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 reiiable 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):

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

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