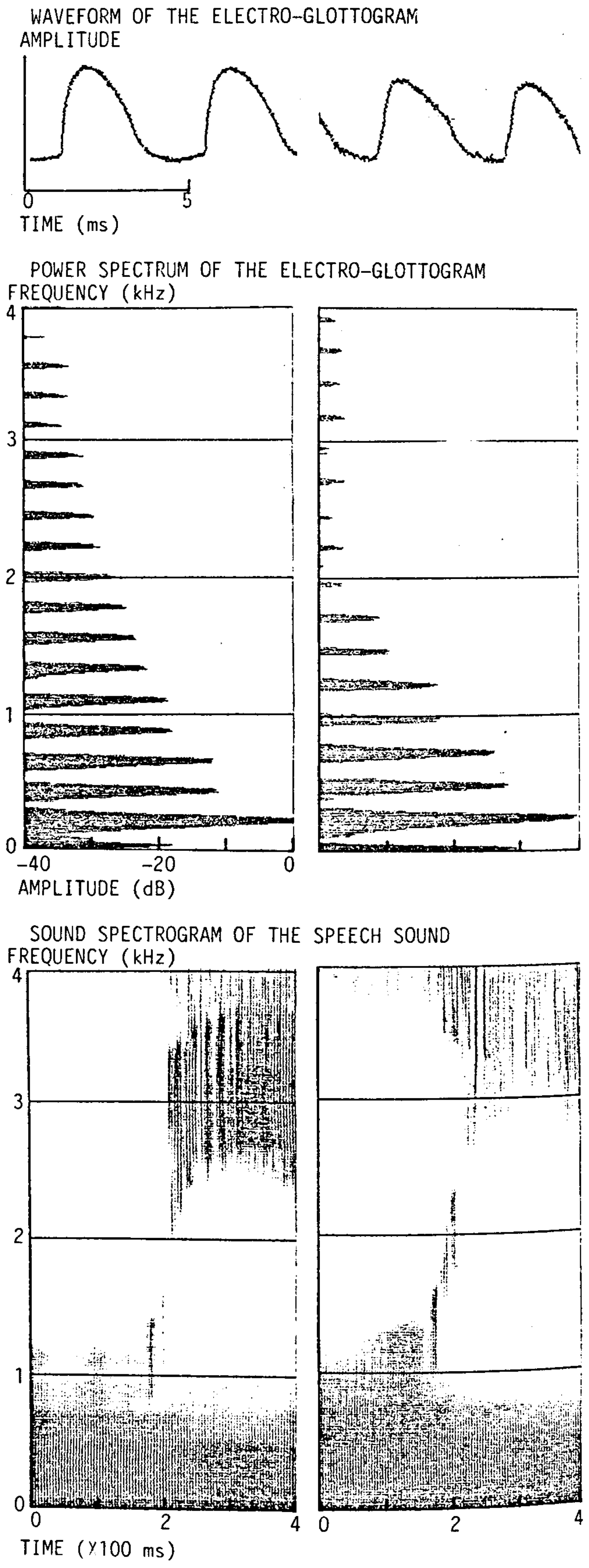


Figure 2. A pair of examples of the electro-glottograms and their power spectrums for a normal and a defective voicing, which was simulated by a female adult, and sound spectrogram of the speech sound for the utterance of Japanese vowel sequence.

Before the training, the voice of the subject in daily conversation was abnormally high pitch and low loudness, and the tonal qulity was too soft and close to falsetto. For these reasons, the phonemic aspect of speech was not acceptable, even though her articulation was fairly good as she had had speech training in the school for the deaf where she stayed for the previous twelve years. As for the prosodic aspect, the subject spoke in slow tempo with ambiguous word accent, sentence intonation and emphasis of phrase. The waveform of the electro-glottogram did not show wide flat valley or steep rise in each fundamental period, consequently, the harmonic components were found only in the low frequency range (Figure 3a). This is known as one of the common characteristics of the speech of hearing-impaired.

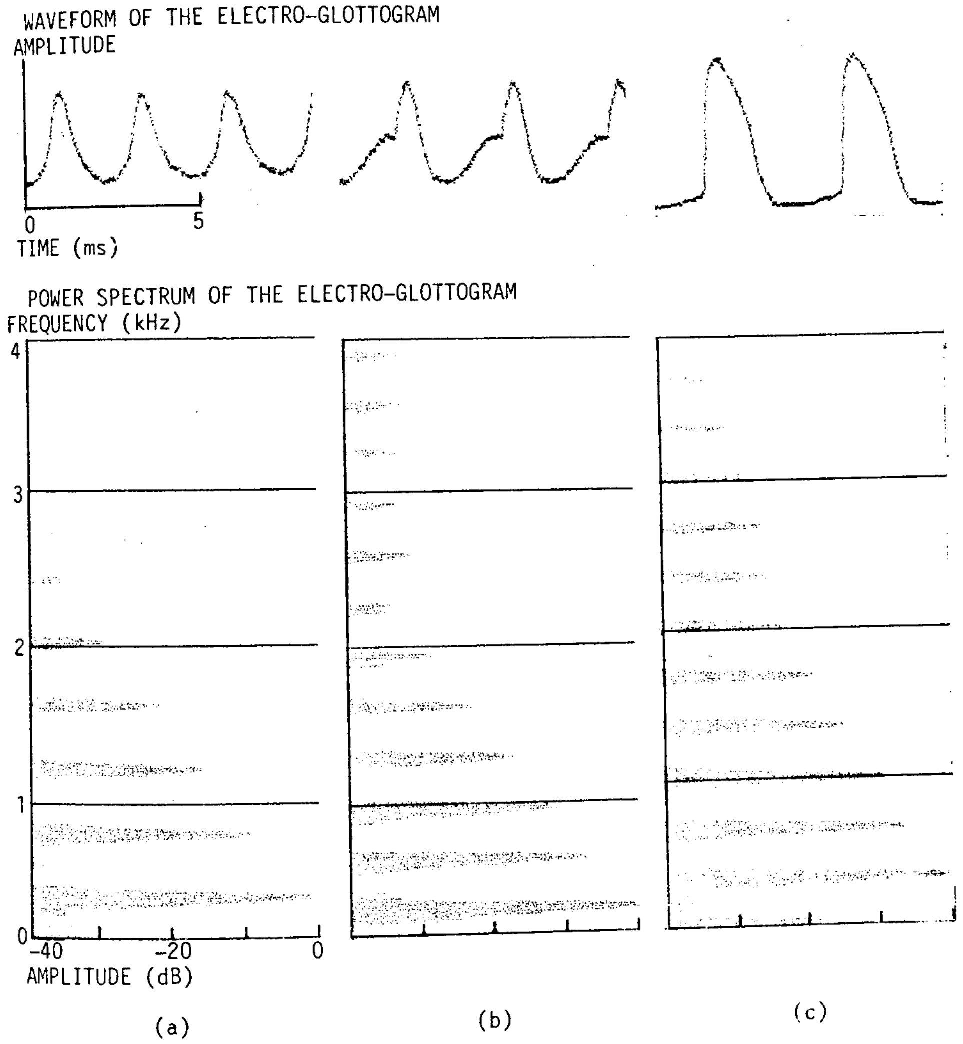


Figure 3. A set of examples of the electro-glottogram and power spectrum in the process of improvement of voice quality by the hearing-impaired subject.

In the preliminary experiment of training of voicing, the subject was instructed to sustain vowel phonation by monitoring the electro-glottogram on the display, and to imitate the instructor's typical waveform especially marking steepness of the rise of the curve in each fundamental period. And the process of improvement of the quality of voice was evaluated based on the degree of richness of the higher harmonic components of the voice source through a spectrographic analysis of both the electro-glottogram and speech sound.

Soon after beginning the training, the subject was able to change the nature of the waveform of electro-glottogram by laryngeal adjustment. One way to produce a steep rize of the curve by ther subject was abnormally tensed phonation. Although the higher harmonic components became richer, the voice quality was unnatural for speech sound (Figure 3b). This is also common to the voicing of hearing-impaired, however, this needs to be checked by the tests other than the sound spectrogram.

A series of training which consisted of two sessions a week, a session being about one hour long, was conducted. After several sessions, the waveform of electro-glottogram become almost normal, resulting in the improved voice quality (Figure 3c). The fundamental frequency became lower towards normal range.

The improvement was achieved for the open vowels [o] and [a] first, but it took another several sessions to stabilize the result, and to achieve a similar improvement for other vowels, particularly for [i] which was the most difficult among the five Japanese vowels.

APPLICATION TO INTEGRATED SYSTEM OF TRAINING

In this study, it was ascertained experimentally that the hearing-impaired subject was able to adjust her voice quality through the electro-glottographic display. Parallel with this training of voice quality, a series of training for refining the articulation was conducted in sequence of vowels, semi-vowels, nasals, flapped, voiced plosives, and other Japanese consonants. Training to achieve a reasonable pitch control for such as Japanese word accent and sentence intonation began when the range of voice pitch of the subject became normal after the series of training of voice quality.

In this way, the electro-glottography for training of voice quality and various other methods for training, for control of pitch and loudness of voice through displays of changes in fundamental frequency and intensity of speech sound, and articulatory training by use of displays of lip movement [7] and lingual contact to palate [8 and 9], were assembled into an integrated program.

It is planned to combine this program of speech training with a system of objective evaluation of speech quality based on acoustical analysis [10], and to extend the range of application to hearing-impaired children in the future.

REFERENCES

[1] A. Boothroyd and M. Decker: "Control of voice pitch by the deaf: An experimental study using a visible speech device," Audiology, 11, 343-353, 1972.

[2] E. Abberton, A. Parker and A. J. Fourcin: "Speech improvement in deaf adults using laryngographic displays," Speech, Hearing and Language, Department of Phonetics and Linguistics, University College London, 1976.

[3] A. J. Fourcin and E. Abberton: "First applications of a new laryngograph," Medical and Biological Illustration, 21, 172-182, 1971.

[4] R. J. Berry, R. Epstein, A. J. Fourcin, M. Freeman, F. MacCurtain and N. Noscoe: "An objective analysis of voice disorders. I," British Journal of Disorders Communication, 17, 67-76, 1982.

[5] R. J. Berry, R. Epstein, M. Freeman, F. MacCurtain and N. Noscoe: "An objective analysis of voice disorders. II," British Journal of Disorders Communication, 17, 77-85, 1982.

[6] D. Childers: "A critical review of electro-glottography," Biomedical Engineering, 2, 131-161, 1985.

[7] S. Hiki and Y. Fukuda: "Mouth shape in the production of [w] and [o] sounds in Japanese," Proceedings of the 9th International Congress of Phonetic Sciences, Section 1, 188, August, 1979, Copenhagen.

[8] S. Hiki and H. Itoh: "Influence of palate shape on lingual articulation," Speech Communication, 5, 141-158, 1986.

[9] S. Yamashita, S. Shibata and N. Murata: "Computer assisted therapeutic package for articulation control (CATPAC)," Logopedics and Phoniatrics: Issues for Future Research, Proceedings of the XXth Congress of the International Association of Logopedics and Phoniatrics, 204, August, 1986, Tokyo.

[10] Y. Fukuda: "Objective evaluation of the prosodic aspects of speech of hearing-impaired children based on acoustical analysis," Logopedics and Phoniatrics: Issues for Future Research, Proceedings of the XXth Congress of the International Association of Logopedics and Phoniatrics, 520-521, August, 1986, Tokyo.

INCORPORATION OF THE FORTIS-LENIS FEATURE IN A QUASIARTICULATORY SYSTEM OF TACTILE SPEECH SYNTHESIS BY ADDING TEMPORAL VARIATIONS

HANS GEORG PIROTH

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

ABSTRACT

A method for electrocutaneous speech synthesis was developed using pulse train sequences with variable intervals that are delivered to 16 electrode pairs along the forearm. The coding was 'quasiarticalatory' in that places of articulation (front - back, high - low)' were mapped quasi-isomorphically to the forearm (distal - proximal, dorsal - volar).
By varying the repetition rate of the pulse bursts (faster - slower) a tactile fortis-lenis equivalent was incorporated and, additionally, a plosive-fricative distinction was defined. So the inventory of tactile consonants was expanded to cover the whole range of the obstruent system of a language such as German. Exp. I uses tactile fricative-vowel equivalents, Exp. II plosive-vowel equivalents to test the learnability of such patterns.

INTRODUCTION

There is a long history of experimental investigations in the field of tactile speech transmission (e.g. [2,3,4,8]). Most of them used mechanical or electrical stimulation devices to transform the acoustic parameters of the speech signal into tactile patterns. The fact that most of these systems failed to reach the level of practical use, demands general reconsideration. According to our point of view in all these investigations the role of articulatory gestures for speech perception seems to be underestimated. Since tactile and proprioceptive reafferent control is present during the period of language acquisition, one may assume that normal speech perception is at least partially governed by the perception of articulatory guestures [1,9]. So it may be argued that a transformation of articulatory rather than acoustic information to the skin would provide a better opportunity to develop a successful tactile speech transmission system [6,7,10].
In its final state a quasiarticulatory system for tactile speech transmission would consist of four components:

(1) Registration of a speaker's acoustic signal.
(2) Analysis of the articulatory parameters from the speech wave.
(3) Transformation of the articulatory information into quasiarticulatory coded tactile patterns.
(4) Presentation of tactile patterns.

The investigation reported here is concerned with the development of the quasiarticulatory coding of tactile stimulus patterns. To yield an approximately geometric mapping of the places of articulation the forearm was selected as the tactile stimulation area. The experiments were executed with the 'System for Electrocutaneous Stimulation' (SEHR-2) presenting current-controlled bipolar pulse train sequences of the basic form shown in Piroth/Tillmann 1984, Fig. 1 [5]. The sixteen channels of the stimulation device (cf. Tillmann/Piroth 1986 [10]) were connected with 16 pairs of round gilded brass electrodes (9 mm in diameter). The smallest distance between the electrodes of a pair was 1 mm.
The electrode arrangement and the order of successive stimulations was defined according to a set of basic criteria for the coding method. First, complete tactile patterns are syllable analogues, i.e. each syllable is in one-to-one correspondence to a complete pattern. Second, a complete pattern is composed of partial patterns representing the consonant and vowel phonemes. In general, vowel patterns move longitudinally along the arm, consonantal patterns circumferentially. (Fig. 1 shows the arrangements of electrodes as well as the stimulation area of the central vowel /ə/.) Third, places of articulation are mapped to the place of tactile stimulation so

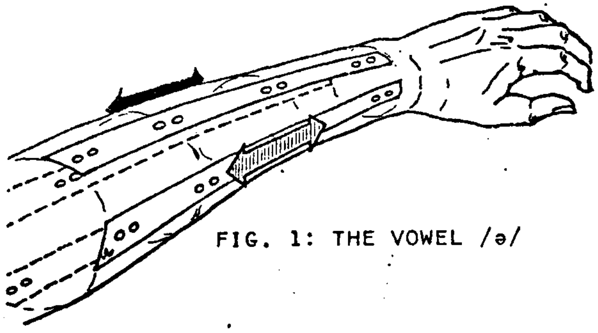


FIG. 1: THE VOWEL /ə/

that front articulations (of vowels and consonants as well) correspond to distal patterns near the wrist and back articulations to proximal patterns near the elbow. High vowels are mapped to the dorsal side. low vowels to the volar side of the forearm. Intermediate places of articulation are coded by stimulating the intermediate tactile areas. Fourth. as the starting point of the circumferentially moving consonant patterns depends on the stimulation area of the preceding or following vowel. an rudimentary form of 'coarticulation' is implemented in the coding method. Fifth. in the present experiments the temporal duration of the tactile syllable equivalents correspond to those of extremely slow and very explicitly uttered natural syllables. Nevertheless. patterns are constructed in a way that overall duration can be shortened by omitting pulses from the pulse train sequences without hereby altering the phenomenal 'gestalt' of the patterns. Former investigations have shown that vowel patterns are easily identified even by untrained subjects and that consonantal places of articulation -although identification is not as good- are recognized well above chance level without training. The following experiments include the fortis-lenis- and the plosive-fricative distinction into the system of consonantal patterns to yield a construction method for the complete system of obstruents, and they use a learning paradigm to improve the identification results by training.

EXPERIMENT I

According to the basic criteria a system of tactile fricatives and the vowel /ə/ was constructed and combined to form syllables. Tactile /fə:/. /sə:/, /ʃə:/. /çə:/, /və:/, /zə:/, /ʒə:/, /jə:/ and their VC-equivalents were presented in a learning test to reveal whether identification of CV-equivalents can be improved by learning. A succeeding control test was run to show whether a transfer of learned skills enhances the identification of the VC-patterns.

EXPERIMENT II

In the same way a plosive-vowel system was used consisting of tactile /pə:/. /tə:/, /kə:/, and /bə:/. /də:/, /gə:/.

Table 1
Fricative Vowel System

	Number of Taps	Tap-duration (ms)	ITI (ms)	Overall duration (ms)
V	8	5.2	20	201.6
FF	8	5.2	15	161.6
LF1	4	5.2	35	160.8
LF2	8	5.2	5/30	181.6

V: Vowel. FF: Fortisfricative. LF1: Lenisfricative (1 ring: simple pattern. LF2: 2 rings: complex pattern).

GENERAL METHOD

Stimuli.
Pulse trains ('taps') of three pulses having the form described above with a constant pulse width of 200 µs. a variable amplitude, a constant inter-pulse-onset interval of 2.5 ms and an overall duration of 5.2 ms were used as basic stimuli. They were arranged to sequences in which the places of stimulation are changed according to the basic criteria cited above. So. the tactile syllable equivalents consisting of fricative patterns and a /ə/-pattern were constructed. Number of taps. inter-tap intervals (ITI). tap-duration and overall duration of the patterns are given by Tab. 1 for each type of stimulus. The local shifts along the first or first and second electrode rings (i.e. the distal ones) in the fricative patterns /f/ and /v/ are shown in Fig. 2, of /s/ and /z/ in Fig. 3.

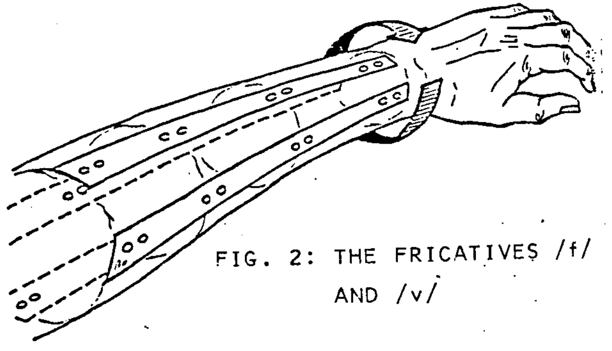


FIG. 2: THE FRICATIVES /f/ AND /v/

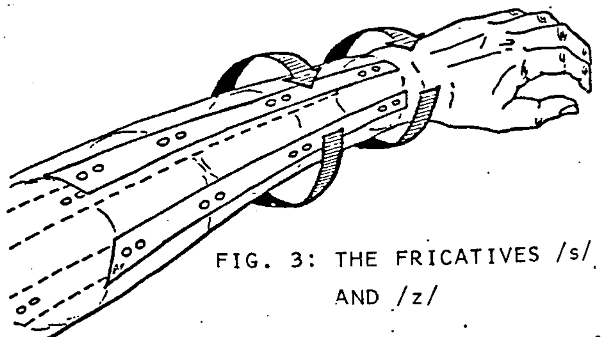


FIG. 3: THE FRICATIVES /s/ AND /z/

/ʃ/ and /ʒ/ resemble /f/ in being simple patterns consisting of one circumferential tap sequence only, but they are constructed as surrounding the second electrode ring instead of the first. /ç/ and /j/ like /s/ and /z/ are complex patterns (two rings) which are presented quasi-simultaneously (i.e. with the proximal ring having a delay of 5 ms) at the second and third electrode rings. As shown in Tab. 1. the fortis-lenis difference is encoded in the inter-tap interval of the sequences: fast moving patterns are fortis. slowly moving ones lenis. (Preliminary investigations [6] had shown that a difference in ITI of 20 ms is perceivable in similar patterns presented without a context.) To keep overall duration constant. the number of taps was halved in the case of /v/ and /ʒ/. Since /z/ and /j/ are presented by stimulation of 8 loci this was not possible. So /z/ and /j/ are 20 ms longer in overall duration. The neutral vowel /ə/ is transformed into an 8-tap pattern moving along the mid pairs of radial and ulnar electrode rows (Fig. 1).
Plosives are built as patterns sweeping between neighbouring electrode pairs as shown in Fig. 4 which represents /p/ and /b/. Analoguously. /t/ and /d/ surround the second ring. /k/ and /g/ the third.

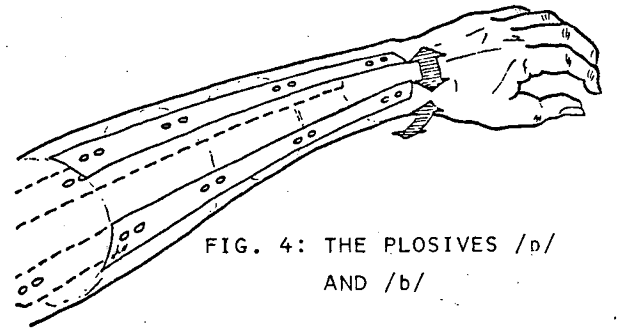


FIG. 4: THE PLOSIVES /p/ AND /b/

Regarding the basic criteria information of the starting point of the vowel pattern is preserved by starting the circumferent pattern at the same electrode row where the following vowel pattern starts. The velocity of the pattern (5 ms/ITI) is clearly below the threshold for discrimination of successive taps and is not used to carry the fortis-lenis information. Instead. this feature is encoded in the neighbouring half of the vowel-pattern as listed in Tab. 2.

Subjects and Procedure.
Four unexperienced Ss participated in Exps. I and II. All Ss were tested singly and were informed about the details of the learning test series and the coding method.
1. Intensity adjustment. Each test session started with a calibration procedure. Ss were asked to adjust subjective intensity to a mid value between absolute threshold and unpleasentness for each place of stimulation. The resulting values were taken as impulse amplitude values in the immediately following test runs.
2. Presentation of the patterns. The inventory of the 8 (or 6) CV-patterns was presented five times in a systematic order to the S as following: for each pattern a phonological transcription was presented via terminal followed by the tactile pattern corresponding to this syllable. After an interval of 4 s the transcription of the next syllable was presented.
3. Feedback tests. For a single FB-test the 8 CV-patterns were presented 6 times in Exp. 1 and the 6 CV-patterns in Exp. 2 were presented 8 times in completely randomized order to yield 48 presentations. By pressing a key on the computer keyboard the S started the presentation of a pattern. After an interval of 1s the S had to name the just presented syllable via keyboard. Then the transcription of the syllable that was presented was given to inform the S whether his answer was correct or not. After the presentation of 5 equal test runs (i.e. 30 or 40 repetitions of each pattern) the test session was finished. The Ss underwent 5 equal test sessions with a pause of 1 or 2 days between each two successive sessions. Finally. in a sixth (control-) session structured in the same way the whole inventory was presented in VC-ordering. Two of the four Ss first underwent Exp. I, the remaining two Ss first Exp. II.

Table 2
Plosive vowel System

	Number of Taps	Tap-duration (ms)	ITI (ms)	Overall duration (ms)
FP	16	5.2	5	163.2
LP	16	5.2	5	163.2
FP+V	16	5.2	8x15 +8x25	403.2
LP+V	12	5.2	4x35 +8x25	402.4

V: Vowel. FP: Fortisplosive (burst).
LP: Lenisplosive (burst).

RESULTS AND DISCUSSION

Fig. 5a presents the average identification rate for all subjects in Exp. I. Fig. 5b gives the computed results that show the recognition of the fortis-lenis feature. (Identification of a fortis-pattern was assumed to be correct when the S after presentation of a fortis-pattern answers with a fortis consonant.)

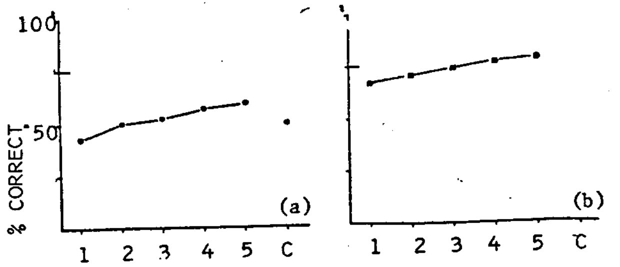


Fig. 5: Results of Exp. I (a) Syllable
(b) Fortis-Lenis Identification

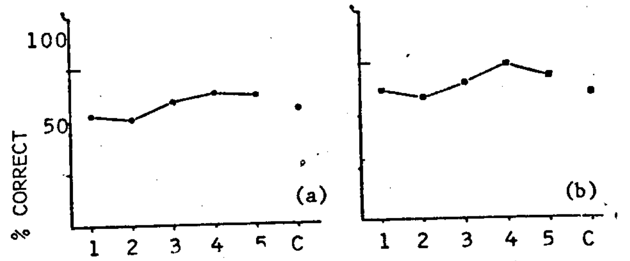


Fig. 6: Results of Exp. II (a) Syllable
(b) Fortis-Lenis Identification

Figs. 6a and b present the results of Exp. II in the same way. To evaluate the effects after a transformation of the dependent variable by

$$y = \text{arc sin } (x/100)^{1/2}$$

a one-factrorial univariate analysis of variance was calculated, first as a trend analysis of sessions 1 to 5 by the method of orthogonal polynomials, then the analysis was expanded to 6 sessions to determine the relevant a priori contrasts. Since the analysis was repeated with fortis/lenis results the level of significance $\alpha = 0.05$ was lowered to α^* by $\alpha^* = 1 - (1-\alpha)^{1/2}$. A highly significant variation over the 5 CV-sessions was found in all cases yielding a linear trend in Exp. I and a cubic one in Exp. II (Tab. 3). For the analysis of contrasts the comparisons of 1st and 5th, 1st and 6th and 5th and 6th session were chosen. According to Tab. 3 the contrasts between the 1st and the 5th CV-session are always significant and show a consistent learning effect. But the results indicate that there is no transfer of learning from the CV-series to the VC-session: in all cases the contrast between the first CV-session and the VC-session is not significant. With a simple exception, the results of the 5th CV-session and the VC-session differ significantly. This may be due to the fact that only one control-session was run. A series of experiments is in preparation to show whether a transfer of learning is possible if the S has to manage more trials in the altered condition.

Table 3
Results of Exps. I and II

Trend analysis by orthogonal polynomials (sessions 1-5):

Exp. I Syllables	F(4.95)= 4.421	p<0.005**
Linear trend	F(1.95)=17.076	p<0.005**
Exp. I Features	F(4.95)= 6.187	p<0.005**
Linear trend	F(1.95)=23.982	p<0.005**
Exp. II Syllables	F(4.95)= 9.869	p<0.005**
Cubic trend	F(1,95)=10.288	p<0.005**
Exp. II Features	F(4.95)=10.366	p<0.005**
Cubic trend	F(1.95)=18.379	p<0.005**

A priori contrasts (sessions 1-6):

Exp. I Syllables				
1 vs. 6:	t=-0.725	df=114	p=0.470	n.s.
1 vs. 5:	t=-3.829	df=114	p=0.000	**
5 vs. 6:	t= 3.104	df=114	p=0.002	**
Exp. I Features				
1 vs. 6:	t=-0.762	df=114	p=0.447	n.s.
1 vs. 5:	t=-4.102	df=114	p=0.000	**
5 vs. 6:	t= 3.440	df=114	p=0.001	**
Exp. II Syllables				
1 vs. 6:	t=-1.552	df=114	p=0.124	n.s.
1 vs. 5:	t=-3.253	df=114	p=0.002	**
5 vs. 6:	t= 1.701	df=114	p=0.092	n.s.
Exp. II Features				
1 vs. 6:	t=-0.013	df=114	p=0.989	n.s.
1 vs. 5:	t=-2.418	df=114	p=0.017	*
5 vs. 6:	t= 2.404	df=114	p=0.018	*

Reduced level of significance:
p<0.00501 **   p<0.02532 *

REFERENCES

[1] C.A. Fowler, "An Event Approach to the Study of Speech Perception from a Direct-Realistic Perspective", J. Phon. 14, 1986, 3-28.

[2] R.H. Gault, "An Experiment on Recognition of Speech by Touch", J. Wash. Acad. Sci. 15, 1925, 14.

[3] M.H. Goldstein, R.E. Stark, "Modification of Vocalizations of Preschool Deaf Children by Vibrotactile and Visual Display", J. Acoust. Soc. Am. 59, 1976, 1477-1481.

[4] R. Lindner, "Physiologische Grundlagen zum elektrischen Sprachetasten und ihre Anwendung auf den Taubstummenunterricht". Zeitschr. f. Sinnesphysiologie 67, 1937.

[5] H.G. Piroth, H.G. Tillmann, "On the Possibility of Tactile Categorical Perception", M.P.R. v.d. Broecke, A. Cohen, Proc. 10th ICPhS, Dordrecht 1984, 764-768.

[6] H.G. Piroth, "Elektrokutane Silbenerkennung mit quasiartikulatorisch kodierten komplexen zeitlich-räumlich strukturierten Reizmustern", Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München 22, München, 1985.

[7] H.G. Piroth, "Electrocutaneous Syllable Recognition Using Quasiarticulatory Coding of Stimulus Patterns" (Abstr.), J. Acoust. Soc. Am. 79, 1986, S73.

[8] D.W. Sparks et al., "Investigating the MESA (Multipoint Electrotactile Speech Aid): The Transmission of Segmental Features of Speech". J. Acoust. Soc. Am. 63, 1978, 246-257.

[9] H.G. Tillmann, "Phonetik und Phonologie sowie die natur- und geisteswissenschaftliche Erforschung der gesprochenen Sprache", FIKPM 19, 1984, 9-32.

[10] H.G. Tillmann, H.G. Piroth, "An Order Effect in the Discriminability of Pulse Train Sequences" (Abstr.), J. Acoust. Soc. Am. 79, 1986, S73.

ACKNOWLEDGEMENTS

This investigation was supported by the German Research Council (DFG-grant Ti 69/25).
Statistical advice was given by Dr. Alexander Yassouridis, Department of Biostatistics, Max-Planck-Institute for Psychiatry, Munich F.R.G.

VISUAL INFORMATION AND SPEECH ACQUISITION OF THE DEAF

DIRK-JAN POVEL

Psychological Department
University of Nijmegen
P.O. Box 9104
6500 HE Nijmegen, The Netherlands

BEN MAASSEN

Interdisc. Inst. of Child Neurology
University of Nijmegen
P.O. Box 9101
6500 HB Nijmegen, The Netherlands

ABSTRACT

Substantial improvement of speech of the deaf can only be obtained if we succeed in providing them with additional information concerning speech through other than auditory channels. Since there have been many attempts at developing visual aids but very little success [2], we consider some basic issues involved in the development of visual aids for speech training and propose a number of assumptions that guide our undertaken to construct an effective visual aid.

INTRODUCTION

Currently we are involved in a project in which we are developing computer controlled visual aids displaying acoustic information of speech to be used in speech acquisition of the deaf. The basic motivation behind the project is the belief that the only way to substantially improve the results of speech training of the deaf is by introducing (visual) aids that supply additional information and feedback about speech, which can be incorporated in a speeech training program. We belief that the deaf child can only form an adequate and stable internal representation of speech if we supply them information that shows what speech looks like and provides them with an opportunity to examine the results of all sorts of articulatory gestures and to check the successfulness of attempts to produce specific speech acts. Because of the enormous growth in recent years of computational power and graphics facilities, it is now possible to design all sorts of visual aids, evaluate them in a training situation and subsequently adjust the design on the basis of this evaluation, in a most flexible way.

However, the more possibilities there are, the more decisions have to be taken about different design aspects of the aids. Therefore, we present in this paper a preliminary framework that allows to see the different dimensions of the problem and indicates the type of questions that should be asked (and answered).

TWO BASIC QUESTIONS

To develop a visual aid means to answer two questions: WHAT information to display and HOW to display it. Attempts to answer these questions lead to the most fundamental aspects of speech perception and production as well as to basic questions concerning the essential differences between the processing performed by the ear and the eye. The answer to these questions is in part also determined by the view one holds with respect to the method of speech education of the deaf. Although educational aspects will undoubtedly play an important role in the ultimate form of the aids, in this paper we will not deal with these aspects but confine ourselves to the more fundamental issues relating to how relevant aspects of speech can be made visible for speech training purposes.

THE ULTIMATE GOAL: THE IDEAL SPEECH VISUALIZER

What would be the ultimate goal of a project like this? As we see it, the ideal speech visualizer for speech training should completely take over the missing auditory function of the deaf child.

In order to attain this ideal two problems should be solved. First, we must find ways to present the acoustic information about speech in a form which is digestible by the eye. In view of the great differences between the way information is processed by the ear and the eye, it will be necessary to do a lot of preprocessing that transforms the acoustic information into a form suitable for the eye. Secondly, we will have to find ways to inform the child how the visual display relates to speech. For the deaf child acquiring speech needs to understand how different visual dimensions and combinations of these dimensions are related to speech production. Even more importantly, information must be supplied about how these dimensions are used to produce different speech-related acts. This means that the display must also present normative information. We will return to this point below.

To what extent this ideal can be realized is at present unpredictable.

But even a superficial study of the differences between visual and auditory perception on the one hand and of the coding of speech in the acoustic signal on the other, makes clear that this is a most complicated undertaken.

A MORE PRACTICAL APPROACH: SOME BASIC ASSUMPTIONS

In order to make the problem somewhat more manageable we will conceive of the speech signal as describable in terms of a limited number of parameters that are related to basic aspects of speech production. The traditionally used parameters associated with intensity, fundamental frequency, timing and spectral composition seem most useful because they can, after some transformations, be related to the basic aspects of speech production: respiration, phonation and articulation. As a first step towards developing a theoretical frame we will formulate some assumptions with respect to the way acoustic information about speech should be mapped in the visual mode. These assumptions form together the starting-point for our approach followed in this project. As such, the assumptions can be seen as requirements for the visual aids to be developed.

Assumption 1

There must be a unique and fixed relation between acoustic and visual parameters. This assumption is based on the consideration that a stable internal representation can only be formed if different dimensions of the process are uniquely connected to specific aspects of the visible output. It implies that one should not use the same diagram to display different acoustic parameters. It also implies that if a certain acoustic feature is once associated with some visual attribute this should not be changed later.

Assumption 2

The visually presented information concerning speech should be as complete as possible. The following arguments form the basis for this assumption. In the first place having all information available concerning a skill to be developed is the natural situation, independent of the question whether the subject uses all the information all the time. Secondly, there is a practical argument which is based on experiences with visual aids that display only one feature. In working with a device that displays for instance fundamental frequency, one will inevitably be confronted with the child that does produce the required pitch or intonation contour but only at an intensity of 110 dB or with a very bad voice quality. This problem is inherent of mono-feature displays and can only be solved by displaying all relevant information simultaneously.

Assumption 3

The visual aid should display the information in an integrated fashion rather than in a parallel one. The main argument for this assumption is that if the information is displayed in parallel, for instance in different windows on the screen, somewhat like the dials on the dashboard of a car, the subject will probably have great difficulties to monitor all the information simultaneously. Therefore we believe that we should try to develop a system that displays all relevant information in one multi-dimensional form, making use of different independent visual dimensions like form, colour, texture, size etc. A considerable body of research has shown that information presented in independent dimensions (such as those just mentioned) are processed almost in parallel. That is, with two dimensions one can convey almost twice as much information as with one [1].

THE DESIGN OF AN AID

In this way we approach the ideal formulated before, in the sense that all relevant information is displayed in a form that is easily accessible to the eye. But, as mentioned above, this is still only part of the answer, because displaying the information in itself does not say anything about the meaning of the information; it does not show how the information is related to speech. It should be noted though that a straightforward display of speech related information, without any normative function, can be a most useful aid since it can help the deaf child to learn the correspondence between visual and articulatory dimensions.

When those relations are acquired the pupil has learned to control specific aspects of the visual display and thus may be said to have developed some aptitude in controlling articulatory structures relevant for speech production. But the child still does not know the relation between the visual image and speech. So now the pupil must be shown how the different visual dimensions are used in the formation of specific speech acts. Here we can think of giving information about the range of values that are used in speech production, like the range of acceptable intensity, fundamental frequency, F1, F2 etc.

The introduction of normative information

One way to introduce a norm is by using a split screen and showing an example of a model speech act on the upper half, which can be imitated by the pupil on the lower half. The models can either be produced on the spot by the teacher or can be build into the apparatus. Although the teacher-produced model is attractive in the sense that it fits within the usual teacher-pupil interaction, in the case of speech training it has several limitations. For instance in displaying intensity-related aspects of a model the distance between speaker and microphone is crucial, but difficult to control especially if the child uses a headphone-microphone combination. Other aspects like fundamental frequency and timbre are even more problematic in this respect because the ranges of values on these dimensions differ greatly between adults and children. To normalize these differences does not seem easy to accomplish. Therefore at present we concentrate at displaying internally stored criteria of acceptability.

Suppose we display intensity of speech as the brightness of the display on the screen. Then, speaking too softly or too loudly would be indicated by the display becoming almost invisible, respectively unpleasantly bright (with well chosen relation between intensity and brightness). This example shows that with a well chosen visual dimension, the normative information can be presented in a most natural way.

For other aspects of the speech signal this may not be readily feasible. Consider for instance the way normative information is build into the Vowel Corrector developed by Povel [4] and Povel & Wansink [5], an aid for teaching vowels to the deaf. This device displays vowels, either spoken in isolation or in mono-syllabic words, as light spots on a screen such that different vowels project at different areas of the screen. When vowels are entered into this apparatus the spot moves over the screen roughly in accordance with the momentary value of F1 and F2 that respectively determine the X and Y coordinate of the spot. In this mode the device only displays some speech-related aspects of the spectrum, but does not indicate the relation between location of the spot on the screen and speech characteristics. This information is presented by indicating on sheets fixed to the screen the outlines of the areas that correspond to different vowels.

It should be noted that this way of displaying spectral information seems most useful because it combines the two functions mentioned above: it shows relevant parameters of speech in a way that is easily interpretable by the eye, and at the same time it shows the relation between the displayed information and certain speech acts, thus fulfilling its normative function.

MOTIVATION

Besides the two functions just discussed, there is yet another aspect that is probably very important in displaying visual information for speech training purposes. This concerns the desirability that the information be presented in a for the child attractive way, for instance in the form of interesting games, thus maintaining motivation during training. Although we believe that this aspect needs attention, we feel that it is even more important to construct a curriculum incorporating the aid, in which tasks are defined that the child can perform successfully, thus maintaining inherent motivation to learn to speak.

CONCLUSION

To summarize we believe that in displaying visual information for speech training purposes, one should aim at displaying as much relevant information as possible in an integrated visual display using independent visual dimensions that are uniquely related to speech parameters. Further, the device should incorporate norms as to how the different dimensions are used in forming specific speech acts. All these aspects should be part of a training curriculum in which attention is given to factors stimulating motivation. Apart from the specific problems to display the separate parameters, we think that the main challenge will be to combine the different requirements in a workable visual aid.

Currently we are working on two parallel lines. In the first one we develop aids for separate aspects of speech. Here we concentrate on displaying segmental, rather than on supra-segmental information on the basis of the results of the work of Maassen & Povel [3] which has shown that an improvement of intelligibility is mainly found after correcting segmental aspects of speech. In the second line we are building aids that combine different aspects in one complex multidimensional display. Examples of displays will be shown during the presentation.

REFERENCES

[1] Attneave, F. (1959). Applications of information theory to psychology. New York, Holt, Rinehart and Winston.

[2] Lippmann, R.P. (1982). A Review of Research on Speech Training Aids for the Deaf. In: N.J. Lass (Ed.). Speech and Language Advances in Basic Research and Practice. Vol. 7. New York, Academic Press.

[3] Maassen, B., & Povel, D.J. (1985). The Effect of Segmental and Suprasegmental Corrections on the Intelligibility of Deaf Speech. Journal of the Acoustical Society of America, 78, 877-886.

[4] Povel, D.J. (1974). Development of a Vowel Corrector for the Deaf. Psychological Research, 37, 51-70.

[5] Povel, D.J., & Wansink, M. (1986). A Computer Controlled Vowel Corrector for the Hearing Impaired. Journal of Speech and Hearing Research, 29, 99-105.

GLOTTAL DETERMINANTS OF DEAF VOICE QUALITY

BEN MAASSEN

Interdisc.Inst.of Child Neurology
University of Nijmegen
P.O.Box 9101
6500 HB Nijmegen, The Netherlands

DIRK-JAN POVEL

Dept.of Experimental Psychology
University of Nijmegen
P.O.Box 9104
6500 HE Nijmegen, The Netherlands

ABSTRACT

Vocal fold vibration of nine deaf children was recorded with help of an electro-laryngograph. Analysis of the laryngographic waveforms yielded several parameters that can be used as an objective measure of instability of voice and deviating voice quality characteristics like breathiness, hoarseness and cul-de-sac. The analysis algorithms will be implemented in a visual speech training aid for the deaf.

INTRODUCTION

Even highly trained, experienced phoneticians exhibit great variability in their evaluations of deaf speech characteristics. This is especially true for suprasegmental aspects and voice quality [8,4,13]. As indicated by the labels typically used in descriptions of deaf voices, such as "too high, monotonous, breathy, nasal, cul-de-sac", voice quality is conceived of as the overall auditory colouring of an individual speaker's voice, to which both laryngeal and supralaryngeal features contribute. The latter refer to long-term muscular adjustments or "settings" of the articulatory organs [9]. For instance, cul-de-sac or pharyngeal focus of resonance [2] is caused by a tendency to retract or "back" the lingual body. Differentiating aspects of voice quality and articulation is complicated by the differential susceptibility [9] of individual speech segments to the biasing effect of a given supralaryngeal setting. That is, a nasalized voice has a different effect on nasal sounds (/m,n,ŋ/) than it has on vowels, plosives, fricatives and affricates. Also the decomposition of voice quality i glottal pulse shape and supralaryngeal effects is problematic, especially when the description is based on perceptual judgment.

To analyze the purely laryngeal aspect of voice quality we have used the electro-laryngograph (ELG) [5]. The ELG is an instrument that measures the electrical impedance of the vocal cords, thereby providing information about opening and closure durations. The laryngographic signal has been validated as a measure of vocal fold contact area by comparing the signal with registrations of subglottal air pressure, and measurements of glottal opening by means of photoglottography and high-speed filming [1,3]. The most stable characteristic of the Lx signal is the steep slope corresponding to the beginning of the closure phase [10], which provides an easy reference point for determining glottal pitch period (see Figure 1). Apart from the fact that a simple algorithm suffices to extract fundamental frequency, a major advantage of using the Lx signal instead of the acoustic speech signal is that period-to-period fluctuations in the waveform can be detected. Thus, measures of jitter (period-to-period frequency fluctuations) and shimmer (period-to-period amplitude fluctuations) are easily obtained, thereby providing information on regularity of voice.

In this paper we present a study of nine deaf children that were selected by a speech therapist to represent a broad range of vocal abnormalities. Laryngographic recordings of these children were analyzed in an attempt to extract perceptually and articulatory relevant parameters. The analysis algorithms will be implemented in a visual speech training aid to be used in therapy [15].

VOICE QUALITY ANALYSIS

Recording Procedure

Nine congenitally deaf children were selected by their speech therapists to represent a broad range of voice abnormalities. These children, 7 boys and 2 girls, ranging in age from 9 to 15 years, with a hearing loss of more than 100 dB (Fletcher Index) in the better ear, read aloud a series of phonetically balanced sentences and words. The acoustic speech signal together with the Lx signal were recorded on two tracks of a Revox tape recorder. Since recording on tape introduces phase distortions, especially in the low frequencies, the Lx signal was re-recorded while running the tape in reverse. (The acoustic speech signal was also re-recorded to preserve temporal alignment on a single tape.) After low-pass filtering (cutoff frequency 4 kHz, slope 24 dB/octave) both signals were fed into a stereo A/D convertor (sampling rate 10 kHz for both channels) and stored in computer memory. In the present study only Lx signals were analyzed, but the figures also show the corresponding acoustic speech signals.

Apart from the nine deaf speakers, two adult, hearing speakers, one male and one female, were recorded and analyzed. These speakers differed only with respect to fundamental frequency. An excerpt of the male voice is presented in Figure 2a for comparison purposes.

Determining Pitch Periods

All analyses of the Lx signal were performed in the time domain. Isolating individual pitch periods started by calculating the first derivative (Lx'). If Lx' exceeded a criterium value, which was about one tenth of the Lx amplitude, that point was taken to correspond to a steep slope at the beginning of a new pitch period (see Figure 1) and marked accordingly. When no steep slope was found, a positive zero-crossing was taken instead. By setting a minimal spacing between period markings of 16 sample points (which corresponds to a maximum frequency of approximately 600 Hz) and a maximum spacing of 120 sample points (corresponding to a minimal frequency of about 80 Hz), and at the same time have steep slopes take precedence over positive zero-crossings, a reliable pitch detection algorithm was obtained. In a comparison of the outcome of the algorithm (the period markings) and the original Lx signal no errors could be detected. After isolation of single pitch periods, very low frequency-components were removed by subtracting from each sample point the mean value of the period it belongs to. The thus adjusted waveforms were further analyzed to obtain an objective description of deviations in voice quality.

Types of Voice Quality Deviations

We will now present the different types of deviating glottal pulse waveforms that occur in our speech samples, together with the measures to describe them.

1. Instability of voice. The accessibility of individual pitch periods permit detection of stability of voicing and laryngeal articulation [14] within a very short time window. A grossly instable voice is displayed in Figure 2b. Here, within a single period, fundamental frequency drops from 280 Hz to 135 Hz. The example is taken from a 14 year old boy, who typically produced such patterns at the beginning of voiced segments following silence.

In Figure 2c an articulation error is displayed. Another 14 year old boy attempted to say the Dutch word /bana:n/ ("banana"), but produced /banda:n/ instead. During the /d/ vocal fold vibration drops to zero, which may be caused by supraglottal pressure build-up during the erroneous complete closure of the vocal tract, or by an incorrect abduction-adduction gesture of the vocal folds.

2. Jitter, shimmer, low-frequency components. On the microlevel, irregularity of successive pitch periods is expressed by high jitter and/or shimmer values. Jitter is calculated by dividing two successive period durations, shimmer by taking the logarithm of amplitude ratios. Whereas for normal voices under sustained phonation jitter values of 0.5% - 1.0% and shimmer values below 0.20 dB are obtained [7], in the example of Figure 2d, produced by an unintelligible 9 year old boy, jitter and shimmer rise to 25% and 11 dB respectively, giving the voice a hoarse quality. In addition, Figure 2d shows a low-frequency component, indicating vertical displacement of the whole larynx. This may be caused by retraction of the tongue during cul-de-sac voicing.

3. Deviations of isolated waveforms. Figure 3a displays a breathy voice. Calculated were the relative closure duration (duty cycle), defined as the relative position within the pitch period of the negative zero crossing (i.e. C/L, see Figure 1), and the relative area below the positive curve (A/(L*P), see Figure 1). Like Hasegawa et al. [6], we took the cosine of the duty cycle, to magnify the important range around 0.5. In normal voicing the cosine of the duty cycle centers around 0, the relative positive area around 0.25; in this breathy voice we found values of 0.7 and 0.15 respectively.

Figure 3b displays a different type of breathiness. Here, the breathiness is caused by insufficient steepness of the positive slope. Relative closure durations were calculated by dividing the number of samples between the onset of the period (in most cases not far from the onset of closure) and the waveform peak by the pitch duration. Whereas a value of 0.10 is typical (see also [10]), in Figure 3b a relative closure duration of 0.25 was found.

In Figure 3c a terribly hoarse voice is displayed. To capture the irregular character of these waveforms, the number of times the second derivative (Lx'') exceeded a criterium value, was counted. During normal voicing, very few pitch periods contained direction changes exceeding the criterium; in the example of Figure 3c a mean number of 2 per period was obtained.

Finally, in Figure 3d a falsetto voice is presented. Note the sinusoidal character of the waveform. A falsetto voice is not only of very high pitch (540 Hz in this example) but also breathy according to the relative closure duration criterium (mean value 0.20).

DISCUSSION

Analyses of glottographic signals obtained from nine deaf children yielded several parameters that are related to voice quality. Fourcin (personal communication) displays "raw" Lx waveforms to deaf children for training purposes. Since we believe that interpretation of the Lx waveform is problematic, especially for the youngest age groups, we are currently implementing the analysis algorithms described above in a speech training device [15], such that voice quality can be displayed in a simplified visual trace. For instance, jitter and shimmer, which indicate hoarseness, might be transformed to the visual dimension texture.

The speech training curriculum proposed by Ling [11], in which teaching of respiration and phonation precede articulation, stresses the importance of simplifying voice quality and displays for young children. Moreover, in previous experiments [12] we showed that voice quality and articulation - rather than temporal structure and intonation contour - are the most important dèterminers of deaf speech intelligibility.

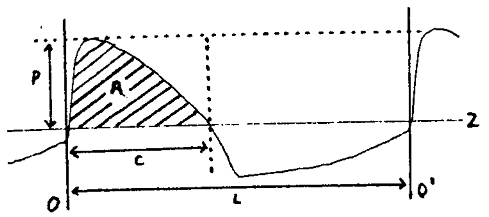


Figure 1. One period of a normal Lx waveform. Indicated are: 0,0': start of closure at the onset of pitch periods; L: length (number of samples) of one period; Z: zero-line; P: maximal positive value; C: closure duration (until negative zero-crossing); A: area below the positive part of the curve.

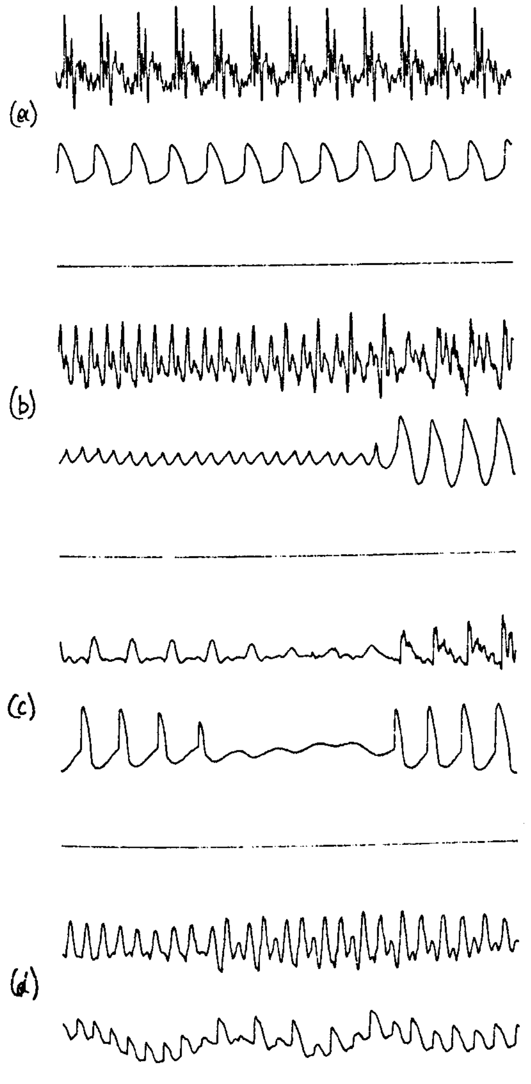


Figure 2. Sample registrations of Lx signals (lower traces) together with the acoustic speech signals (upper traces). Registration (a) id from a normal, male voice; (b), (c) and (d) represent incorrect period-to period fluctuations.

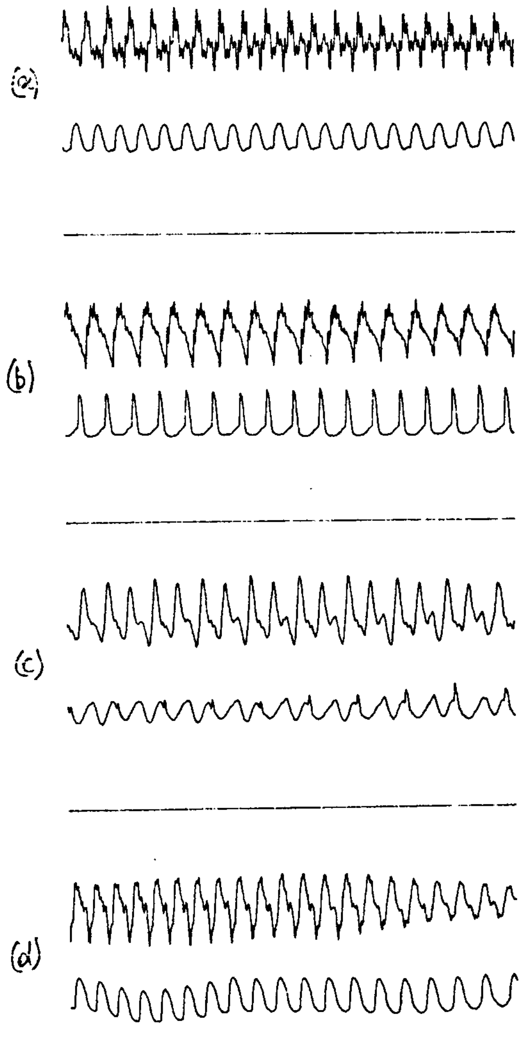


Figure 3. Sample registrations of Lx signals (lower traces) together with the acoustic speech signals (upper traces). In these examples the most noticeable deviation resides in the waveform of individual Lx periods.

Acknowledgement
This research was made possible by a grant from the Institute for the Deaf, St. Michielsgestel, The Netherlands.

REFERENCES

[1] Baer, Th., Löfquist, A. & McGarr, N.S. (1983). Laryngeal vibrations. A comparison between high speed filming and glottographic techniques. J.Acoust.Soc.Am., 73, 1304-1308.

[2] Boone, D. (1966). Modifications of the voices of deaf children. Volta Review, 68, 686 - 692.

[3] Boves, L. (1984). The phonetic basis of perceptual ratings of running speech. Doctoral dissertation, University of Nijmegen.

[4] Calvert, D.R. (1961). Some Acoustic Characteristics of the Speech of Profoundly Deaf. Unpublished doctoral dissertation, Stanford University.

[5] Fourcin, A.J. & Abberton, E. (1976). The Laryngograph and the Voiscope in Speech Therapy. In: E. Loebell (ed.) XVth International Congress of Logopedics and Phoniatrics, Basel: Karger.

[6] Hasegawa, A., Mashie, J., Herbert, E., Brandt, F., Metzler, M. M., Pickett, J. (1984). Real-time extraction of vocal quality parameters from electroglottographic signal. Paper presented at the 107th meeting of the Acoustical Society of America, Norfolk, Virginia, May 784.

[7] Heiberger, V. L. & Horii, Y. (1982). Jitter and shimmer in sustained phonation. In: N.J. Lass (Ed.) Speech and Language: Advances in basic research and practice (vol.7). New York, Academic Press.

[8] Hudgins, C. V., & Numbers, F. C. (1942). An investigation of the intelligibility of the speech of the deaf. Genetic Psychology Monographs, 25, 289 - 392.

[9] Laver, J. (1980). The phonetic description of voice quality. Cambridge University Press, London.

[10] Lecluse, F. L. E. (1977). Elektroglottografie. Doctoral Dissertation, University of Rotterdam.

[11] Ling, D. L. (1976). Speech and the Hearing-Impaired Child: Theory and Practice. Washington, D.C.: The Alexander Graham Bell Association for the Deaf, Inc.

[12] Maassen, B. & Povel, D.J. (1985). The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. J.Acoust.Soc.Am., 78, 877 - 886.

[13] Markides, A. (1970). The speech of deaf and partially-hearing children with special reference to factors affecting intelligibility. British Journal of Disorders of Communication, 5, 126 - 140.

[14] Metz, D. E., Whitehead, R. L. & Whitehead, B. H. (1984). Mechanics of vocal fold vibration and laryngeal articulatory gestures produced by hearing-impaired speakers. Journal of Speech and Hearing Research, 27, 62 - 69.

[15] Povel, D.J. & Maassen, B. (1987). Visual information and speech acquisition of the deaf. (This volume).

THE ACQUISITION OF FINAL CONSONANTS

MARILYN MAY VIHMAN AND CHARLES A. FERGUSON

STANFORD UNIVERSITY

The open CV syllable is the basic, 'unmarked' syllable type in the world's languages and in the phonological development of children. This paper charts the course of acquisition of final consonants by children acquiring a language rich in such consonants and proposes four major characteristics: (1) The number of different consonant phone types in final position is equal to or less than the number in initial position; (2) fricatives and liquids are more likely than stops and nasals to be acquired first in final position; (3) final velars are more likely to be attempted than non-final velars or final non-velars; and (4) final voiced consonants pose a special problem for children, and some children may make use of nasals in attempting to produce them. These characteristics are systematically related to the occurrence of final consonants in children's babbling, to the distribution of final consonants in the world's languages, and to strength hierarchies proposed for consonants.

INTRODUCTION

The open CV syllable is the basic, 'unmarked' syllable type in the world's languages and in the phonological development of children. Reflecting on the vocalizations of their 13-month-old subjects, Kent and Bauer [1] comment on the "primacy" of the CV syllable shape, which may be viewed as a "simplest form...or a kind of atom in the formulation of speech" (p. 527). Although many languages have syllable-final and word-final consonants and even consonant clusters, these final consonants are much less frequent than initial consonants (both types and tokens). Also, final consonants are of very low incidence in babbling, regardless of the language spoken around the child. The acquisition of final consonants can thus be expected to pose a phonological challenge for children, from either a linguistic-universal or a biological-developmental perspective.

The identification and explanation of constraints on types of consonants occurring in final position in the world's languages constitute a significant part of the total characterization of the phonological structure of human languages ('phonological universals'), and analysis of the phenomena of final consonant acquisition can contribute to this 'universal phonology' ([2]).

Word-initial position is typically the position of greatest consonantal diversity in phonological inventory, though medial position in some languages for some classes of segments may be greater; final position is typically the position of least diversity, though preconsonantal position may be even more limited. These constraints may be expressed in terms of strength hierarchies of optimal syllable-initial segment classes and their mirror image for syllable-final position: stops, fricatives, nasals, liquids, glides, vowels ([3], ch. 10). Such hierarchies are intended to show universal relations, but admit of some language-specific variation. Whatever perceptual, articulatory, cognitive processing, and social/conventional constraints account for those hierarchies may be expected to result also in developmental patterns of order of acquisition and types of substitution and assimilation. Thus it may be expected that final fricatives, nasals, and liquids will not only be more frequent than final stops and occur in languages without final stops, but will also be acquired earlier. The present paper explores the actual phenomenon of acquisition of final consonants in the light of this expectation. English is an especially appropriate language for the investigation since the incidence of initial, medial and final consonants in running text is virtually identical (36% initial vs. 32% each medial and final: [4]). We will restrict ourselves here to the analysis of word- (or vocalization-) final consonants, since syllable-final consonants which are not also word-final are extremely rare in children's early productions.

Final consonants in babbling

Several careful accounts of the phonetic characteristics of babbling have documented the relative rarity of final consonants [1, 5-8]. On the other hand, in an analysis of consonant frequency in the babbling and word production of 10 English-learning subjects, Vihman, Ferguson and Elbert [8] found the mean proportion of final consonants to increase gradually with growth in the children's use of words, ranging from a mean of 6% final consonants early on to 17% when 25 or more words could be identified.

Differences have also been reported in the incidence of different manner categories in initial vs. final position in babbling. Oller et al. [5] reported a 10 to 1 ratio of stops to fricatives and affricates in initial position and a 3 to 1 ratio of final fricatives to stops (based on tokens).

Similarly, deBoysson-Bardies et al. [6] reported a 9 to 1 ratio of initial stops to fricatives and affricates and an 8 to 1 ratio of final fricatives to stops in the babbling of their French subject.

In inventories of consonant types used in babble only a slightly higher proportion of fricative and liquid segments were found in final position (29%) as compared with initial position (22%), based on the true consonant categories of stop and nasal (non-continuant) and fricative and liquid (continuant): [8]. Overall, only 19% of all initial consonants were continuant, while 32% of all final consonants were continuant. As increasing numbers of final consonants began to be used in words, the slight initial bias toward continuants in final position was strengthened.

Recent work on the transition from babbling to speech has strongly demonstrated the continuity in phonetic tendencies across that transition [5, 9, 10]. Accepting Locke's assertion that the beginnings of phonological development antedate the child's first use of adult-based words [9], it is important to consider the process by which final consonants are incorporated into the system in the course of acquiring a language characterized by heavy use of final consonants.

Final consonants in early word use

In general, final consonants are rare in early words, as the finding of continuity from babbling to speech leads us to expect, and the range of occurring segments is correspondingly small. In her longitudinal study of the phonetic inventories of early words for 33 children Stoel-Gammon [11] found that the typical inventory of initial phones tended to be about twice as large as the typical inventory of final phones.

The total incidence of initial and final consonant segment types in words and babble reported in Vihman et al. [8] for two lexical points is given in Table 1. Only 10% of the inventory consonants occurred in final position. While the overall proportion of consonants occurring in words was smaller (40%) than the proportion occurring in babble, a somewhat higher proportion of all final consonants occurred in words (48%). Some growth of consonant use as the children "enter into" English is apparent in the breakdown by lexical stages: At the earliest stage of word use final consonants accounted for only 9% of all consonant segments used, while at the later stage analyzed they accounted for 11%. There are no data available at present comparing consonant incidence in babble and words for other languages. However, the emergent influence of an adult language rich in final consonants appears to underlie these tendencies.

Focus on word-final consonants

Recent work in child phonology has emphasized the individual differences among children learning the same language (e.g., [12]). Differential attention to consonants in final position is one such individual characteristic. Menn [13] described her son Daniel's early phonological strategy as a decision "to disregard almost all information about the initial segments of a stop-final monosyllable" (p.226).Daniel seemed to select his earliest words so as to avoid those with contrasting initial and final consonants; after the first 30 words, he attempted many more words with such a contrast but applied regressive consonant harmony, adapting the initial consonant to the place of articulation of the second. A very similar pattern of development is described for one of three children in Stoel-Gammon and Cooper [14].

Table 1. Incidence of initial vs. final consonant types in babbling and words (based on [4], Table 5).

Initial consonants

stage[1]	babble	words
4-word	123	59
15-word	94	79
Total	217	138

Final consonants

stage[1]	babble	words
4-word	11	7
15-word	10	12
Total	21	19

[1] "Stage" = 4-word point (4+ words used in one session: 10 subjects) and 15-word point (15+ words: 7 subjects). The figures represent the sum of different consonants used 4 or more times by any child in any one of three weekly half-hour sessions.

Vihman and Hochberg [15] found that of 550 early words used by 7 children, a mean of 25% were sometimes produced with a final consonant. Only two children exceeded the mean. An analysis of the early phonological patterning of one of those children, Molly, is presented in Velleman and Vihman [16], supported by acoustic data. At 12 months Molly began to produce a number of obstruent-final words with heavily aspirated final stops or even affricates (e.g., oops, up, hot, book, peek, teeth). In the following month she began to produce nasal-final words as well, developing an idiosyncratic pattern in which the final nasal of the adult form was lengthened and followed by [i] or [ə]: bang [bæŋ :i], down [t'æ n:ə]. This pattern appeared to represent a phonetic rapprochement between the obstruent-final words, with their heavy aspiration, and the nasal-final words. Both word patterns subsequently proved highly productive, even attracting new words with nasals or affricates in other positions (Nicky [ɛn:i]; cheese [(a)Itʃ]). Like the children described in [13] and [14], Molly focused her attention on final consonants, developed a workable production strategy or "word recipe" and then used the patterns arrived at to add large numbers of new words to her lexicon. At present it is not possible to estimate the proportion of normally developing children who focus on final position, but it is probably not large.

CHARACTERISTICS OF FINAL CONSONANTS

Continuants and final position

Ferguson [17] suggested that "production of fricatives is easiest to acquire in post-vocalic, final position or intervocalically, and may precede the acquisition of stops in these positions" (p.661). We have seen that there is some association of continuancy with final position in babble. In an exhaustive longitudinal study of fricative acquisition by 6 subjects (aged 1;5 to 2;3 at the outset) Edwards [18] found that, as in earlier studies, the fricatives were generally acquired relatively late, after stops and nasals. Most of her subjects tended to produce fricatives correctly most often in final position (especially the interdentals, voiceless sibilants, and /v/), though there was considerable individual variation.

Similarly, Stoel-Gammon [11] noted that the inventories of her 15- to 21-month old subjects typically included stops, nasals and glides only, with fricatives and liquids appearing only in the 24-month inventories. Comparing initial and final phones within each manner class, Stoel-Gammon found that presence of a final stop or nasal in an inventory implied the presence of an initial stop or nasal. Fricatives and affricates showed great individual variation. Nine subjects had inventories with initial fricatives preceding final ones, while 7 subjects had inventories with final fricatives preceding initial ones. However, liquids showed a clear-cut association with final position. Of 25 subjects whose inventories contained liquids, only 5 had a liquid in initial position before they had one in final position.

In summary, the evidence (from English data) suggests that liquids are likely to be acquired first in final position, that stops and nasals are likely to be acquired first in initial position, and that fricatives may be too variable for a definite statement.

Velars and final position

Velar obstruents tend to be acquired later than labials and dentals by most children. A few children make relatively high use of velars in their early words, however, and these same children may favor final position. Ingram [19] hypothesized that consonants appearing early in a (child's) word are likely to be anterior, while consonants occurring later in the word will be back. Vihman and Hochberg [15] examined this hypothesis on the basis of data from 7 children. They found that one child used a high proportion of both velar and consonant-final words, but there was no overall correlation between velar and consonant-final word use. Considering stops and nasals only, the children as a group were found to favor initial position less and final position more for velars than for labials and alveolars, though in general the child bias in favor of initial consonants was very strong. Lastly, the children were found to attempt more word-final velars than labials or alveolars, and also more velars in medial and final position than in initial position. However, fully 73% of the adult word-final velars targetted were either produced in non-final position (e.g., dog[gɒ:]) or were spread to non-final position as well by consonant harmony (e.g., book[kuk]). Word-final labials and alveolars were less often subject to these changes. Vihman and Hochberg concluded that "while children are attracted perceptually to words with velars in final position, they show no particular preference for producing velars word-finally" (p.46).

Stoel-Gammon [11] found that while the presence of labials or alveolars in an inventory of final phones implies their presence in initial position, in 7 out of 31 cases (25%) velars were present only in final position. As in the case of fricatives among manner categories, velars were found to involve the most individual differences among place categories.

Final voiced steps

The acquisition of voicing appears to present problems for children in general [20, 21]. Some unusual production strategies have been identified for voiced stops in final position. Clark and Bowerman [22] noted that a typical progression in the acquisition of final consonants is (1) omission, (2) production of voiceless stops and nasals, and only later (3) production of voiced stops. Voiced stops may be devoiced in early production attempts, sometimes with distinctive lengthening of the preceding vowel. Clark and Bowerman documented for two children a stage intermediate between (2) and (3), in which final voiced stops were systematically replaced by the homorganic nasals, sometimes followed by the corresponding voiceless stop: rug [rʌŋk], bib [bIm] (Damon, aged 1;8-1;10); egg [æŋk], seed [din(t)] (Eva, aged 1;5-1;8). Both children had mastered the production of nasals in both initial and final position and at all three places of articulation before making use of this strategy. It is perhaps worth noting that both children seem to have first applied this strategy to velar-final words, Damon so producing only velar-finals for the first three weeks that the strategy was recorded.

Fey and Gandour [23] reported that their two-year-old English-speaking subject Lasan distinguished between voiced and voiceless obstruents only in the case of final stops. Final voiceless stops were consistently produced with an aspirated release, while final voiced stops were regularly produced with a nasal release: bad [bædn̥], pig and big [bIgŋ]. Fey and Gandour note further that the only noncontinuants to occur finally were nasals, and that the contrasts between stops and fricatives and between alveolars and velars were first made word-finally. Thus Lasan provides another example of a child who chose to focus on word-final position as he expanded his system of contrasts.

It is striking that nasals or nasal release should be used as part of a strategy for producing final voiced stops. This lends further support to the idea of a natural hierarchy of segment classes in a given syllabic position. That is, nasals may be more "natural" in final position than stops, though less so than the continuant consonants.

SUMMARY AND CONCLUSIONS

Study of the acquisition of word-final consonants in English yields the following generalizations.

(1) Word-final consonants are acquired later than initial consonants. At any point in development, the number of different consonant phone types in final position is equal to or less than the number in initial position. However, a few children utilize a strategy of making final position more salient than initial for consonant variety and stability.

(2) Continuants are more likely than non-continuants to be acquired first in final position. Of the continuants, liquids are most likely to be acquired first in final position; fricatives are more variable.

(3) Velar consonants have a special affinity for final position. Final velars are more likely to be attempted than non-final velars or final non-velars.

(4) Final voiced consonants pose a special problem for children, and some children adopt unusual strategies for producing them (e.g., nasal and stop clusters, nasal offglides, vowel lengthening.

These characteristics are sytematically related to the occurrences of final consonants in children's babbling, to the distibution of final consonants in the world's languages, and to strength hierarchies proposed for consonants. This systematic relationship is the essence of Jakobson's influential model of phonological development [24,25]. The child language data give further specification to the relationship and also in effect extend the Jakobson model to pre-speech, where he denied its relevance, and to final position, which he did not consider. The evidence for final consonants also strongly suggests that where there is relative infrequency and variability in phonological systems world-wide we may expect to find corresponding patterns of individual variation among children acquiring a particular language.

REFERENCES

[1] Kent, R.D. & H.R. Bauer, "Vocalizations of one-year-olds." Journal of Child Language, 12 (1985), 491-526.

[2] Perts, D.L. & Bever, T.G. "Sensitivity to phonological universals in children and adolescents." Working Papers on Language Universals, 13 (1973), 69-90.

[3] Hooper, J.B. An Introduction to Natural Generative Phonology. NY: Academic Press, 1976.

[4] Mines, M.A., Hanson, B.F., & Shoup, J.E. "Frequency of occurrence of phonemes in conversational English." Language and Speech, 21 (1978), 221-241.

[5] Oller, D.K., Wieman, L.A., Doyle, W.J. & Ross, C. "Infant babbling and speech." Journal of Child Language, 3 (1976), 1-11.

[6] De Boysson-Bardies, B., Sagart, L. & Bacri, N., "Phonetic analysis of late babbling: a case study of a French child." Journal of Child Language, 8, (1981), 511-524.

[7] Oller, D.K. & Eilers, R.E. "Similarity of babbling in Spanish- and English-learning babies." Journal of Child Language, 9 (1982), 565-578.

[8] Vihman, M.M., C.A. Ferguson, & M. Elbert. "Phonological development from babbling to speech: Common tendencies and individual differences." Applied Psycholinguistics, 7 (1986), 3-40.

[9] Locke, J.L. Phonological Acquisition and Change. NY: Academic Press, 1983.

[10] Vihman, M.M., Macken, M.A., Miller, R., Simmons, H. & Miller, J. "From babbling to speech: A re-assessment of the continuity issue." Language, 61 (1985), 397-445.

[11] Stoel-Gammon, C. "Phonetic inventories, 15-24 months: A longitudinal study." Journal of Speech and Hearing Research, 28 (1985), 505-512.

[12] Leonard, L.B., Newhoff, M. & Mesalam, L. "Individual differences in early child phonology." Applied Psycholinguistics, 1 (1980), 7-30.

[13] Menn, L. "Phonotactic rules in beginning speech: A study in the development of English discourse." Lingua, 26 (1971), 225-251.

[14] Stoel-Gammon, C. & Cooper, J.A. "Patterns of early lexical and phonological development." Journal of Child Language, 11 (1984), 247-271.

[15] Vihman, M.M. & J.G. Hochberg. "Velars and final consonants in early words." In J.A. Fishman et al. (Eds.), The Fergusonian Impact, I: From Phonology to Society. Amsterdam: Mouton de Gruyter, 1986.

[16] Velleman, S. & Vihman, M.M. "Phonological reorganization: A case study." In preparation.

[17] Ferguson, C.A. "Fricatives in child language acquisition." In V. Honsa & M. H. Hardman-Bautista (Eds.), Papers on Linguistics and Child Language. The Hague: Mouton, 1978.

[18] Edwards, M.L. Patterns and processes in fricative acquisition: Longitudinal evidence from six English-learning children. Unpublished PhD thesis. Stanford University, 1979.

[19] Ingram, D. "Fronting in child phonology." Journal of Child Language, 1 (1974), 233-241.

[20] Smith, B.L. "A phonetic analysis of consonantal devoicing in children's speech." Journal of Child Language, 6 (1979), 19-28.

[21] Macken, M.A. "Aspects of the acquisition of stop systems: A cross-linguistic perspective." In G.H. Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.), Child Phonology, vol. 1. NY: Academic Press, 1980.

[22] Clark, E.V. & Bowerman, M. "On the acquisition of final voiced stops." In J.A. Fishman et al. (Eds.), The Fergusonian Impact, I: From Phonology to Society. Amsterdam: Mouton de Gruyter, 1986.

[23] Fey, M.E. & Gandour, J. "Rule discovery in phonological acquisition." Journal of Child Language, 9 (1982), 71-81.

[24] Jakobson, R. Child Language, Aphasia and Phonological Universals. A.R. Keiler (Tr.). The Hague: Mouton, 1968. (Original title, Kindersprache, Aphasie und allgemeine Lautgesetze. Uppsala: Almqvist & Wiksell, 1941).

[25] Hawkins, J.A. "Implicational universals as predictors of language acquisition." To appear in Linguistics, 25.

ON THE TONOSYNTAX OF A HUNGARIAN CHILD'S EARLY QUESTIONS

ILONA KASSAI

Institute of Linguistics
Hungarian Academy of Sciences
Budapest, Hungary

ABSTRACT

The paper reports on the process of question acquisition from the perspective of prosody. The analysis of prosodic errors observed between 1 and 3 years reveals that the child has not acquired yet certain syntactic structures.

INTRODUCTION

Questions are an important means of cognitive development. Therefore the evolution of the verbal means of questioning highlights the intellectual development of the child on the one hand and its linguistic, especially syntactic development on the other.
In the present paper I give an account on a tentative analysis of questions gathered from the spontaneous speech of one child (a girl) produced in interaction with adults and regularly recorded from 1 to 3 years of age. I was particularly interested in the acquisition of the prosodic shape of questions, a topic largely neglected in child language research across the world.

THE SYSTEM TO BE ACQUIRED

In the process of language acquisition Hungarian children are faced with the following basic question types differing in form.

Wh questions. They require a question-word and a specific word order characteristic of emphatic sentences in which the emphasized element (here the question-word) is obligatorily followed by the unstressed verb. The remaining constituents can either follow the unit formed by the focus and the verb as part of the comment or precede it and constitute the topic of the sentence. In case the question-word stands for the predicate it can even be the last element of the sentence. If, in the neutral sentence, the predicate contains at its head some modifier, this latter must be postponed to the verb in the emphatic sentence [1],[2]. As in the case of wh questions the type of utterance is signalled both morphologically and syntactically, prosodically they are not autonomous in the sense that they do not have a specific intonation. They show the same falling contour as statements, with, however, a somewhat wider frequency range. This is a fourth or a fifth while that of statements is a third. This slight Fo-difference seems to contribute to the reeognition of questions [3]. As for stress patterns, this question type is usually realized with a single heavy stress located on the question-word. (In the Hungarian language word stess affects the first syllable.)

Yes/no questions. This question type has two varieties. The one is constructed by an interrogative particle added to the verb or the nominal predicate. Due to morphological marking the prosodic shape does not differ essentially from that of emphatic statements. Moreover, in present-day Hungarian this variety occurs rarely as main clause. Its use is more and more restricted to subordinate clauses. The other, almost exclusively used variety is expressed by means of intonation. The basic form from which all the remaining forms can be derived seems to be the rise-fall movement appearing on the last three syllables in questions containing only one trisyllabic or multisyllabic word. The magnitude of the rise is about a musical third while that of the fall is a fourth (Fig. 1a,b).

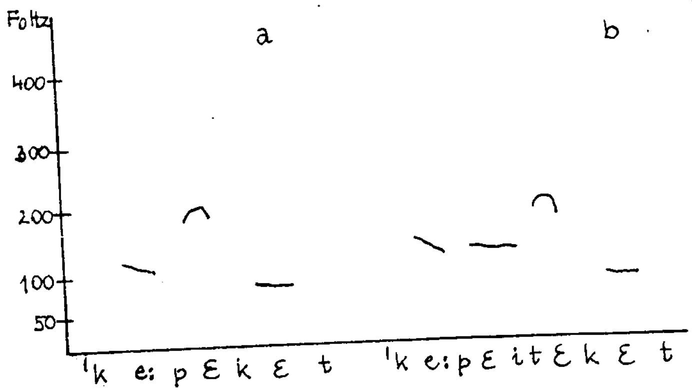


Fig. 1

In questions containing a bisyllabic word both the rise and the fall take place on the last syllable (Fig. 2). Finally, in monosyllabic questions only the rising part of the pattern is realized (Fig. 3).

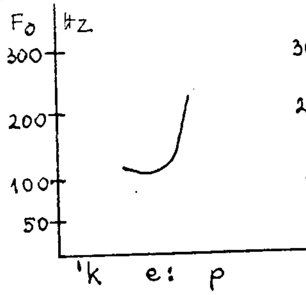


Fig. 2

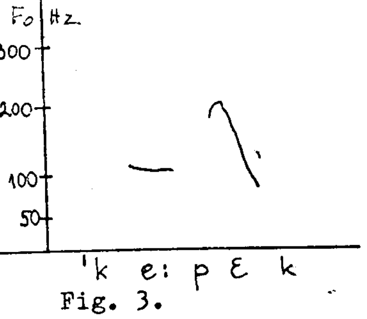


Fig. 3.

This fairly simple picture becomes more complicated in case the question contains more than one word and more than three syllables. The intonation of such question is determined by the number of syllables of the last stress group regardless of the number of words it contains. The rule is as follows. If the last stress group is monosyllabic it displays the contour of monosyllabic questions. If the last stress group is bisyllabic, it shows the pattern characteristic of bisyllabic questions. And, lastly, if in the last stress group there are three or more syllables, the intonation pattern is that of the corresponding one-word question. In other words, the prosodic shape of a morphologically unmarked yes/no question consisting of more than one word and more than three syllables strongly correlates with its topic-comment structure. For the word order the following holds. The constituent bearing the main stress is usually the first element of the sentence but it can also be located in sentence-medial or sentence-final position. If the focussed element is other than the verb, wherever is stands in the sentence, it must be followed, as a rule, be the verb.

Tag questions. They are constructed from a statement and the interrogative morhpeme ugye 'isn't it' added to either the beginning or the end of the statement. In the first case the intonation can only be falling, while in the second there is a choice for the interrogative morpheme to be realized as a statement or a bisyllabic question. The phonetic difference conveys an attitudinal one: the falling contour means that the speaker expects a positive answer. The rise-fall pattern refers to the speaker's desire to elicit a positive answer.

Elliptic questions. This question type shows a rising contour. Syntactically it is elliptic, i.e. only one part of the intended content is expressed. The elliped part can be completed from the nonverbal context as either a wh question or a yes/no question.

THE PROCESS OF QUESTION ACQUISITION

Within the recorded material I analysed questions in order to find out how the interrogative system of the adult language emerges and evolves in the child. The examination has revealed first of all that the productive use of questions is preceded by the imitation of adult models. Imitation has two forms. Part of adult questions was rehearsed by using the prosodic component only: the child hummed the intonation of the question. The other and greater part of imitative questions had an accurate intonation but only approximate segments. The first form of imitation decreases as the child's phonological competence progresses but the second form remains for long enough and occurs whenever the child does not understand or cannot answer the question addressed to her. The main function of imitation seems to be the learning of the verbal means of questioning. Besides, imitative questions may serve to maintain contact with the adult, thus they can perform a discourse function.

As for the order of emergence of the question type treated above the following may be assumed. First appear yes/no questions expressed by means of intonation (1;6,22). Wh questions come next in the developmental order (1;9,18). Tag questions do not appear before 2;4. Elliptic questions come last of all, at 2;7,20.

The yes/no question constructed with particle, as expected on the basis ofits adult use, did not appear as main clause in the periode examined.

The analysis of the formal aspect of question acquisition has revealed the following prosodic and syntactic tendencies. Prosodically wh questions do not cause any problem to the child as their intonation is almost identical with that of statements already acquired. However, from 2;2 in certain utterances one can hear an extra stress on the last syllable, which is in contrast to adult realizations but very characteristic of Hungarian children's performance. At closer examination it turns out that extra stress occurs mainly in longer, multiword utterances beginning with the stressed question-word. Another characteristic of the child's Wh questions is the topicalization of unstressed constituents which results in shifting the question-word towards the end of the sentence. These different strategies have the same goal: to give the end of the sentence perceptual prominence. The explanation for it might be the child's desire to provoke an answer or to get the partner's attention by all means. From among word order changes required by this question type verb-object ordering takes place at once but postposition of the verbal modifiers shows inconsistencies and does not stabilize until the end of the periode examined. The observed difference in the application of the inversion rule concerning subject and verbal modifiers may be given a multiple cue explanation. One seems to be handled by the structure of the Hungarian language which, according to recent research on syntax, is identified as a "topic prominent" language [1]. Another explanation might be the generally observed fact that children between 2 and 3 years, independently of the word order rules of their mother tongue, are inclined to place the verb at the beginning of the sentence. Accordingly, all the child has to do is to add a question-word to statements.

Yes/no questions expressed by intonation, though they appear first, are found to cause the child more difficulties than any other type. As demonstrated above, this question type shows three distinct intonation patterns according to the number of syllables contained in the word constituting the question by itself. This basic distributional rule seems to be acquired early and accurately. Nevertheless when the question contains more than one word its intonation patterning becomes dependent upon the location of the emphatic stress which, in turn, is dependent on the topic-comment articulation of the question. In multiword questions one can often detect intonational mistakes: the child uses a pattern contradictory to the topic-comment structure signalled by one or several of the following factors: stess assignment, word order, nonverbal context. The analysis of prosodically mistaken questions has shown some regularities in the seemingly chaotic patterning. There are utterances in which the child uses the pattern required by the number of syllables of the last word independently of its stressed or unstressed nature. In a few examples the intonation mistake can be considered as the consequence of misplaced stress. A part of the mistakes is supposed to be triggered by the non-application of the obligatory word order of emphatic sentences. On the other hand, it often happens that the child corrects herself within the same discourse turn and produces the appropriate prosodic solution. For tag questions the source of trouble is the sentence-initial position of the question morpheme prescribing a falling contour contradictory to the semantic content of the question morpheme. Elliptic questions are produced correctly.

DISCUSSION

The findings strongly suggest the conclusion that children below 3 are in the process of learning the complex rule-system governing the prosodic articulation and the topic-comment articulation. However two facts allow some other explanation too. Self corrections and the marked tendency in all erroneus items to shift the stress or the intonation peak to the last syllable make one think of an unconscious endevour to ensure continuity in discourse.

(For a more detailed version of this paper see Hungarian Papers in Phonetics, vol. 17)

REFERENCES

[1] K.É. Kiss, Topic and focus: The operators of the Hungarian sentence. Folia Linguistica 15, 1981, 305-330.

[2] F. Kiefer, On emphasis and word order in Hungarian. Bloomington, 1967.

[3] I. Fónagy, K. Magdics, A magyar beszéd dallama [The melody of speech in Hungarian], Budapest, 1967.

ALTERSMÄSSIGE BESONDERHEITEN BEI DER ANEIGNUNG DES QUANTITÄTSSYSTEMS DER ESTNISCHEN SPRACHE

LAINE VESKER

Lehrstuhl für Pädagogik
Republikanisches Institut für Lehrerweiterbildung
Tallinn, Estn. SSR, UdSSR, 200105

ANNOTATION

Die Quantität gehört zu den ersten Komponenten der Wortstruktur, die sich das Kind aneignet.Der Aneignungsperiode ist eine Fluktuation zwischen allen Quantitätsstufen charakteristisch. Analoge Abweichungen bei den Kindern mit allgemeiner sprachlicher Unterentwicklung sind häufig, schwer überwindbar und gehören zu den Merkmalen der Sprachstörung.

EINFÜHRUNG

Unter den theoretischen und angewandten Zielen der Kindersprachenforschungen interessiert uns die Rolle der alters – mäßigen Besonderheiten bei der Diagnose und Überwindung der Sprachstörungen: die Abweichungen in der Sprache können als Merkmale der Sprachstörungen nur im Vergleich zur Norm festgestellt werden: die Berücksichtigung der sprachlichen Ontogenese ist eines der Grundprinzipien der Logopädie /7/.

Bei der Untersuchung der Kinder mit der ausgeprägten sprachlichen Unterentwicklung konnte man neben den Störungen des Silben- und Lautstruktur auch Abweichungen in der Quantitätsstruktur der Wörter feststellen, über die in der Fachliteratur keine angaben zu finden waren. Auch in der logopädischen Praxis waren bis dahin diese Fehler nicht behandelt worden. Von besonderer Bedeutung dabei ist die Tatsache, daß die Quantität ein Universalmerkmal aller estnischen Wörter ist und in der Sprache eine phonologische Funktion ausübt.

Man unterscheidet drei Dauerstufen der Segmentalphoneme, die die Wortbedeutung und die grammatischen Formen differenzieren. Die Phoneme mit verschiedener Quantität können im Wort auf verschiedene Weise kombiniert werden: [vilì] (das Getreide), [vil̆lì] (die Blase,Gen.Sing.), [vìlli] (die Blase,Akk.Sing.), [vīlì] (die Feile,Gen.), [vîli] (Akk.); [sa̋k̀] (die Ernte), [sâG] (die Säge), [sak̀] (die Zacke); [valèD] (die Lüge,Nom.Pl.), [valet̀] (die Lüge,Akk.Sing.).

Phonetisch unterscheiden sich die Laute der 1. und der 2.Quantität voreinander durch Dauer, der Kontrast zwischen der 2. und der 3.Quantität ergibt sich vielmehr aus der gespannten Koartikulationsverbindung./ 1/ Mit der Vergrößerung der Quantität in der ersten Silben, verkürzt sich die phonetische Dauer des Vokals in der zweiten Silben: folglich verbreiten sich die Quantitätsmerkmale auf das gesamte Wort (auf die 1...3-silbige Einheit)/3/,

Laut E. Oksaar eignen sich die Kinder das Quantitätssystem der Sprache sehr früh an – im Alter von 26...27 Monaten, früher als das ganze Lautsystem. Seit dem

Altervon 28 Monaten sind keine Abweichungen mehr zu finden. Bis zu diesem Zeitpunkt betraffen die Abweichungen, die es gibt, nur die 2. und 3. Quantitätsstufe und kommen nie zwischen den beiden anderen Stufen vor.

E. Oksaar erklärt den Früherwerb der Quantität durch deren wichtige Rolle in phonologischen System der Sprache und die Bedeutung für die Kommunikation./4 /

In unserer Arbeit wollen wir die Untersuchungsergebnisse über 328 Kinder im Alter von 1,5 (17 Monaten) bis 7 Jahren darlegen. Bei der Charakterisierung der Aussprache der Kinder von 17 Monaten bis 2 Jahren verwènden wir die Notizen aus den von den Müttern geführten Sprachtagebüchern. Bei 325 Kindern im Alter von 2...7 Jahren untersuchten wir die Aussprache von 160 Einzelwörtern mit verschiedenen Quantitätsstrukturen.Die Kinder sollten das Spielzeug oder die auf den Bildern dargestellten Gegenstände nennen.

ALTERSMÄSSIGE BESONDERHEITEN BEI DEN ANEIGNUNG DER QUANTITÄTSSTRUKTUR: DER WÖRTER

Bei zwei der drei Kinder, deren Sprache die Mütter (Sonderpädagogen) aufgeschrieben hatten, waren nur vereinzelte Fälle inkorrekten Gebrauchs der Quantität zu finden. Beim dritten Kind (Evelin) war der Prozeß der Aneignung der Quantität besser zu sehen. Die Mutter hat regelmäßig Notizen gemacht. Wir haben dort 37 Wörter mit verschiedenen Quantitätsersetzungen gefunden: entweder wurde das Wort mit einer nichtadäquaten Quantität ausgesprochen, oder es traten nur Verwechslungen der Quantität der Laute in den ersten Silben auf, und die Quantität des Wortes blieb unverändert. Man konnte verschiedene Varianten der Verwechslung der Quantität finden: [tul̆lè] pro [tulè], [kal̆là] pro [kalà], [kik̆kù] pro [tiGù], [ak̆ke]pro [ak̆kèn], [am̆mà] pro [rāmàt], [nöp̄] pro [nö̂p].
Es überwog die Vergrößerung der Quantität und sie trat oft bei den Wörtern der 1. Quantität, seltener den Wörtern der 2. Quantität auf (Q1→Q2 - 21 Fälle,Q2→Q3 - 5 Fälle). In allen Wörtern vergrößerte sich die Quantität der Konsonanten.

Es ist möglich, die stufenweise Aneignung der Quantitätsstruktur der Wörter zu beobachten. Häufig verwendete das Kind dabei in kurzer Zeit die falschen und die richtigen Varianten nebeneinander.

Am 6. März, im Alter von 17 Monaten sprach Evelin zweimal das Wort tita Q2 (die Puppe oder das Kleinkind) richtig und einmal in der 3.Quantität: [tit̆ta]. Weiter: am 13.März -[tiDà] Q1; am 26.März - [tīDà] Q2; am 7.April - [tiDà] Q1; am 8. April -[tit̆tats] Q3; am 3. und dem 10.Mai - [tit̆tà] Q2 (richtig!); am 10.Sept.- zweimal richtig, einmal mit der Q3: [tit̆tà naĕrà], [tit̆ta] laŭlà], [tit̆ta kal̆li].

Am 10.Mai -[maǹà](vanaema - die Großmutter); am 26.Mai -[mañ̆nà]Q2; am 30.Mai -[mañ̆nà]Q2; am 16.juuni -[vanà].

Am 16.Mai. Evelin: ai-ai [put̆tu]- die Mutter: ei [pût̆tu] - Evelin: [äl'à put̆tu].

Die letzten Notizen haben wir vom 10. Oktober im Alter von 23 Monaten. An diesem Tag hat die Mutter 59 Phrasen aufgeschrieben (ingesamt 100 Wörter). Wir fanden folgende Ersetzungen der Quantität: [is̆sì tul̆lì] pro [iZà tulì], [is̆si kap̆pà] pro [iZà maGàB], [sok̆ki pro sok̆kìD], [tit̆ta kal̆li] pro [tit̆tà kal̆lì], [ak̆ke pro ak̆kèn], [tak̆kù] pro [taskù], [nöp̄] pro [nö̂p].

Alle Wörter in der Sprache des Kindes waren ein- und zweisilbige, man konnte auch die Vereinfachung der Wortstruktur feststellen: Assimilation und Auslassen der Laute; Palatalisierung, statt r der Laut l.

Es konnte auch das Auslassen der Silben beobachtet werden, in einigen Wörtern waren beide Silben gleichbetont.Wahrscheinlich gehört Evelin zu dieser Gruppe

der Kinder, denen beim Spracherwerb die Reduktion der Silben charakteristisch ist.

Bei den Kindern im Alter von 2 Jahren konnten wir bei 15 die Verwechslung der Quantität feststellen im Durchschnitt 1...8 Fälle pro Kind. Sie traten sowohl in den ersten als auch in den nichtersten Silben auf und waren in den Wörtern mit größerer Silbenzahl zu finden.

Tabelle 1

Zahl der Kinder mit Quantitätsersetzungen in jeder Altersgruppe

Alter	2	3	4	5	6	7
Gesamtzahl	25	18	50	60	70	102
Zahl der Kinder mit Quantitätsersetzungen	15	11	6	6	3	1

Weil bei den 2jährigen Kindern im Vergleich zu den anderen Kindern die Quantitätsverwechslungen relativ zahlreich und verschiedenartig waren, bringen wir hier alle Ersetzungsarten mit Beispielen . (Q1→Q2 bedeutet Veränderung der Quantität des Wortes, Q2=Q2 - Verwechslung der Quantität der Laute in der ersten Silbe, die Quantität des Wortes bleibt unverändert; K – kurze Klusile in der nichtersten Silben; K̀ - lange Klusile in den nichtersten Silben).

1. Q1→Q2 [hop̆pune] pro [hoBùne]; [lap̆piDas] pro [laBìDas], [mom̆mì] pro [momàt], [tat̆tì] pro [täDì].

2. Q1→Q3 - [tat̆ti] pro [täDì].

3. Q2→Q1 - [kiZù] pro [kīZù], [ńiñuk̆keZeD] pro [liñ̆nuk̆keZeD], [suZat] pro [sūZàD]

4. Q2→Q3: [tüttur] pro [tŭDruk], [kîZu] pro [kīZù], [ìlveZ] pro [ĭlvèZ], [lamBaZ] pro [lam̆màZ].

5. Q3→Q2: [lammak̆ke] pro [lamBak̆ke].

6. Q2=Q2,Q3=Q3 -[kis̆sù] pro [kīZù], [pork̆kañ̆D] pro [porĞanD], [sap̆pà] pro [saBà], [hil̄] pro [hîr], [roŋk] pro [roŋG].

7. K̀→K -[kuk̆kèGe] pro [kuk̆kèk̆ke], [sēlìG] pro [sēlik].

8. K→K̀ - [nañk̆kut] pro [jäñk̆kùD] , [kapsat] pro [kapsaD], [suZat] pro [sūZàD].

Die Ersetzungen Q2→Q3 (ein gespanntenes Artikulieren der langen Laute) traten bei 2,1% aller Wörter der 2.Quantität auf. Von den Wörtern der 1.Quantität wurden 1,9% durch die Wörter der 2. und 3. Quantität ersetzt. Die anderen Abweichungen zeigten sich noch seltener.

Bei den 3jährigen Kindern war die relative Häufigkeit der Ersetzungen nicht niedriger (2,1%). Wir fixierten auch die zufälligen Ersetzungen K→K̀ und Q2→Q1, insgesamt 1...2 Fälle. Man kann sagen,daß die dreijährigen Kinder die Artikulation der Wörter der 1. und 2. Quantität vollständig erworben haben. Stabiler sind die Abweichungen beim Artikulieren der Wörter mit der 2.Quantität, aber sie kamen nicht bei allen Kindern vor.

Auch bei den 4- und 5jährigen Kindern zeigten sich einige Abweichungen: Q2→Q3 und K̀→K, 1...3 Fälle in der Aussprache eines jeden der 6 Kinder.

Die 6- und 7jährigen machten in der ersten Silben keine Fehler mehr, die zufälligen Abweichungen betrafen nur die langen Klusile im Wortauslaut.

Unsere Angaben zeugen davon, daß die Quantitätsstruktur der Wörter wircklich sehr früh erworben wird; die Häufigkeit der Quantitätsveränderungen bei den Kindern war gering, bei vielen Kindern konnten aber keine Abweichungen fixiert werden. Möglicherweise fiel bei ihnen die Aneignungsperiode in ein früheres Alter.

Im allgemeinen bestätigen unsere Angaben die Ergebnisse von E. Oksaar, darüber hinaus werden einige Einzelheiten präzisiert: die Abweichungen betreffen nicht nur die 2. und 3. Quantität, sondern können sehr verschieden sein, allerdings hat die Fluktuation zwischen der 2. und der 3. Quantität Übergewicht (es gab eine Ausnahme - Evelin im Alter von 1.5...1.11). Es tritt deutlich die Tendenz zutage,Wörter mit den größeren Quantität auszusprechen, dabei vergrößert sich meistens die Quantität der Konsonanten. Weil vor allem bei den Wörtern der 2. Quantität Schwierigkeiten auftraten, können wir diese Wortstruktur als die "kritische" bezeichnen. Obwohl die Quantitätsstruktur im allgemeinen sehr früh erworben wird, gibt es Kinder, die auch noch später d. h. im 3. und 4. Lebensjahr Abweichungen haben.

Die Aneignung der Quantitätsstruktur der Wörter fällt zeitlich mit der Aneignung der Rhythmus- und Silbenstruktur zusammen. Aus den Untersuchungsergebnissen geht hervor, daß der Ton zu den ersten Komponenten der Wortstruktur gehört, die sich das Kind aneignet / 6 /. Die seltenen Abweichnungen von der Norm kommen nur im 2. Lebensjahr vor. Auch der Aufbau der Silbenstruktur fällt in das 2. Lebensjahr / 6 /. Die Quantität, die eng mit dem Ton und der Silbenstruktur verbunden ist,eignet sich das Kind zusammen mit allen diesen Strukturelementen des Wortes an. So erwirbt das Kind die Quantitätsstruktur tatsächlich bevor bei ihm alle Laute vorhanden sind. Die weitere Ergänzung und Vervollkommnung der Segmentalstruktur baut sich auf das erworbene Quantitätsschema auf.

ZUR ANEIGNUNG DER QUANTITÄTSSTRUKTUR DER WÖRTER BEI KINDERN MIT ALLGEMEINEN UNTERENTWICKLUNG DER SPRACHE

Allgemeine Unterentwicklung der Sprache wird als Sammelbegriff verwendet und kann bei den Kindern mit verschiedenen Sprachstörungen (Alalie, Aphasie, Dysartrie) in verschiedenen Ausprägunsgraden auftreten.

Beim Aussprechen der Einzelwörter traten dieselben Abweichungen zutage, die bei den Kindern mit der normalen Sprachentwicklung zu beobachten waren./5 / Der Unterschied bestand vor allem in der Häufigkeit dieser Fehler. Es wurde festgestellt, daß der Charakter und die Häufigkeit der Fehler neben den Stufen der sprachlichen Entwicklung auch von einigen linguistischen Faktoren abhängen. Zu den schwierigsten Strukturen gehören die zweisilbigen Wörter der 2.Quantität mit dem langen Klusil oder dem langen Vokal in der ersten Silbe, die Wörter der 1. Quantität mit kurzem Klusil in der nichtersten Silbe, die Wörter der 3.Quantität mit langem Klusil in der nichtersten Silbe, einige einsilbige Strukturen (paat, laud, saag). Es hat sich ergeben,daß die Abweichungen, die bei den normalentwickelten Kindern häufiger auftraten,bei den sprachgestörten Kindern zu den schwierigsten gehörten (Q2→Q3, K→Ǩ).

Bei der Formierung der Sprache bei sprachlosen Kindern muß man berücksichtigen, daß die Quantität eines der ersten Merkmale der Wortstruktur ist, das erworben wird. Deswegen ist es notwendig in ersten Linie die Hauptstrukturen (ein- und zweisilbige Einheiten) auf der Basis der vorhandenen Laute "aufzubauen".Dabei sind die Schwierigkeitsstufen der Wörter zu berücksichtigen.

LITERATURVERZEICHNIS

1. A.Eek,M.Remmel.Eesti keele foneetika uurimise tulemusi. - Keel ja Kirjandus 1971, nr. 12

2. A. Eek. Observations on the duration of some word structures: II. - Estonian Papers in Phonetics, Tallinn, 1975

3. Ķ.Karlep, L.Vesker. Mõningaid emakeele algõpetuse probleeme logopeedi pilguga. - Noukogude Kool 1972, nr. 5

4. E. Oksaar. Zum Spracherwerb des Kindes in zweisprachiger Umgebung. - Folia Linguistika IV 3/4 1970. The Hague

5. L. Vesķer. Häälıku ja sõnavälte asendustest kõne üldise alaarenguga lastel. - Emakeele Seltsi Aastaraamat 22, 1976

6. A.Гвоздев. Усвоение ребёнком звуковой стороны русского языка. Москва-Ленинград, 1948.

7. P. Левина. Наука о нарушениях речевого развития у детей. - Советская педагогика (I), Москва, I974.

THE ACQUISITION OF PALATALIZATION IN RUSSIAN

Katia McClain
Department of Slavic Languages and Literatures
University of California
Los Angeles, California 90024
USA

The paper treats the acquisition of palatalization for dental and labial stops in prevocalic environments in Russian using data from the 1927 longitudinal study of A.N. Gvozdev which describes the early stages of phonological acquisition by his son. The initial goal was a reanalysis of Gvozdev's data to provide a description of the phonemic as well as phonetic facts in the data. That is, not merely to describe the acquisition of individual sounds, but to describe the child's pre-adult phonological system(s).

To provide a general framework for the acquisition of palatalization by the child, Gvozdev's own explanations, as well as previous explanations in the early stages of phonological acquisition in Russian[7], as well as studies of palatalization in Slavic languages are examined.

Finally, it is shown that the facts and issues in the child's acquisition of palatalization can best be explained by showing which phonemic contrasts have been acquired and by relating the child's acquisition to specific phonetic properties and ambiguities, eg. formant frequencies of vowels, of the adult system.

INTRODUCTION

This paper treats the acquisition of palatalization for dental and labial stops in prevocalic environments in Russian. The data used is from the 1927 longitudinal study of A.N. Gvozdev[7], in which he describes the early stages of phonological acquisition by his son, Ženja.

I will argue that in order to best explain the facts and problems of the acquisition of palatalization, it is necessary to understand the child's pre-adult phonological system. That is, one must not only examine the phones in isolation, as Gvozdev did, but also the development of phonemic contrasts and syntagmatic contraints. Furthermore, the child's developing system must be examined within the context of the relevant facts, phonemic and phonetic, of the adult system.

Phonology of Russian

The Russian phonological system includes five vowel phonemes, the front vowels /i e/, the back vowels /a o u/, and a series of consonants which fall into classes according to place and manner of articulation. These consonants may utilize palatalization either contrastively or as an obligatory feature. This paper will examine the dental stops /t d n/ (and their palatalized counterparts /t' d' n'/) and the labial stops /p b m/ (and their palatalized counterparts /p' b' m'/). Palatalization functions distinctively for dental and labial stops in Russian before the phonemes /i a o u/ and in final position. Before the phoneme /e/ in native Russian words the phonemic distinction is neutralized. Only the palatalized variant of the consonant appears. There is therefore an asymetry in the distribution of phonemic palatalization before different vowel phonemes in Russian.

The effect of palatalization on vowel phonemes in Russian is very strong. Even though there are only five vowel phonemes, it is traditional to distinguish at least two phones for each phoneme conditioned by the presence or absence of palatalization of the surrounding (especially preceding) consonants.[4] [15]

DATA

The source for the data in this paper is the diary of Gvozdev, a Russian philologist who observed his son from the age of one year and seven months until eight years of age. I will be concerned with data relevant only to the acquisition of dentals and labials in prevocalic position, for the time period one year seven months (1,7) to two years four months (2,4). Behaviour of the segments in final position was not considered because final segments are often treated in a special way or omitted in the early stages of acquisition.[9] The data will be presented in phonetic transcription.

At the stage 1,7-1,9, sequences where a nonpalatalized labial should appear before a phonemic back vowel are produced correctly by the child [mas'a] ([máslə] 'butter',) [pat'] ([spat'] 'to sleep'). Where a palatalized labial should appear before the vowels /i/ and /e/, the child pronounces the sequences correctly [p'is'i] ([p'išɨ́] 'write!'). For nonpalatalized labial before phonemically front /i/ (phone [ɨ]), the child produces a palatalized labial and the front allophone ([m'is'ka] for [mɨ́škə] 'mouse' (dim.)) For palatalized labial before phonemic back vowel the child produces the palatalized labial and a front vowel [p'et'] for [p'æt' /p'at'/ 'five'. Palatalized dentals before phonemic back vowels are produced correctly [t'ot'a] ([t'ót'ə] 'aunt'). None of the sequences of nonpalatalized dental and phonemic back vowel are correct. The palatalized dental occurs instead [t'am] ([tam] 'there'). Palatalized dentals before /i/ or /e/ are correct: [d'i] ([id'í] 'go!'). Nonpalatalized dentals which should occur before /i/ are mispronounced by the child. [T'i] sequences appear instead of [T°ɨ] [bad'i] ([vʌdí] 'water' gen. sg.).

In the month 1,10-1,11 nonpalatalized labials before back vowels are all produced correctly by the child [s'abaka] ([sʌbákə] 'dog'). Examples of the palatalized labial before phonemic front vowel are produced correctly [kup'il'a] ([kup'ílə] 'she bought'). The child mispronounces the adult [P°i], producing [m'ija] for ([mɨlə] 'washed' neut. sg.). The word requiring a palatalized labial before a phonemic back vowel alternates between front and back vowels [p'ec']-[p'ac'] ([p'æt'] 'five'). Sequences of palatalized dentals before back vowel alternate palatalized and plain phones [s'en'a]-[s'ena] ([žɨ́n'ə] 'Ženja'). Most examples of nonpalatalized dentals before /a o u/ are pronounced correctly [noga] ([mnógə] 'many'). Some forms alternate hard and soft dentals [paduka]-[pad'uka] ([pʌdúškə] 'pillow'). Some have only the incorrect palatalized dental [d'und'uk] ([sundúk] 'box'). Palatalized dentals and front vowels are correct [n'is'ka] ([kn'íškə] 'book' dim. Adult [dɨm] 'smoke' is produced as [d'im].

At the stage 1,11-2,0 sequences of plain labial and back vowel are produced correctly. Palatalized labials before /i e/ are also correct. Adult [P°ɨ] are still incorrect [mam'i] ([mámɨ] 'mama'gen.sg.) Adult [p'æt'] 'five' occurs incorrectly as [p'ec']. Most of the palatalized dentals and back vowel sequences are now produced correctly [d'ot] ([id'ót] 'he/she goes'). Most sequences of plain dental and back vowel are correct. All cases of palatalized dental and /i e/ are correct. Adult [T°ɨ] is either incorrect [t'i] ([tɨ] 'you') or shows alternating forms [d'im]-[dɨm] 'smoke'.

In the following months (2,0-2,4) the child's system of palatalization moves towards the adult system.

At the first pre-adult stage (1,7-1,9), for the labials, palatalization is distributed according to the following vowel: [P'] before /i e/ and [P°] before /a o u/. In contrast to this distribution, dental stops occur as [T'] before all vowels.

At the second stage (1,0-1,11) the labials show no change. [P'] appears before /i e/ and [P°] before /a o u/. The system for the dentals has changed and looks like the system for the labials. [T'] occurs automatically before /i e/. [T°] occurs before /a o u/ in most cases.

At the third pre-adult stage (1,11-2,0) there is no change for the labials: [P'] before /i e/ and [P°] before /a o u/. The system for the dentals has changed again. [T'] is still mandatory before /i e/. However, now [T'] may now occur before /a o u/ as well as [T°]. The child has begun to distribute palatalization according to adult phonemic constraints instead of according to vowel context. Furthermore, the appearance of alternation in [d'im]-[dɨm] suggests that the dentals will soon adopt contrastive palatalization before /i/.

Gvozdev claims that by the end of the stage (1,7-1,9) both plain and sharp labials and sharp dentals have been acquired. The plain dentals are not acquired until 1,10.

The data at this stage shows that indeed, both hard and soft labial phones have appeared. However the labials appear in complementary distribution, [P'] before /i e/ and [P°] before /a o u/. Therefore, it cannot be said that the phonemic contrast of palatalization has been acquired for the labials, or that the contrastive adult phonemes /P'/ and /P/ are truly present in the child's system.

The same situation obtains for the dentals in the period 1,9-1,10. Any hard dentals which are produced appear before /a o u/. Only soft dentals appear before /i e/.

The need to distinguish the two levels in acquisition (phonetic and phonemic) has been recognized in earlier works. Menn notes that there is a difference between "the ability to hit a phonetic target accurately and the more "cognitive" acquisiof the information that the two phones contrast phonologically." [14]

An interesting fact arising from the data is that the marked palatalized dental phones appear earlier that their unmarked plain counterparts. Gvozdev indicates only that the child may be missing a particular articulatory function and therefore cannot pronounce the plain phones.

Jakobson[11] offers a possible explanation. Part of his theory of language acquisition is the principle of maximal contrast. According to this theory, the first sound a child acquires is an "a" type vowel. A labial is the first consonant acquired because it provides for a maximum contrast with that vowel. Because a labial is a grave consonant, one of the next consonants to be acquired will be a dental, providing the opposition grave/acute. The fact that dentals appear first as [+palatalized] is not a problem, and indeed is crucial to this theory: "...the initial inclination of children to palatalize dentals can also be accounted for. Dentals are opposed to the labials by their distinct lightness and since palatalization...intensifies the lightness of the consonant, the palatalized dental sound offers the optimal degree of lightness." Jakobson indicates that the early appearance of palatalized dentals has been noted not only for Russian, but also in French, Polish, Estonian and Japanese. Further work in cross-linguistic phonology will verify the accuracy of Jakobson's hypotheses.

Further Discussion of Dentals and Labials

There is a paradox in the acquisition of palatalization in the early stages. Although nondistinctive variation arises first in the labials, the distinctive opposition occurs first in the dentals. Interestingly, these facts correlate well with the facts of adult Russian and other Slavic languages in which palatalization occurs more for dentals.

In adult Russian, dentals show the use of distinctive palatalization more than labials. Data from Avanesov shows that in final position, soft labials are becoming hard while dentals are not[4], showing that in final position dentals show more contrast.

As mentioned above, all dentals and labials are palatalized before /e/ in native Russian words. The situation in foreign borrowings is different. Before /e/ in these lexemes a consonant may appear as [-pal.]. However, as noted by Holden[8], the tendency to appear as [-pal.] is not equally utilized by all consonants. Labials assimilate (return to their neutralized state) before /e/, while the dentals maintain the distinctive contrast. Thus Holden suggests that, "...the opposition of palatalization vs. nonpalatalization is most weakly developed for velars, more developed for the labials, and most developed for the dentals."

Vowel Context and Asymmetry

Earlier we showed that in the child's pre-adult systems the application of palatalization works differently in the environment of different vowels ([+pal] before /i e/, [+pal] before /a o u/).

If we examine adult Russian, such an asymmetry is not clear. Back vowels allow distinctive palatalization, but before the two front vowels palatalization works differently.

There are facts about other Slavic languages that do show the asymmetry. For example in contemporary standard Bulgarian distinctive palatalization is found only before back vowels. [3]

Another example of this asymmetrical application of palatalziation before the two types of vowels can be found in the history of Slavic, in the dispalatalization of Ukrainian. The development, as noted by Jakobson [10], was that all palatalized consonants were dispalatalized before the front vowels, while they retained their palatalization before /a o u/. Since hard consonants also existed before /a o u/, it became the environment which allowed more distinctive palatalization.

EXPLANATIONS FOR PALATALIZATION PATTERNS

Factors affecting acquisition noted in the literature include articulatory constaints, phonological processes, avoidance, asymmetries in the adult system and others. This section will explore two kinds of explanations for the child's acquisition of palatalization: ambiguities in the adult acoustic signal and contextual constraints which cause assimilation or allow contrast. This phonetic model seems to most accurately explain the acquisition of palatalization in this child.

Ambiguities in the Russian Vowels

In a model of sound change proposed by Andersen, ambiguities in the utterances of adults open the way to possible reanalysis by a new generation of speakers. In a system with two features, "the language learner, who has to interpret its acoustic manifestations [must] make a number of decisions... how many phonological oppositions are involved... which of the constituents is superordinate and which subordinate."[1]. Andersen notes that the child may make different choices then the adult, choices made plausible by ambiguities in what he hears.

As stated by Ladefoged, "vowels can be described as points on a continuum in a way that is not true for consonants..."[12]. A continuum that needs to be divided into meaningful units is inherently ambiguous. The question, then, is what kind of division of the continuum the child is going to make. In this case, can Ženja's division of the vowels into the groups /i e/ vs. /a o u/ be given a phonetic explanation?

Fant (1970) provides the basis for an articulatory classification which permits the separation of vowels into the two groups [i e] and [a o u]. Fant shows that the distance of the maximum constriction from the front of the vocal tract is one of the most important dimensions for the vowels. The distance may be seen in Table I (from [5]).

Table I: Distance of the Main Constriction from the Front End of the Vocal Tract (in cent.)

[i]	[e]	[ɨ]	[u]	[o]	[a]
4	4	7.5	11	12	13

It is clear from Table I that /i e/ can be grouped together, apart from the rest of the vowels.

Fant also utilizes the front to back cavity volume ratio which he shows separate [u o a] from the other vowels. [5]

Palatalization consists of a constriction in the palatal region, precisely where the vowels /i e/ have their maximum constriction.

Looking at the acoustic shape of Russian vowels, we see the following formant values (from [5]):

Table II: Formant Frequencies

	i	e	ɨ	u	o	a
Fl	240	440	300	700	535	300
f2	2250	1800	1480	1080	780	625

Here again, a division of vowels into the two groups /i e/ and /a o u/ is possible according to the height of the second formant.

A high second formant is the most important cue for palatalization. Vowels with a high second formant might well be expected to function in a special way with respect to palatalization.

The child's pattern of palatalization for the labial stops in all three stages is thus easily explained. He palatalized before /i e/, vowels that sound like and are articulated like palatal sounds, and does not palatalize before other vowels.

Dentals have a high second formant transition similar to the high second formant trajectory produced by palatalization, and that is easily confused with palatalization. For example, Andersen [2] has suggested that this kind of confusion has led speakers of certain Czech dialects to reinterpret palatalized labials as dentals. The child might, in a similar way, re-interpret all dentals, which have a high second formant characteristic of palatalization, as palatalized.

In the first stage, the child palatalizes all dentals. This is consistent with the high second formant transition of the dentals, and seems particularly likely re-interpretation given that he is hearing a language in which palatalized dentals occur. It seems that the palatalization of some dentals is overgeneralized to include all of them.

At the second stage the dentals have changed their distribution of palatalization. Before front vowels, both the consonant and the vowel have a front tongue constriction and a high second formant, forming a gesture and an acoustic shape similar to palatalization. In the environment before front vowels, the distinction between palatalized and nonpalatalized dentals is quite subtle acoustically.

Dentals before back vowels now appear as nonpalatalized. The child has thus begun producing dentals in two different ways, but does not use palatalization distinctively. The acoustic and articulatory characteristics of the following vowel, rather than the acoustic and articulatory characteristics of the consonant, now come to determine whether dentals are palatalized or not. The pattern for the dentals at the second stage is therefore the same as that for the labials.

In the third stage, there is no change for the labials or dentals before front vowels. Dentals before back vowels now may occur as palatalized or nonpalatalized. Palatalization causes a high second formant, while back vowels have a low second formant. Palatalization will therefore cause a steep downward glide of the second formant. In the absence of palatalization this very steep glide will not occur. Therefore, the distinction between palatalization and nonpalatalization should be highly audible before back vowels. It thus seems logical for the child to develop the contrast first for dentals before back vowels. The fact that back vowels allow more phonemic palatalization, and that dentals utilize phonemic palatalization to a greater degree than labials, is true also of adult Russian and other Slavic languages. The adult asymmetries and the child's acquisition patterns are both subject to the same phonetic constraints. (For further discussion of parallels in child and adult systems see [6].)

To return to our original question, given that the vowels are potentially ambiguous, what would lead Ženja Gvozdev to divide the vowel continuum into the groups front/back. Fant's analysis, utilizing maximum constriction and second formant height, shows that the Russian vowels really can be divided naturally into these groups. Therefore, it is not surprising that the child does so.

The substitutions made by the child become clear within Fant's framework. The child hears the adult sequence [Cɨ] and produces [C'i]. As indicated in tables I and II above, [ɨ] can be grouped with [i e] on the basis of both articulatory and acoustic factors (the point of maximum constriction and the height of the second formant). Furthermore, [ɨ] is and allophone of /i/ in the adult language. This apparently leads the child to reinterpret [ɨ] as [i].

The problem with palatalized labials before back vowels is more complex. As pointed out in literature on child language, children often deal with difficult combinations by avoiding them. [13] Ženja produces only one example of /P'/ before the back vowels [p'ec'] ([adult [p'æt'] from /p'at'/). He maintains the correct palatalization but fronts the vowel. Although Fant does not include [æ] in his tables, he does say that "the centralization of /u/ /o/ /a/ phonemes in positions between two sharp [+pal] consonants resulting in the allophones [ü] [ö] and [æ] is manifested by a higher F2."[5]. Since a raised F2 is a cue for front vowels and palatalization, it is not surprising that the child reinterprets the combination of a palatalized labial and the front allophone of a back vowel as palatalization plus a front vowel.

CONCLUSION

This paper has presented the facts of acquisition of palatalization for dental and labial stops in prevocalic environment for one Russian child. It showed that the general facts of acquisition can best be explained not only by showing which phones have been acquired, but by showing which phonemic contrasts and syntagmatic constraints are relevant to the child's system. The child's development of palatalization has been shown to be related to the articulatory and acoustic properties of the adult system.

REFERENCES

[1]Andersen, H. 1973. Abductive and deductive change. Language 49: 765-793.

[2]Andersen, H. 1978. Vocalic and consonantal languages. In H. Birnbaum, L. Durović, et al, (eds.), Studia Linguistica Issatschenko, 1-12.

[3]Aronson, H. 1968. Bulgarian inflectional morphophonology. The Hague.

[4]Avanesov, R. I. 1972. Russkoe literaturnoe proiznošenie. Moscow.

[5]Fant, G. 1970. Acoustic theory of speech production. The Hague.

[6]Greenlee, M. and J. Ohala. 1980. Phonetically motivated parallels between child phonology and historical sound change. Language Sciences, vol. 2, no. 2: 283-308.

[7]Gvozdev, A. N. 1961. Voprosy izučenija detskoj reči. Moscow. (Originally published as Usvoenie rebenkom rodnogo jazyka. In Detskaja Reč'. 1927, and Usvoenie rebenkom zvukovoj storony russkogo jazyka. Moscow. 1948.

[8]Holden, K. 1976. Assimilation rates of borrowing and phonological productivity. Language 52: 131-147.

[9]Ingram, D. 1979. Phonological patterns in the speech of young children. In P. Fletcher & M. Garman (eds), Acquisition: Studies in First Language Development. 133-148. Cambridge.

[10]Jakobson, R. 1962. Remarques sur l' évolution phonologique du russe comparée à celle des autres langues slaves Selected Writings II. 7-116. The Hague. (Originally published in Travaux de Cercle Linguistique de Prague, II. 1929.)

[11]Jakobson, R. 1968. Child language, aphasia, and phonological universals. The Hague. (Originally published as Kindersprache, aphasie und allegemeine Lautgesetze. Uppsala Universitets Arsskrift 1-83. 1941.

[12]Ladefoged, P. 1971. Preliminaries to linguistic phonetics. Chicago.

[13]Menn, L. 1978. Phonological units in beginning speech. In A. Bell & J Hooper (eds.), Syllables and segments. 157-171. Amsterdam.

[14]Menn, L. 1980. Phonological theory and child phonology. In G. Yeni-Komshiam, J. Kavanagh, & C. Ferguson (eds.), Child Phonology, vol. 1. 23-41. New York.

[15]Oliverius, Z. F. 1967. Fonetika russkogo jazyka. Prague.

EMOTIONALLY EXPRESSIVE PREREQUISITES OF LANGUAGE UNITS IN RUSSIAN SPEECH

E.N. VINARSKAYA

Maurice Torez Moscow State Pedagogical Institute of Foreign Languages
Ostozhenka 38, Moscow, USSR, 119034

ABSTRACT

Innate emotionally expressive reactions: baby cries, cooing and babbling influenced by the social environment are transformed into language specific intonational signs of emotional expressiveness: vocalizations, increasing sonority segments, pseudo-words and pseudo-sensegroups. They are further transformed in Russian language environment into stressed and unstressed vowel allophones, CV syllables, syllabic rhythmic structures and communicative types of sensegroup.

The material presented here is a result of the synthesis of natural and humanitarian studies of emotions on the one hand and of speech in relation to child early development in Russian culture and language environment, on the other |1,2,3,4,5,6,7,8|.

The quality of innate emotional states and their intensity change simultaneously according to the zone principle. The zone of low values of emotional excitement level is qualitatively indefinite. The zone of its moderate values is emotionally positive while the zone of its high values is emotionally negative. Baby cries appear as part·of emotionally negative states caused by the baby's biological discomfort: hunger, thirst, cold, overheat, etc. Social regulation and normalization of baby cries begin with the decrease of their intensity. This is achieved by the baby by the age of 2 months as a result of the imitation of his mother's voice in the process of their emotional interaction. Intensively moderate baby voice reactions as signs of communicative-cognitive behaviour, start to be opposed to intensive cries as signs of defensive behaviour. The emerging of innate intensively moderate reactions of cooing are beginning to correspond to the dynamic range of spoken speech intonation. Cooing sound tambres have zone characteristics which are conditioned by the zone structure of periodically developing emotional states related to the baby's communicative-cognitive behaviour. In the zone of relatively low values of emotional excitement level the tambre quality is indefinite (ə-tambre); in the zone of moderate values the tambre quality is differentiated according to the tendency in the emotional excitement development. Its growth and consequent increase in the tone of speech tract muscles call forth the advanced movement of the tongue and the spreading of the lips as in smile; this results in the occurrence of emotionally positive "И"-tambre. On the contrary, the decrease of emotional strain and consequent decrease of muscular tone lead to the retraction of the tongue and the protrusion of the lips as in cry, which results in the occurrence of emotionally negative "у"-tambre.

Imitating his mother's voice in the course of their emotional interaction, the baby transforms universal biological tambres of cooing into the tambres of language-specific vocalizations. А,И,У,ə vocalization tambres and their emotionally expressive zone transitions: э, о, ы as well as ь, ъ, ʌ in further development give rise to basic positional allophones of Russian vowels. The characteristic opposition of stressed-unstressed vowels in Russian |9| can be understood from tambre emotionally expressive regularities: unstressed vowels are language derivatives of the vocalizations that express states of low emotional and, consequently, muscular tone (detachment, disinterest).
A further step in the development of emotionally expressive speech means is ensured by the emergence of the baby's innate reactions of crying and laughter. Crying sound elements caused by the decrease of emotional excitement - sobs - can be defined as decreasing sonority segments (VC). While laughter elements caused by the increase of emotional excitement can be defined as increasing sonority segments (CV). The manifestations of great emotional excitement - burst of laud laughter and sobbing (which, as is well known, turn easily into each other) are segments of increasing-decreasing sonority (CVC). Laud laughter and sobbing are opposed to light sobs and gigles with no obvious structure.

Being able to cry and laugh, i.e. to produce sound segments of changeable sonority, the baby starts imitating similar sound complexes in his mother's speech: her laughter, crying as well as syllabic speech units. Syllables consisting of vowels and consonants are, in fact, segments of changeable sonority. Babbling, appearing at the age of 6 months, give favourable grounds for such imitation efforts.

Babbling segments of changeable sonority are quite variable; they are normalized under the influence of Russian speech standards perceived by the baby from his mother's speech. Since the basic structural unit of Russian speech is the CV syllable |4,5|, already with a year-old baby the predominant babbling units are CV segments |10|. These segments are further normalized according to the degree of contrast between their composite initial noisy and final vocal elements. Social normalization of noisy maxima in CV segments is over when they are transformed into CV syllables which are characterized by

a number of syllabic contrastive features.
The baby's physiological bias toward repeating every babbling segment until it has faded, emphasized accentuation of words in his mother's speech, an abundance of words with choree structure - all these factors favour the transformation of babbling segments into CV babbling pseudowords at the age of 9-10 months. Their social normalization is realized in various ways |10|. Choree pseudo-words (CVcv) which prevail in the beginning, by the age of 13-14 months become quantitatively equal to jambus pseudo-words (cvCV); by the age of 18 months jambus pseudo-words become prodominant.
Such pseudo-word structures are in full accord with the baby's emotional state, when of primary importance are moderate emotions of the positive zone, characterized by the increase of emotional strain. The number of CV segments, making up a pseudo-word, is normalized as well: by the age of 8 months a pseudo-word comprises 4-5 segments on the average, while by the age of 12-16 months their number is reduced to 2,5 segments which is close to the average number of syllables in Russian word-forms 2,3. Finally, the qualitative structure of pseudo-words is normalized. The prominence of a segment is achieved by its duration, laudness or pitch. The older the baby is, the more often segment prominence is realized by a complex of several means among which duration is dominant. This fact conforms to the nature of word stress in Russian.
Babbling pseudo-words have various zone structures which preaetermine tneir emotional expressiveness. Jambus emotionally positive pseudo-words (cvCV) are genetically related to the increase of emotional excitement, while choree emotionally negative pseudo-words (CVcv) are related to its decrease. Pseudo-words of indefinite temporal structure represent a relatively low level of excitement in the development of emotional states while pseudo-words of the cvCVcv type represent a relatively high level. In the course of emotional interaction with the baby adults constantly attract his attention to various objects, thus "marking" them by their own emotions |11|. The baby masters the rhythmic patterns of Russian words as normative variants of pseudo-word zone characteristics.
In later melodic babbling a second year old baby uses sequences of babbling pseudo-words which correspond to the phonetic structures of sensegroups in the speech of adults. Of fundamental significance in these pseudo-sensegroups are melodic parameters. The increase of the speaker's emotional excitement, as a result of his wish to receive certain emotional information, calls forth the occurrence of pseudo-sensegroups of rising pitch movement. The absence of such a wish is marked by the decrease of emotional excitement and results in the occurrence of pseudo-sensegroups of falling pitch movement. The first, emotionally positive, type of pseudo-sensegroup is realized in questions, requests, apologies, thanksgivings, encouragements and so on. The second, emotionally negative, type of pseudo-sensegroup occurs in the transmission of information to the listener (various kinds of nominations, statements and so on). Pseudo-sensegroups of emphasized rising-falling pitch movement are typical of affect-volitional states, they are opposed to pseudo-sensegroups with vague melodic structure (these are pseudo-sensegroups of high and, correspondingly, low zone levels of emotional excitement).
Generalization of various emotionally expressive pseudo-sensegroups according to the character of the pitch movement transforms them into Russian normative communicative types of sensegroups: complete, incomplete, interrogative, exclamatory.
The table below presents systemic correlates of emotionally expressive intonational signs and the corresponding phonetic forms of native (Russian) language that are based in speech social environment.

Table

Systemic Correlates of Emotionally Expressive Intonation Signs and Russian Phonetic Forms

Intonational signs of emotional expressiveness	Phonetic forms of language signs
Vocalizations	Stressed and unstressed vowels
Increasing sonority segments	CV syllables
Pseudo-words	Rhythmic syllabic word structures
Pseudo-sensegroups	Melodic contours of communicative types of sensegroups

Thus, emotionally expressive intonational means can be transformed into phonetic forms of the native language only under the influence of normalizing affects of the social environment.
Mastering phonetic forms of language signs is ahead of mastering their meaning. Hence the phonetic forms themselves have two functions in speech: mastered linguistic function and a prior one - emotionally expressive.

REFERENCES

|1| Л.С. Выготский "Проблемы развития психики" Соб. соч., т.III, 'Педагогика", 1983.
|2| П.В. Симонов "Эмоциональный мозг", Наука, 1981.
|3| В.К. Вилюнас 'Психология эмоциональных явлений" МГУ, 1976.
|4| Р. Якобсон. "Звуковые законы детского языка и их место в общей фонологии. Избр. работы 'Прогресс", 1985.
|5| H. Truly, J. Lind. "Cry sounds of the newborn infant". Acta paediatrica scandinavica. Suppl. 163, 1965.
|6| E. Sedláčkova "Development of the acoustic pattern of the voice and speech in the newborn and infant". Academia, Praha, 1967
|7| "Talking to children". Ed. by C. Snow and Ch. Fergusson. Cambridge University Press, 1977.
|8| Е.И. Исенина. "Дословесный перид развития речи у детей". Саратов, СГУ, 1986.
|9| Л.Р. Зиндер "Общая фонетика". Высшая школа, 1979.
|10| Е.Н. Винарская "Раннее речевое развитие и проблемы дефектологии", "Просвещение", 1987.
|11| А.Н. Леонтьев. "Деятельность, сознание, личность", Политиздат, 1975.