UDC 681.3.019:621.39

N. Bykov, Cand. Sc. (Eng.), Prof.; V. Kovtun, Cand. Sc. (Eng.); N. Savinova, Student

RELIABLE METHOD OF SYLLABLE SEGMENTS ALLOCATION IN SPEECH SIGNAL

New method which promotes reliability of syllable segments allocation in speech signal is offered, algorithm and device, for realization of the offered method are suggested.

Keywords: *recognition of language, speech signal, segmentation of signal, sign of appearances, frequency band, device for syllable segmentation, syllable features extraction algorithm.*

Introduction

Usage of information regarding syllables in hierarchial systems of speech recognition allows to decrease the dependence of such systems on the speaker and on the context [1], and, as a result, to improve the quality of recognition. We suggest new method aimed at increasing of reliability of component segments allocation in speech signal, the paper considers the elaborated algorithm and device realizing the suggested method.

Problem statement

In spite of the fact that nowadays there exist numerous commercial systems of automatic speech recognition, the problem dealing with elaboration of methods and means of speech recognition which are not to be adapted for individual voice peculiarities of the speaker is of paramount importance. One of the approaches to the problem of development of speaker-independent recognition systems is the application on upper levels of hierarchial system as recognition elements of sound tyres phonetic characteristics of which don't greatly depend on the speaker and context, for instance, syllables, semisyllables, voiced, fricatives, pauses, etc. In [2] it has been revealed that usage of, for instance, only syllabic information allows already at upper level of recognition to reduce 2-4 times the number of candidates to be classified. Such information is the duration of syllables and their number in the utterance. One of the main parameters, used for delimitation of syllables in speech signal, is its energy [3, 4]. The kernal of the syllable is determined in the place of maxima energy localization, limited by considerable (40 or 50 dB) drops of energy. But in some papers, for instance [3], frequent allocation by this characteristic, of erroneous syllables, formed by high-energy fricative or sonorous sounds was noted. This fact is proved by experimental recordings. In [6], the function of "volume", obtained as weighted sum of amplitudes of signals of 22 frequency channels, located in critical bands is used as the parameter for syllable feature allocation. The obvious drawback of such method of syllable character formation – considerable material and computational expenditures as well as low reliability.

In order to eliminate the drawbacks inherent to above-mentioned methods, our aim is to develop a new method that would improve the reliability of speech signal segments allocation, which correspond to speech syllables as well as to develop algorithm and device able to realize this method. In the next section of the research we will consider mathematical model enabling to determine information characters of component segments. The given model will serve as the base for development of method, algorithm and device intended for allocation of these segments.

Mathematical model and method of syllable characters allocation from speech signal

It is possible to reveal the mechanisms of speech signal spectral parameters representation in space of invariant characters adding to the processes of peripheral auditory system the processes occuring in central hearing system, as it is shown in Fig 1.

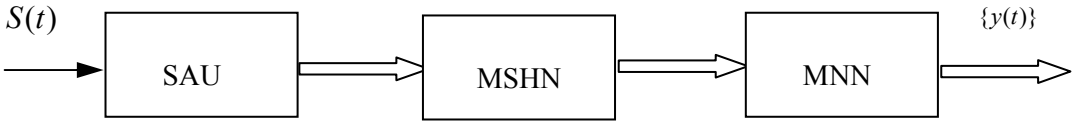

Fig. 1 Generalized model of hearing system

In generalized hearing system model, presented in Fig 1, spectral analysis unit (SAU) shows frequency – selective properties of the model and is a set of filters. Speech signal $S(t)$ is transmitted at the input of these filters. As a rule, filters cover frequency band 250-6400 Hz, where energy of speech signal is concentrated.

Model of sensor hearing newrons (MSHN) describes the functionning of hearing newrons, connected with hair cells of basilar membrane of the ear. It takes into account such hearing effects as dynamic compression of the input signal, its one half-periodic rectification and regulation of single amplification. Mechanisms of hearing perception, presented by given models have already been studied, however numerous attempts of their application in recognition systems did not give desired results. The given paper considers the improved model of hearing system, taking into consideration the functionning of neural network, illustrated by its model (MNN).

Peculiarities of speech patterns formation by neural network model, having signals y(t) at the outputs of neurons have been analyzed in [7, 8]. The analysis, carried out, showed that characters of speech signal are to be searched among the elements of autocorrelated matrix of speech signal spectral parameters:

$$\left\| y_x \right\| = x \cdot x^T = \begin{Vmatrix} x_1 \cdot x_1 & x_1 \cdot x_2 & \dots & x_1 \cdot x_n \\ x_2 \cdot x_1 & x_2 \cdot x_2 & \dots & x_2 \cdot x_n \\ \dots & \dots & x_i \cdot x_j & \dots \\ x_n \cdot x_1 & x_n \cdot x_2 & \dots & x_n \cdot x_n \end{Vmatrix} \quad (1)$$

Proceeding from the results obtained, new method of allocation of component segments characters has been developed. Envelope signals in frequency bands Δ_1 = 800 – 2500 Hz and Δ_2 = 250 – 540 Hz are used as initial parameters for their formation. The resulting parameter, further used for allocation of syllable characters, is obtained by means of correlation method and written in the following form:

$$U_c(t) = U_{\Delta 1}(t) \cdot U_{\Delta 2}(t), \quad (2)$$

where $U_{\Delta 1}(t)$ envelope energy in frequency band Δ_1, and $U_{\Delta 2}(t)$ – envelope energy in band Δ_2.

Frequency range of the first band filter 3, equal 250-540 Hz, is selected due to the fact, that it lacks the energy of high-energy fricative sounds, such as /ш/ and /ч/, which create erroneous component nuclei as well as because greater part of energy of all voiced sounds, including vavels is concentrated in this range, however, in the given range, the energy of sonorous sounds, such as /л/, /м , /н/ corresponds to the energy of vavels, that is why, determination of component segments only by speech signal by-passing in the given range will be accompanied by numerous mistakes. Hence, frequency range of the second band-pass filter 4, is selected within the limits of 800-2500 Hz, where the energy of voiced sounds minimum two times exceeds the energy of sonorous sounds.

While perfoming the operation of envelope multiplication $U_{\Delta 1}(t)$ and $U_{\Delta 2}(t)$ in resulting temporal function the of curve sections in the region of voiced sounds takes place due to correlation of their energies in both ranges, and erroneous maxima of energy, caused by the presence of considerable portion of fricative sounds energy in the band of 800-2500 Hz, are eliminated by their multiplication by practically zero value of fricative sounds amplitude in the band of 250-540 Hz.

Algorithm and device for syllable characters allocation

In accordance with the description of the device intended for component segments allocation the algorithm of its operation includes the following steps:

1. Input of speech signal.
2. Filtration of the signal by means of two Batterword band filter of the fourth order in operation bands 250 – 540 Hz and 800 – 2500 Hz correspondingly.
3. Detection of filters output signals in order to obtain enveloping.
4. Multiplication of enveloping output signals of the filter.
5. Differentiation of the resulting signal.
6. Comparison of the results obtained with positive and negative threshald voltages and allocation of logic signal.
7. Formation from the obtained signal, logical signals for positive and negative half-periods of differentiated signal.
8. Allocation of syllable segments and syllable centres by means of logical addition and multiplication of the obtained logic signals correspondingly.

Scheme of the algorithm is given in Fig 2.

Such denotation are used in the algorithm:

ff₁ = filter (p₁,p₂, s) and *ff₂ = filter (p₃,p₄, s)* – functions of filtration in frequency bands Δ_1 = 800 – 2500 Hz and Δ_2 = 250 – 540 Hz correspondingly; $U_{g1} = abs\ (ff_1)$, $U_{g2} = abs\ (ff_2)$ – envelope signals in indicated bands.Other designations are explained in the description of operation of character allocation device.

The device intended for allocation of component segments in speech signal operates in the following way (see Fig 3).

Speech signal is detected by acoustic sensor, then it is transformed into electric signal and is sent to the input of the amplifier. Electric signal, amplified to the value, sufficient for operation of further stages, enters the inputs of two band-pass filters having the bands $\Delta_1 = [p_1;p_2]$ = 800 – 2500 Hz i $\Delta_2 = [p_3;p_4]$ = 250 – 540 Hz. Amplitude detector, the input of which is connected to the output of the band-pass filter, allocates U_{g2} envelope of speech signal in 250 – 540 Hz band. Amplitude detector, the input of which is connected to the output of the second band-pass filter, allocates U_{g1} envelope of speech signal in 800 – 2500 Hz band. Voltages U_{g1} and U_{g2}, sent on the inputs of signals multiplier are multiplied, and as a result of multiplication at the output of multiplier voltage U_n, appears, it equals:

$$U_n = U_{g1} \cdot U_{g2}.$$

Frequency bands of the first and second band-pass filters are selected so that as a result of realization of multiplication operation only the energy of voiced sounds is correlated this bands to elimination in envelope U_n maxima, corresponding to sections of high energy fricative sounds. This voltage passes to the input of differentiator 9, characters generator 8 (Fig 3), at the output of which voltage U_{ng} is generated, being proportional to the derivative of voltage U_n. Voltage passes to the first inputs of threshold circuits 10 and 11. Positive threshold voltage U_{n1} passes to the second input of threshold circuit 10, and negative threshold voltage U_{n2} passes to the second input of threshold circuit 11, U_{n1} and U_{n2} are selected so that $|U_{n1}| = |U_{n2}| \approx 50$ mV, false operation of threshold circuits due to background noise while transition across zero values of U_{ng} is eliminated. Clipping signals are obtained from the outputs of threshold circuits 10 and 11, these signals are reduced to standard levels of digital signals. They are sent at the inputs of ABO 12 circuit, at the output of which digital signal $U_д$. . Fronts of signal that passes to the input of T trigger with direct

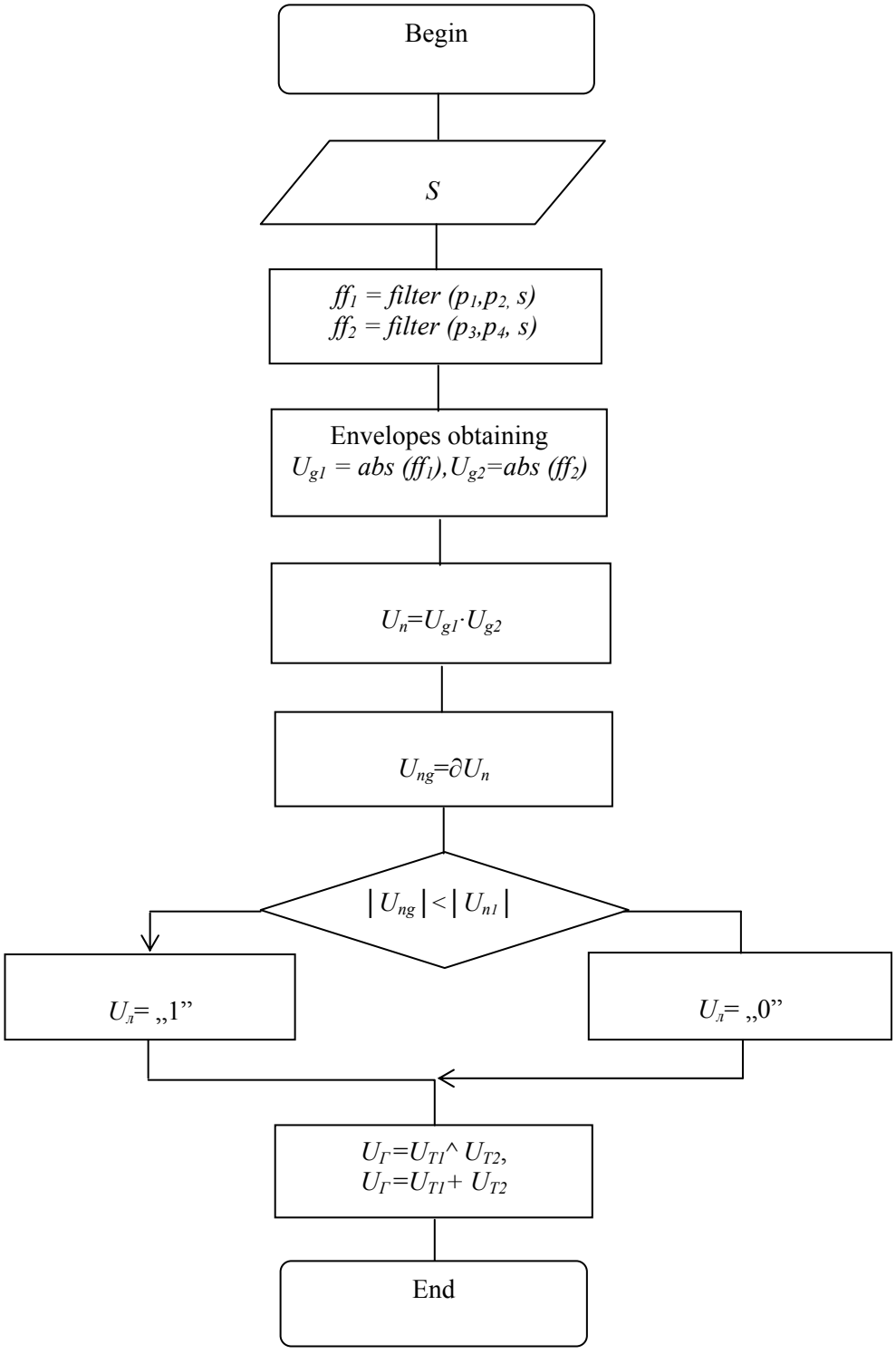

Fig. 2 Algorithm of syllables character allocation

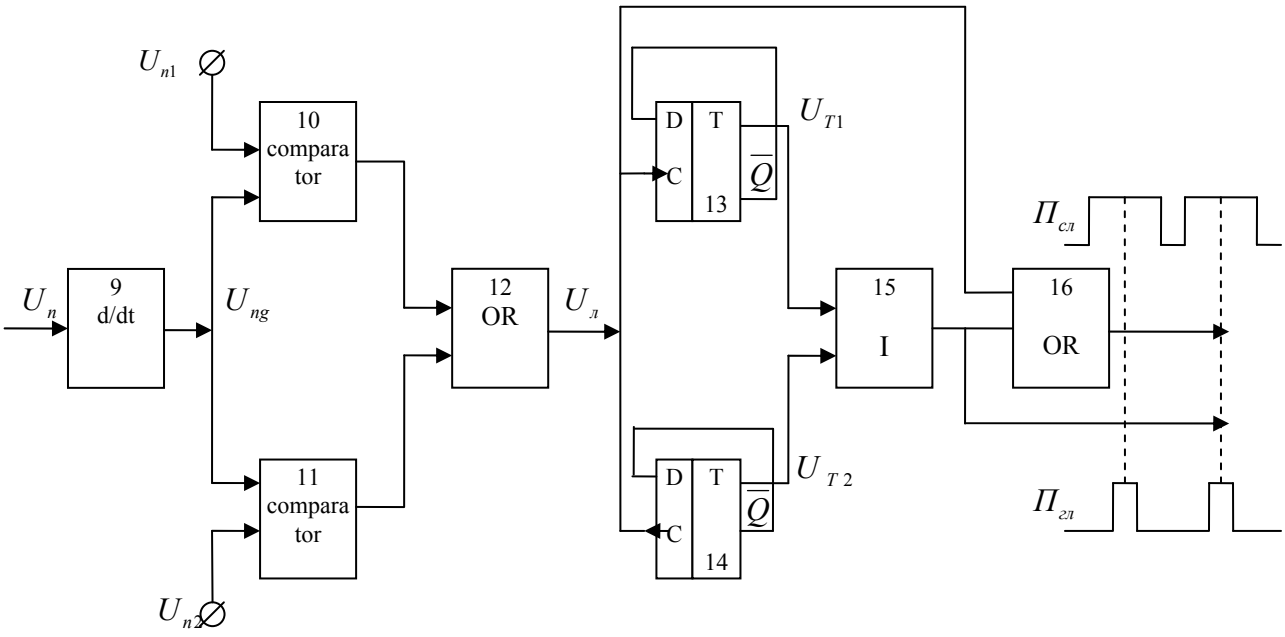

Fig. 3 Functional diagram of generator of segment boundaries of syllables

dynamic control 13 each time reverse it into opposite state, and as a result, digital signal U_{T1} is formed at the output. T trigger with reversed dynamic control 14, is bridged by drops of input signal $U_л$ and generates signal U_{T2} at the output. Signals U_{T1} and U_{T2} are sent at the inputs of coincidence circuit 15, at the output of which short pulses $U_г$ are formed, they localize the centres of composite nuclei in the word. These pulses are sent at the first input of ABO 16 circuit, as well as at the output of the device for allocation of component segments in the word, being character $П_{гл}$ for determination of voiced sounds centre location in a syllable. Digital signal is sent to the second input of ABO 16 circuit. This signal is united, by means of circuit ABO with the signal $U_г$, as a result, signal $П_{сл}$ is generated at the output of circuit 16, signal $П_{сл}$ allocates segments in a word, corresponding to location of syllables in a word. Duration of the signal $П_{сл}$ and the number of signals $П_{гл}$ are used by recognition system for classification of the vocabulary by the subsets, formed according to the given characters.

For testing of the suggested method the experiment was carried out, during the experiment 650 syllables were segmented using the above-mentioned method. Statistic processing of experimental data enabled to calculate the reliability of the given method, it is 96.4 %, whereas the reliability of other methods applying the equivalent test sample is 76%.

Conclusions

The suggested method and algorithm, intended for syllables character allocation, based on improved model of human hearing system, enables to improve the reliability of speech signal segmentation into composite segments and determine such characters as duration, location and number – application of such characteristics allows to increace recognition rate, reducing the number of alternatives for search 2-4 times, as well as to improve the reability of the device. The developed device, realizing the given method, is more reliable and is less complex as compared with already existing ones, and can be used for development of autonomous systems of speech recogninion.

REFERENSES

1. Быков Н.М. Методы и средства измерения и преобразования информации в системах машинного распознавания речи. – Дис. на соискание уч. ст. канд. техн. наук. – Винница, ВПИ, 1985. – 243 с.

2. N.M. Bykov, I.V. Kuzmin, A.I. Yakovenko. Development of effective strategy of pattern recognition. – Proceedings of SPIE, 2001, Vol. 4225, pp.76 – 83.

3. Джелинек Ф. Разработка экспериментального устройства, распознающего раздельно произносимые слова. // Тр. ин-та инженеров по электронике и радиоэлектронике.: Пер. с англ. 1985. – Т. 73. – № 11. – С. 91 – 100.

4. Биков М.М., Грищук Т.В. Методи підвищення дикторонезалежності опису і розпізнавання мовної інформації в мережі INTERNET // "Інтернет – Освіта – Наука – 2002", третя міжнародна конференція ІОН – 2002, 8 – 12 жовтня 2002 р. Збірник матеріалів конференції. – Вінниця: УНІВЕРСУМ – Вінниця, 2002. – Том 2.– С. 329 – 332.

5. Методы автоматического распознавания речи / Под ред. У. Ли.– М.: Мир, 1983. – Т.1. – 200 с.

6. Ruske C., Schotola F. An approach to speech recognition using syllabic decision units. – Proc. 1978, IEEE ICASSP, Tulsa, 1978. – N.Y., 1978, pp. 772 – 725.

7. Ковтун В.В. Вибір інформативних ознак в задачі ідентифікації диктора // МКІМ – 2002. Міжнародна конференція з індуктивного моделювання. Львів, 20 – 25 травня 2002: Праці в 4-х томах. – Львів,ДНДІ, 2002. – .Т.1, ч. 2 – С. 280 – 287

8. Биков М.М., Грищук Т.В. Розпізнавання мовних образів з використанням нейромережевого підходу // МКІМ – 2002. Міжнародна конференція з індуктивного моделювання, Львів, 20 – 25 травня 2002: Праці в 4-х томах. – Львів: Державний НДІ інформаційної інфраструктури, 2002. - Том 1., Ч. 2. – С. 203 – 207.

Nikolay Bykov – Professor of the Department;

Vyacheslav Kovtun – Assistant Lecturer of the Department;

Nataliya Savinova – student.

The Department for Computer Control Systems, Vinnytsia National Technical University