Pre-processing Voice Signals for Voice Recognition Systems

Gulmira K. Berdibaeva[2], Oleg N. Bodin[1], Valery V. Kozlov[1], Dmitry I. Nefed'ev[1],
Kasymbek A. Ozhikenov[2], Yaroslav A. Pizhonkov[1]
[1] Penza State University, Penza, Russia
[2] Kazakh National Research University. K.A. Satpayev, Almaty, Kazakhstan

***Abstract* – The article considers the pre-processing voice signals for voice recognition systems based on the use of artificial neural networks. Based segmentation preprocessing is put in the speech signal according to a phonetic transcription of language, in order to reduce the amount of data supplied to the input of the neural network, which considerably improves its input data sensitivity. Application of numerical methods in processing will reduce acoustic noise impact on the speech signal segmentation, which will more accurately identify the areas of classification. Simulation results of the speech signal partition into components are shown, i.e. the selection of phonemes which will be the voice message classification.**

***Index Terms* – Voice control, speech recognition, eigenvalues, artificial neural network.**

I. INTRODUCTION

VOICE Control System technology is a major area of research for several decades due to the almost unlimited scope of their possible applications, from household appliances to the special-purpose systems. Unfortunately, their distribution on the practice is still constrained by insufficient high-grade automatic speech recognition in conditions of intense external acoustic noise.

For example, *Google's* speech recognition system is based on cloud technologies. Voice commands and requests submitted for recognition in the cloud system, and then, sent to the various instruments and devices in the recognized form [1,2]. The principle of operation of such system is similar to the principle of work of *Siri* - the voice recognition technology implemented in Apple's devices [3]. Microsoft released a virtual voice assistant *Cortana* with elements of artificial intelligence, which can replace a standard search engine, and will be called by clicking the "Search" button [4]. The request you can print manually or set voice. The necessary information will find it on the basis of the search results in *Bing, Foursquare* systems and among the user's personal files. The disadvantage of this system is the lack of support for the Russian language.

The absence in Russia own independent speech recognition system (SRS) makes end-users dependent of foreign development, which requires a constant internet connection to access databases that are stored abroad. The result is that there are risks of restricting access to these resources, and therefore there is the need to create SRS, which will not be subject to these factors.

Consumers of domestic SRS are users who need to simplify the communication with technique, and also people with disabilities.

II. PROBLEM DEFINITION

Based on the foregoing, development of hardware and software is actual that allows control different instruments, devices, instruments, and virtual objects using your voice. The problem of recognition of speech commands is a central element of the voice control system and one of the priority directions in research of artificial intelligence. Voice Control process includes the following basic steps:

- registration of the voice commands, implemented using a microphone and a PC sound card;
- pre-processing and voice recognition commands, implemented with the help of the instruction decoder;
- execution of the voice instruction, realized by the actuator.

Through the use of advanced technologies computer simulation it is possible to improve the accuracy of automatic segmentation of speech signals for the recognition problem of voice commands based on the artificial neural networks with pre-processing input data.

At the stage of pre-processing the input data we see the following tasks:

- only use the general characteristics of the speech signal, because usually at this stage there is no specific information about the content of the speech utterance;
- work not only with isolated words, but also with continuous speech;
- resistance to external noise occurring when recording speech signal or present in the communication channels;
- the level of type I error (the number of missed true boundaries as a result of manual markup to the total number of segments of boundaries) not more than 20%.

The purpose of this article is to investigate the pre-processing voice signals for voice recognition systems based on artificial neural networks.

III. THEORY

Currently, voice recognition systems market is represented by a small number of software vendors. The known speech recognition methods have several disadvantages, such as:

- high dependence on noise , thereby significantly decreasing the probability of detecting a phoneme and

978-1-5090-6688-9/17/$31.00 ©2017 IEEE

hence its recognition [5];

- speech signals detection method, dependent on the random factor, in particular based on a comparison of a large number of different parameters.

One of the most important tasks to the automatic speech processing systems is the segmentation problem according to the phonetic transcription of the language. For example, for a voice recognition task characteristic features of voice should be calculated on certain segments of the speech signal.

Pre-segmentation is needed to solve the problem of automatic speech recognition. Segmentation accuracy largely determines the reliability of automatic speech recognition. Manual segmentation can be an alternative to automatic segmentation on phase of the study and at the stage of pretreatment. However, it requires the presence of experienced linguists, as well as a significant investment of time and effort from both of the lack of breaks in continuous speech words, and because of coarticulation. Coarticulation process occurs at the boundary consistently produced sounds, it greatly facilitates the correct perception and understanding of speech, but it complicates the task finding the boundaries of the segments. In addition, it is practically impossible to accurately reproduce the manual segmentation results due to the subjectivity of the human auditory and visual perception. Such problems do not arise with the automatic segmentation, which, though not infallible, but it gives reproducible results [6].

Thus, the challenge is to develop reliable SRS that is less susceptible to these factors. Block diagram of such SRS is on th Fig.1.

There are several types of speech segmentation algorithms. Consider an algorithm that does not use a priori information about the phrase, and wherein border segments are determined by the degree of change of the acoustic characteristics of the signal. One disadvantage of this approach is a significant effect of acoustic noise. Therefore, at the stage of pre-treatment is necessary to use methods of reducing the impact of noise on the speech signal segmentation.

As a method for reducing the effect of noise on speech recognition it is suggested to use a method eigenvalues decomposition, which is based on an eigenvalues analysis of the autocorrelation matrix or data matrices [7].

This method provides better resolution and parameter estimation than other parametric methods, especially at low signal / noise ratio, when these methods are not able to distinguish between similar frequency sine wave or other narrow-band spectral components.

The method of signal matrix autocorrelation decomposition on the eigenvalues best suits the task of measurements signals parameters in the noise on the basis of the fact that the analysis of the eigenvalues of the autocorrelation matrix is a division of the information on the two vector subspaces - a signal subspace and the noise subspace.

The autocorrelation sequence (ACS) consisting of M complex sinusoids described as

$$r_{xx}[k] = \sum_{i=1}^{M} P_i \exp(j2\pi f_i k\Delta t) + \rho_\omega \delta[k]$$

where P_j – power i-th sinusoids, and ρ_ω – variance of the white noise.

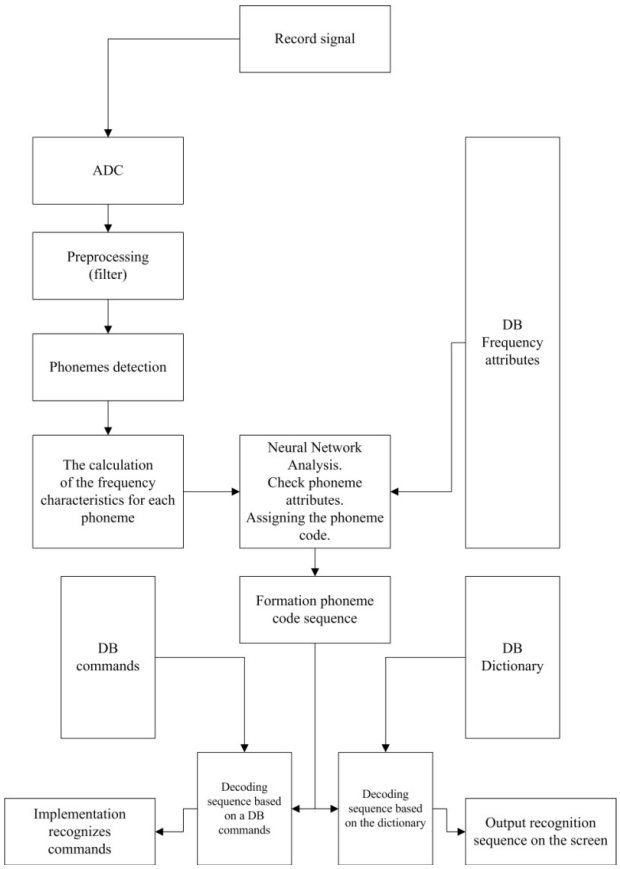

Fig. 1. Block diagram of the proposed SRS.

Toeplitz autocorrelation (p+1)×(p+1) – matrix in the case of complex sinusoid in white noise has the following structure:

$$R_p = \sum_{i=1}^{M} P_i s_i s_i^H + \rho_\omega I,$$

where $\boldsymbol{I}$ – identity (p+1) × (p+1) - matrix, and $\boldsymbol{S}_i$ – a signal vector of dimension p+1, indicative of the i-th frequency sinusoid. $\mathbf{R}_p$ matrix can be represented as the sum of the autocorrelation matrix of the signal $\mathbf{S}_p$ and the noise autocorrelation matrix $\mathbf{W}_p$:

$$R_p = S_p + W_p,$$

where $S_p = \sum_{i=1}^{M} P_i s_i s_i^H$, and $W_p = \rho_\omega I$.

Signal Matrix will have the following eigenvalue decomposition:

$$S_p = \sum_{i=1}^{p+1} \lambda_i v_i v_i^H,$$

where λ_i – eigenvalues, v_i – eigenvectors, and $\lambda_1 \geq \lambda_2 > \ldots > \lambda_{\rho+1}$ [4].

The signal matrix autocorrelation decomposition on the eigenvalues can be used to obtain the spectral estimates or, more precisely, improved procedures for frequency estimates. Preservation alone information corresponding to the signal subspace eigenvectors effectively increases the signal / noise ratio, because it eliminates the power contribution of the noise component subspace.

For the harmonic component in the white noise process is observed

$$y_n = x_n + \omega_n = -\sum_{m=1}^{2p} a_m x_{n-m} + \omega_n ,$$

which matrix notation is as follows:

$$R_{yy} A = \sigma_w^2 A. \quad (1)$$

This expression is the equation of its own process, in which the variance of the noise σ_w^2 is an eigenvalue of the autocorrelation matrix R_{yy}, vector ARMA – A parameter is an eigenvector associated with the eigenvalue σ_w^2, to determine the value of ARMA - parameters in the case where the known value of the autocorrelation function. This equation is the basis of the expansion of the procedure on eigenvalue, and to determine the exact frequency and power p real sinusoids in the presence of white noise, if you know exactly 2p+1 values of the autocorrelation function. Since rely only known values of the autocorrelation function, information about each phase of the sine wave is lost.

After finding the eigenvectors and corresponding eigenvalues, are determined the polynomial coefficients

$$z^{2p} + a_1 z^{2p-1} + \ldots + a_{2p-1} z + a_{2p} = 0. \quad (2)$$

The roots of this polynomial $z_i = \exp(j2\pi f_i \Delta t)$ set frequency sine waves.

$$f_i = arctg\left[\frac{\mathrm{Im}\, z_i}{\mathrm{Re}\, z_i}\right] \cdot \frac{1}{2\pi\Delta t}.$$

After determining the frequency at the roots of the polynomial (2) it is possible to determine the power of sine waves. The values of the autocorrelation function of the $R_{yy}(1)$ to $R_{yy}(p)$ can be written in matrix form

$$FP = r. \quad (3)$$

$$F = \begin{bmatrix} \cos(2\pi f_1 \Delta t) & \cdots & \cos(2\pi f_p \Delta t) \\ \vdots & & \vdots \\ \cos(2\pi f_1 p \Delta t) & \cdots & \cos(2\pi f_p p \Delta t) \end{bmatrix},$$

$$P = \begin{bmatrix} P_1 \\ \vdots \\ P_p \end{bmatrix}, \quad \text{and} \quad r = \begin{bmatrix} Ryy(1) \\ \vdots \\ Ryy(p) \end{bmatrix}.$$

The matrix ***F*** is composed of members, depending on the frequency sine waves, which are determined by finding the roots of the polynomial (2). The power of these sine waves can be found by solving the system of equations (3) relative to the power of the vector ***P***.

Obviously, this method is particularly adapted to solving the problem of estimating the oscillations parameters, as it considers the physical signal parameters, and the same time excludes the effect of noise by dividing information signal and noise subspace.

IV. EXPERIMENTAL RESULTS

Due to the fact that the digitized voice message is too large set of data in order to submit it directly to the input of neural network is necessary to preprocess the data in order to reduce their volume, as well as to identify the sites for which classification will be.

To implement this approach, the voice message is separated into components - phonemes which can distinguish peaks increase and decay level the signal. Pre-treatment carried out in MATLAB using the *envelope()* function, which returns the upper and lower limits of the input sequence, the value of its analytical signal. The analytical signal is using the discrete Fourier transform and the Hilbert transform. This function first removes the mean value, and then adds it back when calculating the maximum and minimum values [8].

The original speech signal, “digitized” 8 kHz is shown in Fig.2.

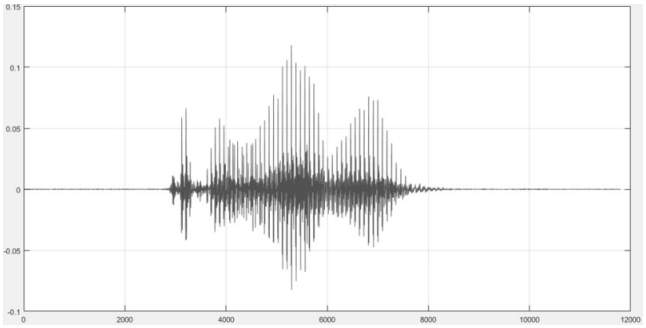

Fig. 2. The word “pravo”

The authors propose to use the “envelope” of the speech signal to identify phonemes. The first envelope is based on the RMS value of the speech signal. Using the function

envelope(nsignal,ceil(length(nsignal)/27.2),’rms’);

We obtain the speech signal envelope shown in Fig.3.

Then we find the second envelope, which is built on the peaks previously received the envelope. Using the function

envelope(up,ceil(length(nsignal)/12.2),’peak’);

we obtain the envelope shown in Fig. 4.

Fig. 4 points marked coordinates on which the separation into components, that will be the selection of phonemes which will be the classification of the voice message.

An example of a voice message separation program into its component phonemes is shown in Fig. 5.

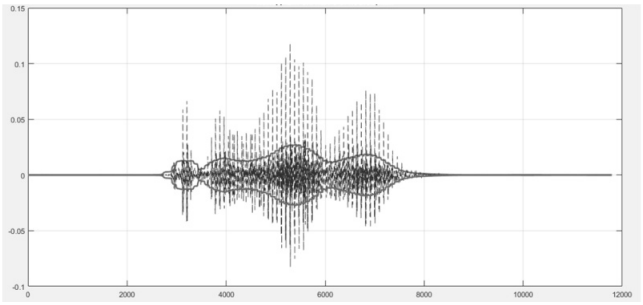

Fig. 3. The envelope for RMS

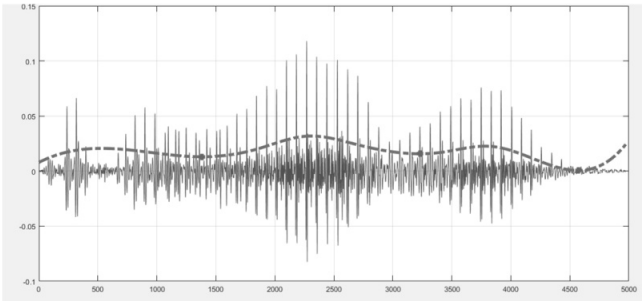

Fig. 4. The envelope on the peaks

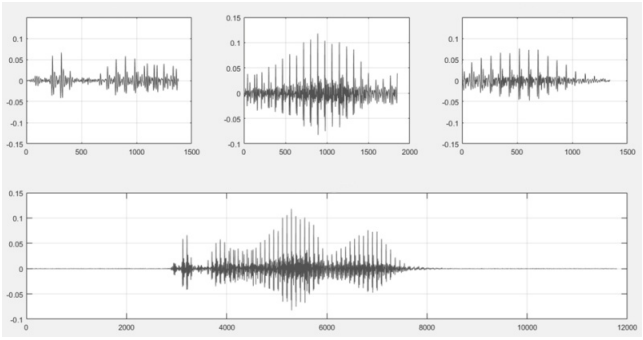

Fig. 5. Example program of the separation into phonemes

V. Discussion of Results

Thus, the authors propose to divide the voice message to phonemes, based on which the input vector will be designed to the neural network. This approach greatly improves the susceptibility of the neural network to the input data in comparison with the known approach, when the input is directly input voice message. The proposed approach is analyzed not only the input sequence, but also the number of phonemes, which simplifies the task of speech recognition and, consequently, decreases the time of recognition of the speech signal, and also improves the accuracy of determining membership input message to a particular class.

The problem of phonemes recognition similar to the classification task, therefore, this problem can be solved by using neural networks designed for data classification. One of these networks is the Kohonen neural network [9], which can recognize the clusters in the data set as well as the closeness of classes. Thus, it is possible to improve the understanding of the data structure, then to specify neural network model. If the data classes will be recognized, they can be identified by the neural network. Kohonen network can be used in the classification problems when classes are already set - then the advantage is that the network will be able to identify the similarities between the different classes.

In our case, the neural network is input segmented voice message, and the neural network must recognize what classes are phonemes, of which it is composed.

VI. Conclusion

The analysis of modern means of voice signals recognition for voice control has shown that the current cloud technology used in this sector, which implies the presence of a permanent Internet connection. The result is that the creation of an autonomous speech recognition system is an actual problem.

Proposed by the authors approach the construction of the voice control system increases the reliability of voice commands recognition through the use of numerical methods for segmentation of speech signals by eliminating the noise component, and subsequent neural network analysis. This increases the susceptibility of the neural network to the input data in comparison with direct feeding to the input of the neural network of speech commands.

References

[1] https://ru.wikipedia.org/wiki/Google_Now
[2] Schuster M. Speech Recognition for Mobile Devices at Google// LNCS. 2010. Vol. 6230. P. 8–10.
[3] https://ru.wikipedia.org/wiki/Siri
[4] https://ru.wikipedia.org/wiki/Kortana_(voice assistant)
[5] Savchenko V. V. Speech recognition method for phonetic decoding of words with background noise cancellation / Information technology. – 2016. – No. 1. – Pp. 76-80.
[6] O. A. Vishnyakova, D. N. Lavrov phonemic segmentation Algorithm based on the analysis of the rate of change of energy of a discrete wavelet transformation / Information technology. – 2011. – No. 4. – S. 146-152.
[7] Kozlov, V. V. Determination of parameters of harmonic signals in terms of the noise and interference based on the method of decomposition of signal eigenvalues / V. V. Kozlov // Modern problems of science and education (electronic journal). – 2013. – № 6. – URL: http://www.science-education.ru/113-10860.
[8] Sergienko A. B. Digital processing of signals – SPb: Piter, 2003. – 604 p.
[9] Medvedev V. S., Potemkin V. G. Neural networks. MATLAB 6 / Under the General editorship of V. G. Potemkin. – Moscow: Dialog-MIFI, 2002. – 496 p.
[10] Pat. US 8175883 B2, Int. C1 G10L21/00 (2006.01). Speech recognition system and method / Invertor: Grant R., Gregor P. // Assignee: Nuance Communications Inc. – Pub. Date 08.05.2012.

Yaroslav A. Pizhonkov,
Russia, Penza, 1994
Secondary,
Student, Penza state University,
Digital signal processing.