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-parameters 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 EPhoS 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-oriented 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 analyzed 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

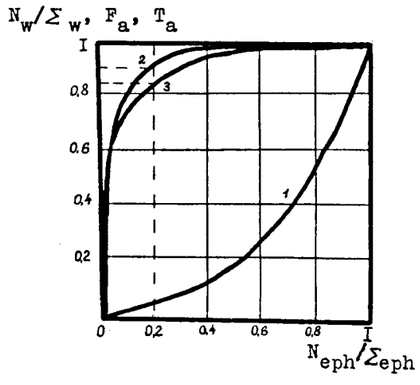

Fig. 1. The dynamics of appearing (1), accumulating of frequence (2) and duration (3) of the original EPhoS.

the results of sence analysis for that purpose.

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 (appea - ring 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 phone - mes. 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).

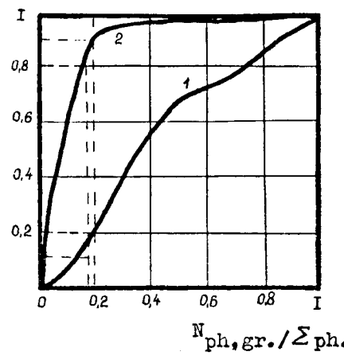

Fig. 2. Dependence of the number of the EPhoS (1) and the number of the recognized words (2) on the number of the phoneme groups.

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.