ONE OF THE METHODS OF AUTOMATIC SYLLABLE SEGMENTATION FOR CONNECTED SPEECH

RODMONGA POTAPOVA

Dept. of Experimental Phonetics, Institute of Foreign Languages, Moscow, USSR 119034

ABSTRACT

The present report proposes a method of automatic speech segmentation using syllable templates based on soft-hardware method of speech analysis to cope with several difficulties, i.e. indistinct syllable boundaries; absence of data on the amount of phonemes in a syllable, localization of phonemes on temporal axis; relative complexity of processing. The report describes one of the possible approaches to continuous speech segmentation, which is of great importance in solving tasks of automatic recognition and understanding of a spoken message in the process of "man-to-computer" communication using a natural language and spoken speech as the basis.

INTRODUCTION

Method based on syllable templates is widely used in automatic speech recognition systems. At present three approaches to automatic speech recognition using syllable templates are known:

-a) input speech is segmented into syllable-sized units which are matched against stored syllable templates;
-b) words synthesied from syllable-sized units are matched against input words;
-c) input speech signal is analyzed, segmented into soundlike (or smaller) segments with subsequent forming of syllable units.

From literature it is known that in the first type of methods the difficulty is that they are liable to segmentation errors, while the difficulty in the second and the third approaches is an increase in the complexity of processing. Though the method using syllable templates is rather effective because it takes into account most of coarticulation phonemes models without considerable extention of memory and increase of processing rate, it is limited now, first of all, by the way of presentation of input material (separately pronounced words) and limited number of speakers [2]. Complexity of speech recognition as the result of increase of a number of speakers and extention of lexicon, clearly shows the advantage of the syllable segmentation method based on changes of feature parameters of speech wave [4].

It is proposed that the method can integrate with the method using syllable templates because, for example, in the Russian language there are comparatively small amount of main syllable types ($n \approx 200$). It is possible to form plausible hypotheses on phoneme structures of a segment using data of probable occurence of these syllables and of phonetic correlates of distinctive phoneme features forming these syllables [5, p.98].

PROPOSED METHOD

Well-known principles of syllable recognition of speech are based either on analysis of average signal energy $E_{(t)}$ (the envelope) and determination of minimum and maximum of the envelope, intervals between minima of signal envelope being taken for syllable boundaries and maxima of the envelope being located on the nucleus of a syllabic vowel; or on the analysis of the wave itself by segmenting it according to maxima and minima with subsequent forming of syllable units based on typical properties of segments, mainly, of an energy character [3].

In the method of syllable segmentation there are several difficulties:

-indistinct boundaries between syllables;
-absence of data on the amount of phonemes in a syllable and on the localization of these phonemes in time;
-heuristic approach in forming syllable units from segments;
-relative complexity of processing;
-insufficient self-descriptiveness of parameters which have very often nothing to do with parameters of vocal apparatus with subsequent false maxima and minima which are the result of energy changes in bands depending on peculiarities of vocal apparatus more than on peculiarities of some parameters.

The present method of automatic speech recognition using syllable templates is free from these defects. It is based on

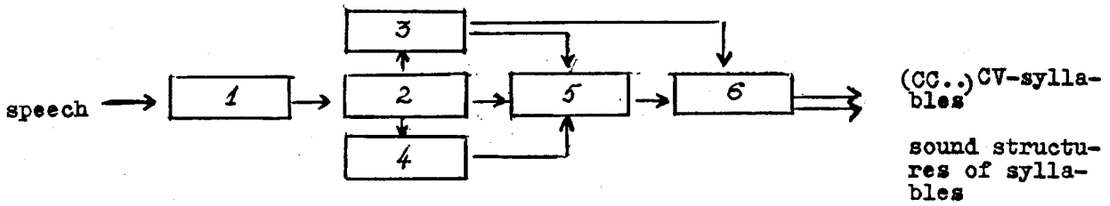

Fig.1. Block diagram of automatic speech segmentating process using syllabic units

soft-hardware method of analysis of speech sounds. This is achieved:
-a) by segmentation of several parameters from a speech sound, (formant frequency, intensity averaged by analysis interval, frequency averaged by transition of signal through zero, pitch, etc) by analog or digital processing;
-b) by segmentation of a sequence of parameters correspondent to speech phrase;
-c) by analysis of a sequence of segment;
-d) by summery up derivatives of parameters normalyzed and averaged by all parameters in time;
-e) by obtaining segmental function $S_{(t_j)}$;
-f) by location of extreme maxima of this function in a signal which are characterized by intensity decrease;
-g) by taking the extreme maxima for the boundaries of open syllables and extreme maxima $S_{(t_j)}$ between boundaries of syllables for the boundaries of sounds in every syllable.
The value of the function at the moment t_j is:

$$S_{(t_j)} = \sum_{i=1}^{n} \frac{k_i \Delta A_{i,t_j}}{\bar{\Delta} A_{i,t_j}}$$

A_i is the i-th feature, where k_i is the weight value of a parameter, n is the number of speech parameters utilized.
Exact boundaries of (CC...)CV-syllables are determined by the highest peaks of S_{t_j} in time, that are characterized by intensity decrease.
Peaks of $S_{(t_j)}$ inside the syllable boundaries determine boundaries between sounds.
The organization block diagram is illustrated as follows (Fig. 1).
-Signal processed is put into segmenter of speech parameters (block 1) as described in [1].
-From input register sequence of parameters correspondent to input utterance is transferred to storage (block 2).
-Stored data are processed by blocks 3 and 4.
-Block 2 using the sequence of parameters finds and stores location on temporal axis and quantity of maxima of $S_{(t_j)}$.
-Block 4 finds temporal parameter intervals of decrease of signal intensity.
-Block 5 locates absolute maxima of $S_{(t_j)}$ on temporal intervals of decrease of signal intensity.
-Block 6 finally determines boundaries of open syllable and boundaries between sounds in a syllable using data obtained from blocks 3 and 5.
Thus in the output of the system we find boundaries of open syllables and sound structures in every syllable. Accuracy of boundary measurements is determined by discrete time quantization of data transfered from separater of informative parameters (block 1).
As it has been demonstrated by numerous experiments, phonotactic information is of great importance. In a number of automatic speech recognition systems such information was not taken into account, which significantly reduced the percentage of correctly recognized words in continuous speech. Information of phoneme co catenation in speech chain forms a filter which passes actual combinations of phonemes and blocks phonotactically impossible ones. Information of phoneme concatenation inter morpheme and word junctures is also of great significance.
The characteristic pronunciation features, phonologic, accentual and rhythmic peculiarities of the utterance in different languages can lead to certain constraints and additional complications in detecting syllabic boundaries. The difficulties increase in case of languages with a high amount of consonants in speech continuum, in which the correct extraction of syllables helps to solve the task of correct recognition and understanding of a spoken message.
Thus, three-level segmentation of continuous speech is proposed: first, (CCC...)CV-syllables are marked along with the exact definition of their boundaries, then syllables are segmented into certain sound types. At the final stage the boundaries of linguistic units are specified on the basis of phonotactics.
Usage of additional acoustic data about the way and place of sound formation allows to define sound structure of segmented syllables based on spectral, temporal and energy parameters which are rather easily and reliably separated by special devices or algorithmic.

REFERENCES

1. T.A.Barašova, B.N.Rudnyi, V.N.Trunin-Donskoj, "Ob avtomatičeskoj segmentacii rečevogo potoka pri vvode rečevogo signala s ustrojstva vydelenija priznakov", V kn.: Rečevoje upravlenije, Moskva, 1972.

2. H.Fujisaki, K.Hirose, T.Inone, "Automatic recognition of spoken words from a large vocabulary using syllable templates", IEEE, 1984.

3. C.Gagnoulet, G.Mercier, R.Vives, J.Vaissiere, "A multi-purpose speech understanding system", IEEE, International Conference on Acoustics Speech & Signal Processing, 1977.

4. R.K.Potapova, "Avtomatičeskaja segmentacija reči na psevdoslogovyje edinicy" ("Automatic segmentation of speech into pseudo-syllabic units", Proceedings of the first Intern. Workshop on Natural Communication with Computers, Warsaw, 1980.

5. "Urovni jazyka v rečevoj dejatelnosti. K probleme lingvističeskogo obespečenija avtomatičeskogo raspoznavanija reči", Leningrad, 1986.