Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/7086
DOI URL: http://dx.doi.org/10.21474/IJAR01/7086

RESEARCH ARTICLE

SPEECH RECOGNITION USING THE EMPIRICAL MODE DECOMPOSITION METHOD.

M. S. Medvedev.
Computer Science Dept., Institute of Space and Information Technology, Siberian Federal University, Kirenskogo, 26, Krasnoyarsk, 660074, Russia.

Manuscript Info

Manuscript History

Received: 12 March 2018
Final Accepted: 14 April 2018
Published: May 2018

Keywords:-
Speech recognition; Hilbert-Huang transform; wavelet analysis; empirical mode decomposition.

Abstract

In this paper the using of empirical mode decomposition for creation of the Russian phoneme models in a system of speech-to-text conversion is considered. The proposed method is compared with the Fourier transform and wavelet transform. The experimental evaluation has shown that this method has advantages in the task of speech features formation (within neural network approach).

Introduction:-

In spite of dynamic development of speech system market, the main tasks of this direction, such as continuous speech recognition, national language support and speaker-independence, remain unsolved, while recognition rate declared by software developers in this area turns out to be overstated in reality and prevents users from comfortable use of the existing software products. It explains topicality of researches and development of new efficient algorithms of speech recognition. Development of efficient algorithms of Russian speech recognition is a key moment in solution of the followings tasks:

- speech-to-text conversion,
- speech understanding,
- voice actuation,
- automatic translation,
- speech recognition in telephony (voice menu).

The main goal of researches described in the article is studying Russian phoneme models and development of a method and algorithms of speech-to-text conversion enabling improvement of recognition quality. In order to solve this task, existing methods and algorithms used to build speech recognition systems were systematized, a system of speech-to-text conversion was implemented in software using developed Russian phoneme models and quality of recognition of the developed system was assessed.

A speech signal is an example of a non-steady process, whereby the very fact of change in its frequency-time characteristics is informative. Methods of speech features calculation using Fourier transform, wavelet transform and Hilbert-Huang transform were considered. We analyzed features and singled out advantages and disadvantages of application of these methods in a task of phoneme speech signals classification.

Corresponding Author:- M. S. Medvedev.
Address:- Computer Science Dept., Institute of Space and Information Technology, Siberian Federal University, Kirenskogo, 26, Krasnoyarsk, 660074, Russia.

Material and Methods:-

2.1 Short Time Fourier Transform

Classical Fourier transform deals with signal spectrum taken in the whole range of existence of a variable. Local frequency distribution is of the biggest interest, while initial variable (usually time) shall be preserved. From positions of exact representation of arbitrary signals and functions, Fourier transform has a series of drawbacks that led resulted in emergence of short time Fourier transform and facilitated development of wavelet transform. These are the main drawbacks:

- limited information value of non-steady signals analysis and almost complete absence of opportunities for analysis of its features, since signal features in the frequency domain "smear" over the whole frequency range of the spectrum.
- emergence of Gibbs effect at function jumps during signal reduction and clipping of signal pieces for local detailed analysis;
- harmonicity of basic functions.

Inability of Fourier transform to perform time localization of short-term frequency changes in signals is partially eliminated by introduction in the transform of a window function having a compact carrier, which makes it possible to present a result of the transform in a form of a function of frequency and time window coordinate:.

$$F(\hat{t},\omega) = \int_{-\infty}^{\infty} f(t) \cdot W(t-\hat{t}) e^{-jt\omega} dt$$

2.2 Wavelet transform

Wavelet transform of signals is a generalization of spectral analysis. Bases used for this purpose were called wavelet - functions of two arguments - scale and shift. Wavelet transform ensures two-dimensional representation of the studied signal in the frequency domain in plane frequency-location, whereby scale of a basic function argument serves an analogue of frequency, and location is characterized by its shift. It enables differentiation between big and small signal features and their simultaneous localization on a time scale.

Introduction of wavelet transform helps to reduce influence of Heisenberg uncertainly principle on obtained time-and-frequency signal representation. Any function can be decomposed at some given resolution level (scale) to a row of the following form:

$$f(x) = \sum_{k=0}^{2M-1} s_{j_n,k} \varphi_{j_n,k} + \sum_{j=j_n}^{j_{\max}} \sum_{k=0}^{2M-1} d_{j,k} \psi_{j,k}$$

$\varphi_{j_n,k}$ and $\psi_{j,k}$ – scaled and shifted versions of scaling function φ and "mother wavelet" ψ ;

$s_{j,k}$ – approximation coefficients; $d_{j,k}$ – detail coefficients.

This method was chosen for studying its applicability in creation of phoneme speech models. Features obtained as the result of it also characterize signal in time and frequency plane (figure 1).

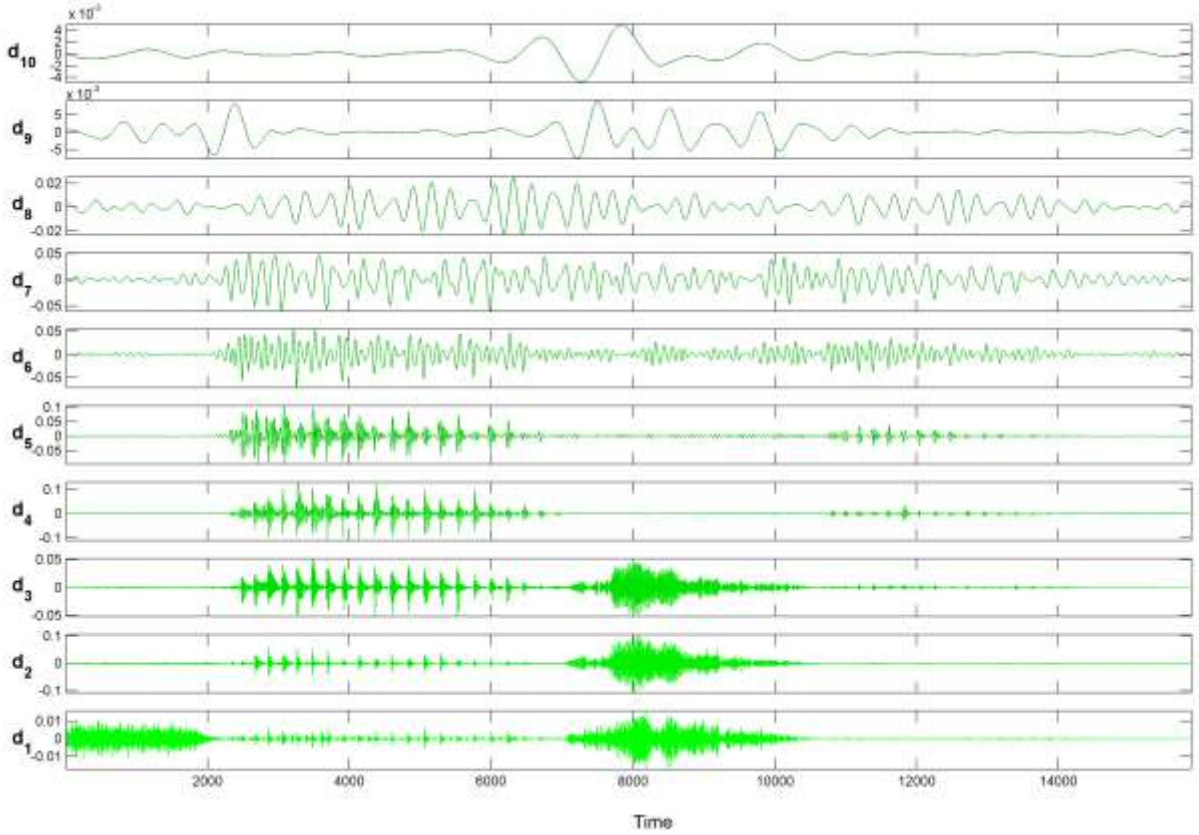

Figure 1:- Wavelet decomposition of speech signal.

2.3 Hilbert-Huang Transform:-

Hilbert-Huang transform is time-frequency analysis of data (signals) requiring no a priori functional basis transform. Basis functions are obtained adaptively immediately from data using procedures for empirical mode functions sifting. A scheme of Hilbert-Huang transform can be divided into two stages. At the first stage experimental data is decomposed to a row of internal mode functions (IMFs). These functions are presented as a transform basis. They are found using a method of empirical mode decomposition (EMD), which can be described in the following algorithm:

1. Searching for local extremums of analyzed signal.
2. The upper and lower signal envelopes are built using its minimum and maximum values found at the previous stage.
3. Function of mean values m(t) between the envelopes is determined:

m(t)=(emin(t)+emax(t))/2,

emin(t) – lower signal envelope, *emax(t)* – upper signal envelope.

4. Difference between the signal and mean-value function provides the component with sifting, a local high-frequency component (detail function) d(t):

d(t) = x(t) − m(t),

x(t) – analyzed signal, *m(t)* – mean-value function.

5. Operations 1-4 are repeated using detail function *d(t)* as a reference signal *x(t)*. Together with increase of a number of iterations, the mean-value function *m(t)* goes to zero and function *d(t)* - to an unchangeable form to be taken for the most high-frequency mode function.
6. Function *imf (t)* determined using this method is deducted from the reference signal to result in balance r(t), containing low-frequency components:

r(t) = x(t)-imfi(t)

x(t) – analyzed signal, *imfi(t)* – internal mode function.

7. The next internal mode function is searched for using procedures described above with the difference that balance *r(t)* will be an input signal.

8. Algorithm of empirical mode decomposition stops when the balance, ideally, contains no extremums. It means that balance r(t) is a constant or a monotone function.

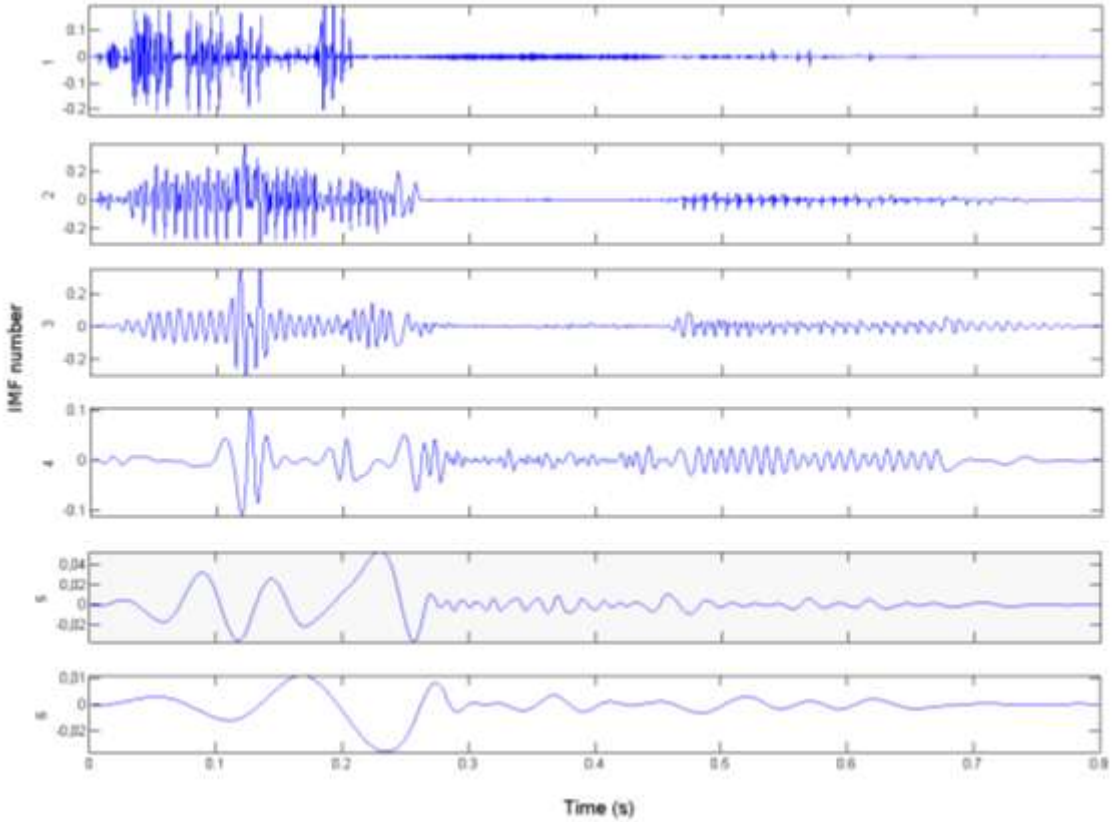

Figure 2:- The empirical modes of speech signal

Extracted internal mode functions possess unique local frequencies, whereby each of the components contain lower frequency components than the one extracted before. Figure 2 contains diagrams of 6 mode functions obtained using empirical decomposition of speech signal for a word signal.

Theory:-

Numerical studies held during our work have shown that parameters of internal mode functions and their number considerably change even within one phoneme (within a selection). For this reason it seems impossible to use readings of obtained mode functions as features for neural network training.

As an alternative to formation of descriptive phoneme features were considered values of energy calculated for every obtained empirical mode. In this case mode decomposition with subsequent calculation of mean energy value for every mode function was applied to each phoneme speech frame. Thus, a vector can be formed of calculated energy coefficients, size of which will be equal to number of calculated internal modes.

Yet this approach does not take into account a distinctive feature of the very EMD method, which lies in adaptive building of signal decomposition functions depending on its nature. Mode functions calculated at iterations with the same ordinal number can contain components differing in frequency composition for different signals, in spite of the fact that their vibrations will be more low-frequency than for a mode function obtained at previous iterations.

In order to form a vector of phoneme features, a specter of instantaneous Hilbert frequencies can be calculated for internal mode functions obtained by decomposition (Hilbert-Huang transform).

After performance of Hilbert transform for every internal mode function, analyzed signal *x(t)* can be presented as a substantial part of a complex form:

$$x(t) = RE\left[\sum_{j=1}^{n} a_j(t) \cdot \exp\left(i\int \omega_j(t)dt\right)\right]$$

This approach enables to obtain frequency-time representation of analyzed speech signal (figure 3).

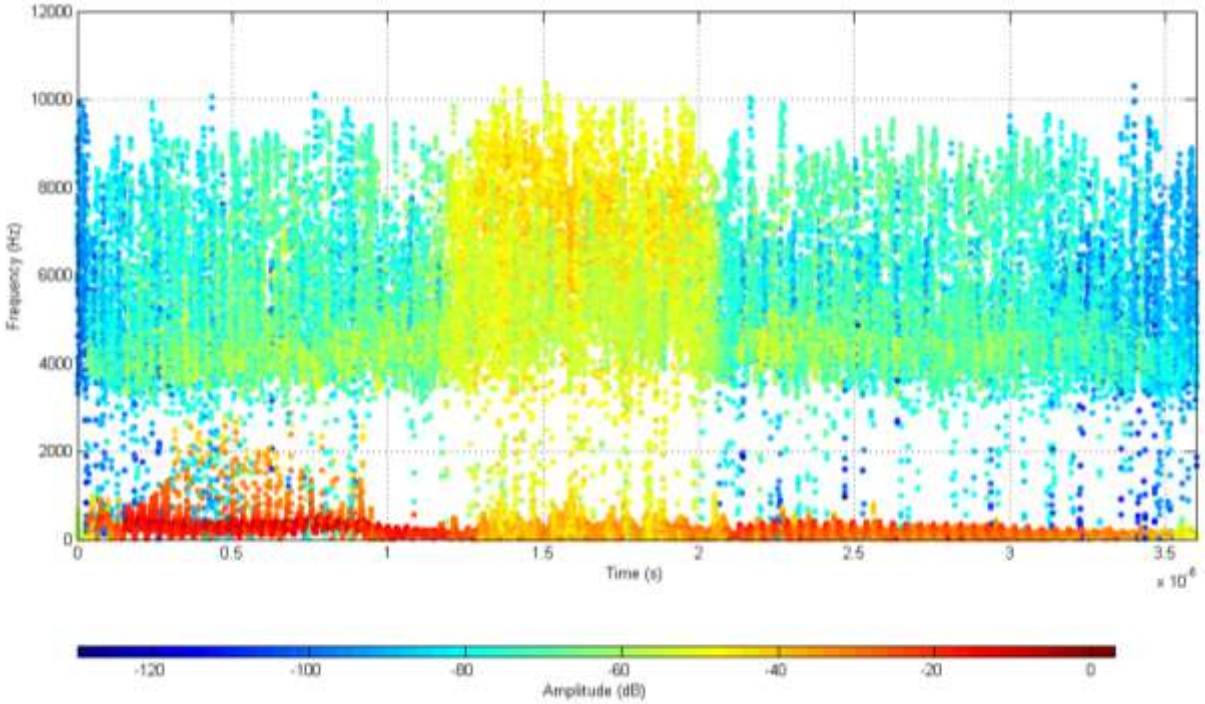

Figure 3:- Frequency-time representation of speech signal using Hilbert-Huang transform

Results:-

In spite of the fact that it is obvious that frequencies of definite spectrums prevail within separate phonemes, frequency features fluctuate in time. That is why to build a vector of features, we decided to consider summary specter of instantaneous frequencies throughout the whole duration of phoneme signal. Amplitude values in the interval of 20 frequency samples were averaged to reduce length of a vector with spectral features; this method is used to calculate a vector out of 60 values of frequency to be used in speech signal classification.

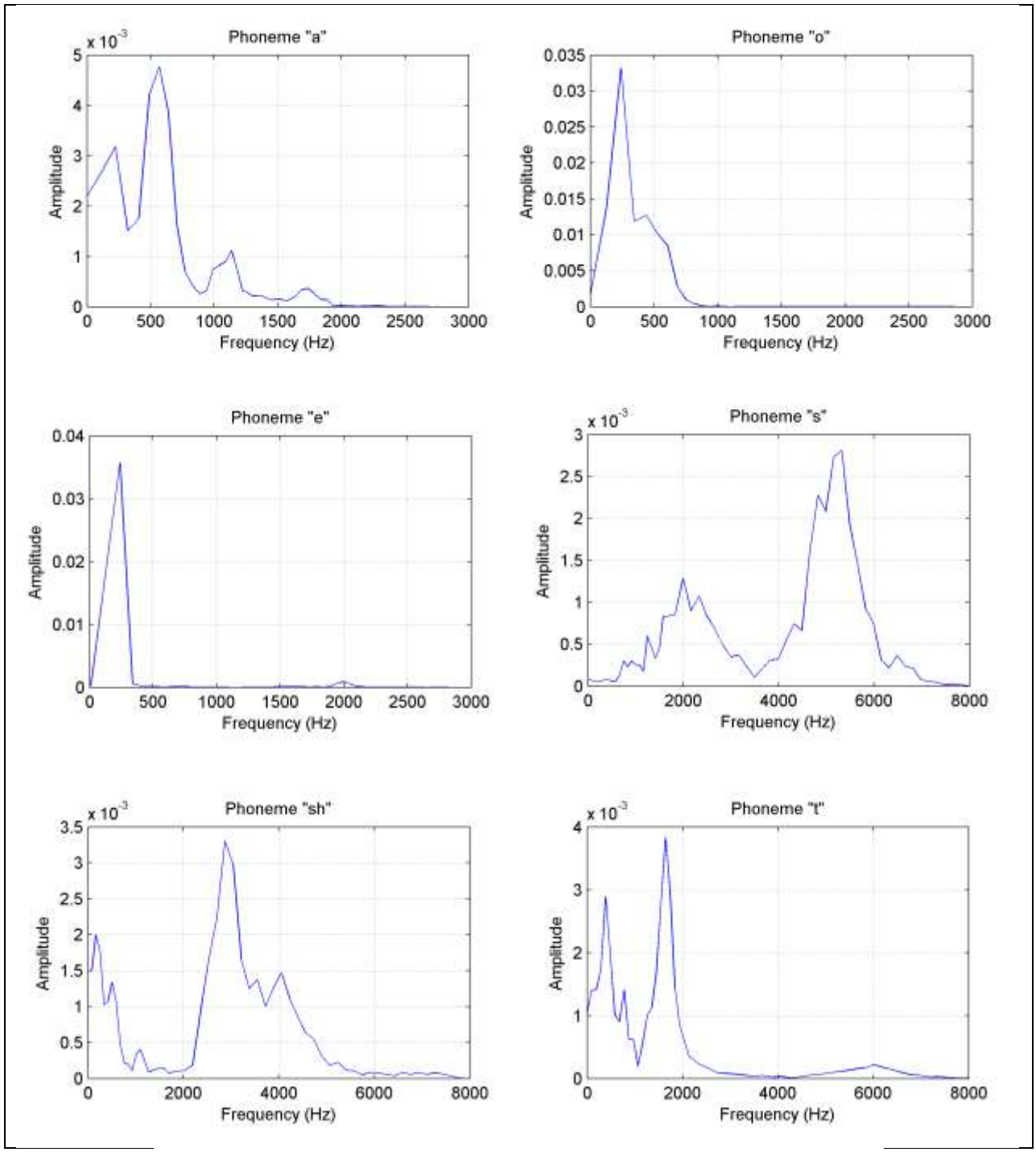

Figure 4:- Hilbert frequency analysis based on empirical modes of phoneme signals.

Number of internal mode functions used to build phoneme features was experimentally limited to 3, which enables singling out of frequency features required for the purposes of classification and reduces time necessary for their calculation, stopping EMD method at earlier iterations.

Based on results of held theoretical researches, we proposed and specified a phoneme model using a method built on Hilbert-Huang transform. Parameters of phoneme models were defined to be applied in a task of speech recognition.

Figure 4 contains diagrams of spectrums obtained based in Hilbert frequency analysis of empirical mode functions of Russian phoneme signals.

Based on presented algorithms, we developed a system of speech-to-text conversion in Matlab environment, enabling to conduct experimental research and optimize developed methods and algorithms (figure 5).

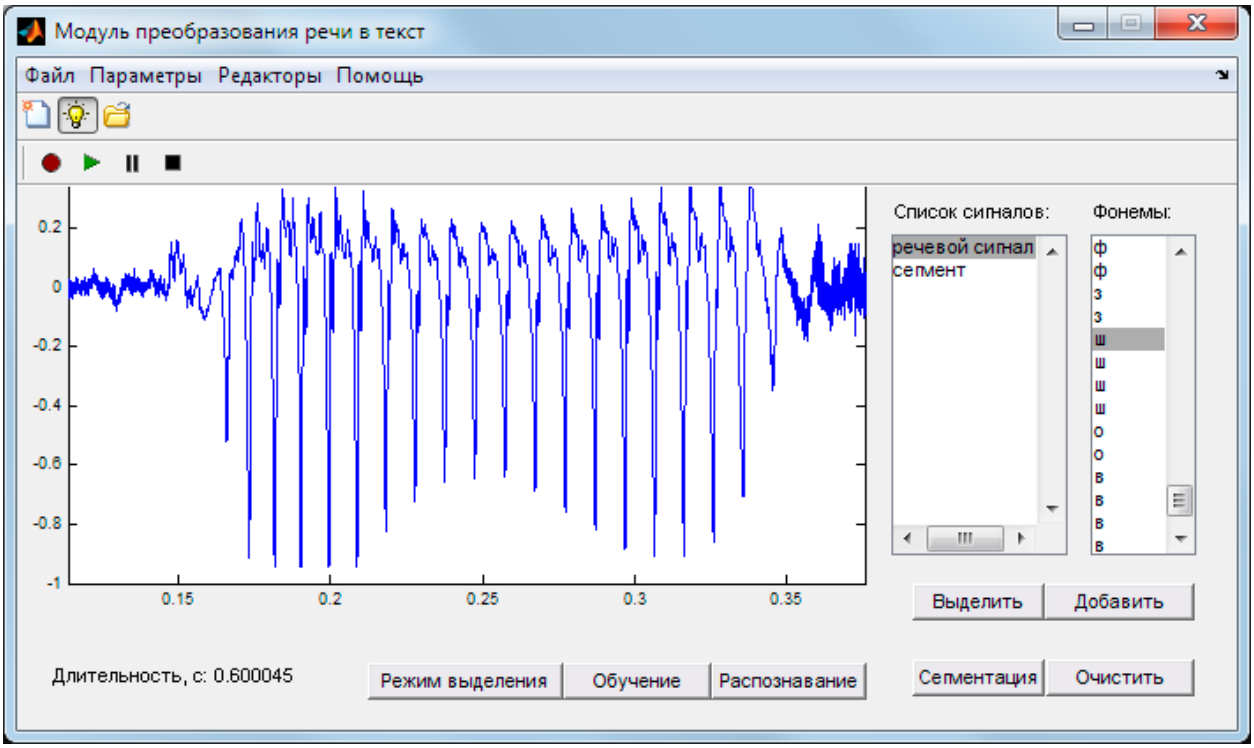

Figure 5:- User interface of the speech-to-text application

Conclusions:-

Comparative analysis of phoneme recognition quality was held using various approaches to speech features formation: using wavelet transform and method developed on the basis of Hilbert-Huang transform. Analysis of the results revealed advantages of the latter.

Experiments related to determining of quality of system work have shown that phoneme recognition coefficient amounted to 95.2%. Based on obtained results, we may speak about possibility of application of developed method in formation of phoneme speech features based on Hilbert-Huang transform in the task of speech recognition.

References:-

1. Davydov, A.V., 2005, Digital signal processing: thematic lections, Yekaterinburg: UGGU, Collection of Digital Documents.
2. Dremin, I.M., Ivanov, O.V., Nechitaylo, V.A. 2001, Wavelets and using of them, Advances in Physical Sciences, 171, 5, 465-500.
3. Medvedev, M.S., 2006, Using the wavelet transform in russian phoneme model construction, Bulletin of Krasnoyarsk State University, 193-201.
4. Qin, S.R., Zhong, Y.M., 2006, A new envelope algorithm of Hilbert–Huang Transform, Mechanical Systems and Signal Processing, 20, 1941–1952.
5. N.E., Huang, Samuel, S.P., Shen, 2014, The Hilbert-Huang transform and its applications, World Scientific Publishing Co. Pte. Ltd., 5 Toh Tuck. Link: Singapore.