AUTOMATIC IDENTIFICATION OF ACCENTUAL-RHYTHMICAL STRUCTURE OF SPOKEN WORDS

B. Lobanov, T. Levkovskaya, E. Karnevskaya

Inst. of Egin. Cybernetics Ac. of Sc., Minsk, Belarus

ABSTRACT

A model of automatic identification of accentual-rhythmical structure (ARS) for isolated words is described. The identification of ARS is based on measuring of the duration, intensity and F1 value for each vowel of the word and their comparison with the set of patterns for all possible types of word ARS.

INTRODUCTION

The word ARS is generally defined as the distribution of force and distinctness of the articulation of the vowel sounds in the word. The acoustic correlate of the vowel articulation force is energy of speech signal, i.e. the intensity (amplitude) and duration, and the correlate of distinctness is stability of the vowel spectral characteristics.

It follows from the definition that words with a different position of the stressed vowel in a word display differences in ARS. The stressed vowel has the greatest force and distinctness in a word. If we assume that the stressed vowel articulation force and distinctness equal to 3 points, then the other positions of vowels in Russian words will be characteristics by the following values:

```
v v...v v V v...v v
2 1 1 1 2 3 1 1 1 2
```

It can be seen from the above scheme that the 2 points force is characteristic of the prestressed, initial and final vowels, while the force rest of the vowels is indicated with 1 point.

Information about ARS is a very important component of both the overall and phonemic recognition of a word [1,2]. In the overall recognition the knowledge of the word ARS makes it possible to considerably narrow the range of pretenders for the final decision about the word and thus increase the recognition reliability and speed, which is especially important in working with large vocabulary lists. In phonemic word recognition the knowledge of the degree of reduction of each vowel in the word, obtained as a result of the latter's ARS recognition allows to define more precisely the range of probable vowel allophones for each concrete position in a word and thus increase the probability of their correct choice.

EXPERIMENTAL STUDY

In essence ARS recognition is reduced to the identification of the stressed vowel position in a word. It is commonly known, that the stressed vowel possesses on average the maximal energy (product of duration and mean amplitude). However in real speech conditions this rule is not observed in all cases. As bas been shown by previous investigations, the energy value of a vowel is generally influenced by the following factors:

- the position of the vowel in relation to stress: prestressed, stressed, poststressed;
- the position in relation to word boundaries: initial, word-medial, final;
- whether the 1-st vowel in the word is covered or uncovered by a consonant;
- wether the last vowel in the word is open or closed by a consonant;
- wether the vowel being analyzed pertains to the narrow (u, i, ...) or broad (a, o, ...) articulatory group.

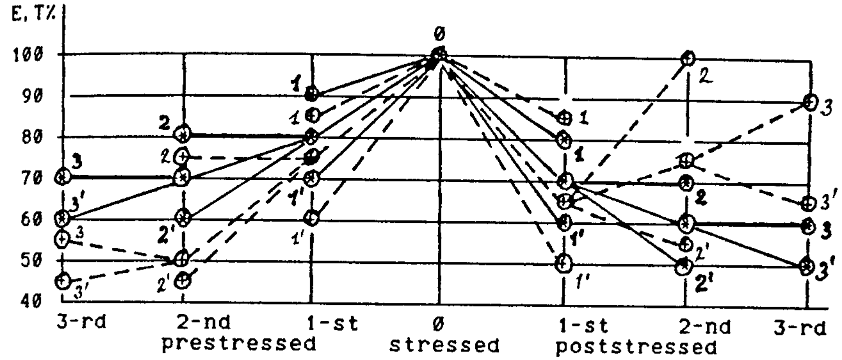

Fig. 1. Experimental distributions of the corrected relative energyof vowels (solid lines - energy, dotted - duration).

The aim of the present experimental study was to assess quantitatively the impact of the above factors on the vowel intensity and duration. The experimental corpus included Russian words containing from 1 to 7 syllables with a different position of word stress and different vowels in stressed and unstressed syllables. The selected words were recorded twice by two native speakers of the language and then digitized with the discretization at 16 kHz and 12 bits. The next step was marking the boundaries between phonemes by hand. Measurings of the vowel duration and intensity as well as the collecting of statistical data were carried out automatically. Besides, F1 frequency was found for each vowel so that a decision could be taken as to what group (narrow or broad) the vowel belongs. On the basis of the data obtained vowel duration and intensity distributions were received normalized with relation to the maximal value. These distributions covered words with different rhythmic structures and with different combinations of narrow and broad vowels: n-N-n, b-B-b, n-B-n, b-N-b, in prestressed, stressed and poststressed syllables. These data were then used for building up the common normalized distributions of the vowels duration, intensity and energy, corrected against the F1 meaning (fig. 1).

In fig. 1. the lines refer to the following ARS:
(1-0) - one uncovered prestressed syllable;
(1'-0) - one covered prestressed syllable;
(0-1) - one open poststressed syllable;
(0-1') - one closed poststressed syllable;
(2-0) - two prestressed syllables with an uncovered initial syllable;
(2'-0) - two prestressed syllables with a covered initial syllable;
(0-2) - two poststressed syllables with an open final syllable;
(0-2') - two poststressed syllables with a closed final syllable.
(3-0) - three prestressed syllables with an uncovered

initial syllable;
(3'-0) - three prestressed syllables with a covered initial syllable;
(0-3) - three poststressed syllables with an open final syllable;
(0-3') - three poststressed syllables with a closed final syllable.

If a word contains more than 3 prestressed or poststressed syllables, it is possible to continue using lines 3(3') - 0 - 3(3'), and in that case the energy value for the 2-nd prestressed vowel (or poststressed) is shifted to the 3-rd syllable if the total number of syllables is 4, or to the 3-rd and 4-th if the number of syllables is 5 or more, etc.

IDENTIFICATION PROCEDURE

The procedure of taking a decision about the ARS of the word being analysed is described in the following way. At the first stage a sequence of pretenders for the vowel phonemes in the word is identified. The phonemes are analyzed as to their belonging to the narrow or broad types according to whether the F1 frequency exceeds a defenite limit (in this case the limit was defined as F1 = 350 Hz). Next the energy value for each vowel is calculated: Ei = Ti * Ai. This is then specified for broad vowels by multiplying by the coefficient Kn = 0,7, and for narrow vowels by the Kn = 1. Among the [E]i thus obtained is the maximal meaning, in relation to which the normalization of all energy values in the set is carried out. The normalized energies obtained are then compared with the set of pattern characteristics [En]i, similar to those in fig.1. The comparison is carried out with the following sample sequences. If the number of syllables is 2, an alternative comparison with one of the following pairs of sequences is made:

$$\left[\begin{array}{ll} (1 - 0), & (0 - 1); \\ (1' - 0), & (0 - 1); \\ (1 - 0), & (0 - 1'); \\ (1' - 0), & (0 - 1'). \end{array}\right. \qquad (1)$$

The meaning (1') here is taken in the cases when the first vowel is preceded and the second vowel is followed not by a pause, but by one or more consonants. For a three syllable word comparison is drawn with one of the three-unit sequences:

$$\left[\begin{array}{l} (1 -1),\ (2 -0), (0 -2); \\ (1'-1),\ (0 -1), (0 -2); \\ (1 -1'), (2 -0), (0 -2'); \\ (1'-1'), (2'-0), (0 -2'). \end{array}\right. \qquad (2)$$

Similar sample sequences are composed for the number of syllables > 3. To take a decision about the type of the word ARS the energy similarity degree is calculated between each vowel in the word being analyzed and the set of energies pattern in the corresponding sequences (1), (2), ... Next the similarity sum is calculated for each of the possible sequences, shown in the brackets. Among the summed measures of similarity there is the one displaying the maximal degree. This sequence is assumed to be the most likely rhythmic structure of the analyzed word.

RESULT AND DISCUSSION

To check the effectivity of the suggested procedure a software model of the ARS automatic identification was worked out. The speech signal is analyzed with the help of a set of 5 octave filters in the range up to 8 KHZ in the time-window of 16 ms with a step of 8 ms. Out of 5 spectral parameters 3 segmenting parameters - P1, P2, P3 - are formed, displaying the maximal value of P1 - for the vowels, P2 - for the consonants and P3 - for the pauses. The segment boundaries are defined by analyzing the segmenting functions Sk obtained from Pk according to the formula

$$Skn = \mathrm{SUM}_{i=1}^{L/2}(Sk,n+i - Sk,n-i)/L,$$

where n-is current timing, L-analysis window.

Within the boundaries determined in this way the vowels duration and intensity were defined, as well as the occurrence of a consonant or a pause before the first or after the last vowel of the word. In order to identify the vowels nature - high or low - estimation of F1 was carried out on the vowel segments by means of counting the number of zero-crossing at the output of the 1-st filter. The information obtained in this way was then used in the software of ARS automatic identification in accordance with the procedure described in the above section. The ARS automatic recognition algorithm was evaluated for its efficiency first on the speech material corpus described in section 2, and then on a new additional testing set of material. The results of the tests were summed up by fixing two types of errors:

Ms-the general percentage of erroneous identification of the word ARS;

Mr-the percentage of errors in the recognition of the ARS in the cases where the total number of vowels in the word and their articulatory type was identified and, besides, the decision about the presence / absence of a consonant at the beginning or at the end of the word was correct.

The results of both tests are estimated on average as: Ms=78%, Mr=94%. As is shown by the results of test evaluation the suggested procedure of the ARS automatic identification can be assessed as effective, on condition, however, that the preliminary phoneme segmentation and marking have been sufficiently correct. It should be noted hereby, that in certain cases, analyses for the degree of similarity (nearness) between the ARS being considered and the sample one, can help reveal the defects of phoneme segmentation and marking in the word. If some of the ARS have close similarity measures, they must be subject to further analysis with a view to making clear whether or not there is a vowel cluster or some mistake in segmentation.

REFERENCES

[1] Zue V. (1985), "The use of speech knowledge in automatic speech recognition", Proc. of the IEEE, vol. 73, N11, pp. 75-90.

[2] Carlson R. (1991), "Duration models in use", Proc. of the XII-th ICPhS, Vol 1, pp. 243-246.