Noise-Robust Speech Signals Processing for the Voice Control System Based on the Complementary Ensemble Empirical Mode Decomposition

Alimuradov Alan Kazanferovich
Penza State University
Penza, Russia
alansapfir@yandex.ru

Churakov Pyotr Pavlovich
Penza State University
Penza, Russia
churakov-pp@mail.ru

Abstract - **Noise-robust speech signals processing is one of the main problems of practical realization of voice control systems (VCS). The offered algorithm of noise-robust processing represents speech signals filtering (voice commands) with the use of the Complementary Ensemble Empirical Mode Decomposition (CEEMD) and the Independent Component Analysis (ICA) methods. A noisy speech signal is adaptively decomposed into frequency components - intrinsic mode functions (IMF) by means of the CEEMD method. The application of the CEEMD method for signals decomposition allows excluding mixing of IMF arising when processing signals containing short-term and disparate in scale areas. From the received set of IMF the mode is defined containing the main noise by means of an assessment of weight energy and noise IMF coefficients. Further the initial noisy speech signal and IMF with the main noise are exposed to processing by means of the ICA method. As a result the filtered speech signal is allocated. The application of the offered filtering algorithm contributes to the increase of VCS noise resistance and accuracy of voice commands recognition. The results of the offered algorithm researches show the effective noise suppression, including small values of signal-to-noise ratio (SNR).**

Keywords - noise-robust processing, speech signals filtering, voice control, Complementary Ensemble Empirical Mode Decomposition, Independent Component Analysis.

I. INTRODUCTION

At present the use of speech signals as the interface of voice control has received a wide popularity in operating systems. In practice all speech signals are noisy to some extent. In the conditions of a modern "aggressive" noise situation and depending on intensity, noise can significantly distort results of processing and analysis of speech signals. For this reason research and development of new highly effective methods of noise-robust speech signals processing, adaptive to modern conditions of noise pollution, are very actual.

The work in the area of noise-robust speech signals processing is conducted rather actively. Today a large number of various methods of noisy speech signals processing is developed:

- methods of adaptive compensation of hindrances;
- methods based on the use of speech signals mathematical models in time domain;
- methods based on the use of speech signals mathematical models in frequency domain;
- methods based on the use of noise spectral characteristics;
- methods based on the use of artificial neural networks models;
- methods based on a person’s speech perception models.

In the field of noisy speech signals processing for VCS methods based on the use of spectral characteristics [1, 2], processing in frequency [3] and time [4] domains are the most popular. The detailed analysis of the presented methods revealed that the problem of residual noise after filtering is not completely solved. The reason of it is impossibility of methods to analyze correctly nonlinear and non-stationary signals of a difficult form (useful and noise) of various intensity levels.

The carried-out review of known filtering algorithms of noisy speech signals using non-stationary data analysis tools [5, 6] revealed the following features. As a rule, in the presented works the classical Empirical Mode Decomposition (EMD) and the Ensemble Empirical Mode Decomposition (EEMD) methods are used [7, 8].

On the basis of the analysis of known algorithms and own researches [9, 10, 11] speech signals filtering algorithm for noise-robust VCS, based on the CEEMD [12] method is offered. Unlike the EMD and the EEMD methods the use of the CEEMD for decomposition allows excluding almost completely mixing of IMF (arising when processing signals containing short-term and disparate in scale areas). It is reached at the expense of a complementarity of the added white noise at decomposition stages. A noisy speech signal is adaptively decomposed into IMF from which a mode containing the main noise is defined by means of a special technique of an assessment of weighting energy and noise modes coefficients. Further, it is used the ICA method for an initial noisy speech signal and IMF with the main noise resulting in the allocation of the filtered speech signal with a minimum level of residual noise.

978-1-4799-7103-9/15/$31.00 ©2015 IEEE

II. Complementary Ensemble Empirical Mode Decomposition

The basis of the CEEMD is the classical EMD method. The mathematical apparatus of the EMD represents an iterative computing procedure of elimination [7]. A peculiarity of the EMD is that the basic functions used for decomposition in elimination procedure are taken directly from an initial signal. Decomposition into IMF allows analyzing short-term local changes in a signal therefore this method can be used for processing of nonlinear and non-stationary speech signals. The received IMF should meet the following requirements:

1) throughout a signal duration, a number of extrema (minimum and maximum) and zero crossings via time axis must be identical or differ no more per unit;

2) the mean value of the upper signal envelope and the lower signal envelope must be equal to zero at any time.

The result of the process of elimination [7] can be presented as follows:

$$s(t)=\sum_{i=1}^{I}c_i(t)+r_I(t) \qquad (1),$$

where $s(t)$ is the initial speech signal; $c_i(t)$ is an IMF; i is an IMF number; I is the amount of IMFs; $r_I(t)$ is a residue.

The shortcomings of the EMD are:

- mixing of several IMFs which are interpreted as one mode;
- receiving of IMFs consisting of signal areas of disparate scales which are in various parts of a signal.

For the elimination of these shortcomings it is used the CEEMD method based on repeated addition of white noise to a signal (with direct and inverse values) and calculation of IMF mean value as the end true result. The addition of white noise allows allocating disparate in amplitude and time signal areas which are in various parts of IMF for receiving all possible variations of modes in the course of elimination. Thus, the analyzed signal represents an association of a signal and white noise (with direct and inverse values) [12]:

$$\begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix}=\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}*\begin{bmatrix} s(t) \\ n(t) \end{bmatrix} \qquad (2),$$

where $s(t)$ is an initial speech signal; $n(t)$ is the added white noise; $s_1(t)$ is the sum of initial data with white noise; $s_2(t)$ is the sum of initial data with inverse white noise.

This approach fully uses the advantage of statistical characteristics of white noise for the detection of weak periodic (quasi-periodic) signals with the minimum value of residual noise.

Thus, the CEEMD method is a more exact way of speech signals analysis than the EMD and the EEMD by means of small in the amplitude white noise addition with direct and inverse values.

III. Speech signals filtering algorithm

The algorithm of speech signals filtering with the use of the CEEMD and the ICA methods is presented in Fig. 1. We will consider the main stages of the offered algorithm work in more detail, source data being as follows: $s(n)$ is a noisy speech signal, n is discrete timing $0<n\le N$, N is a number of discrete samples in a signal, $RMS=0{,}01$ is the level of noise amplitude, $j=1,2,...J$ is a decomposition cycle number, $J=50$ is the amount of decomposition cycles.

Block 1. According to [12] the CEEMD method is intended for receiving a set of IMF and a residue mean values:

$$IMF_i(n)=\frac{\sum_{j=1}^{J}IMF_{ij}(n)}{J} \qquad (3),$$

$$r_I(n)=\frac{\sum_{j=1}^{J}r_{Ij}(n)}{J} \qquad (4).$$

Block 2. After the decomposition completion a noisy speech signal represents a set of $IMF_i(n)$, where i is an IMF number, I is the amount of IMFs. Further the algorithm work is carried out with each IMF individually. For this purpose we expose a number of IMF equal to one ($i=1$).

Block 3. At the first stages of the algorithm work, it is necessary to install the initial thresholds of IMF weight coefficients (energy and noise) which are necessary for IMF with the main noise definition at the subsequent stages. Values of coefficients are established equal to one ($a_e=1$, $a_n=1$).

Block 4. IMF segmentation into fragments is carried out according to the following formulas:

$$S=\frac{IMF_1(n)}{L} \qquad (5),$$

where $IMF_1(n)$ is a signal of the first IMF1, S is the amount of IMF1 fragments, L is the amount of discrete samples in one fragment:

$$IMF_{1,s+1}(n)=IMF_1((s*L)+1;(s+1)*L) \qquad (6),$$

where $IMF_{1,s+1}(n)$ is IMF1 fragment, $s=(0,1,2,...S-1)$ is a fragment number.

Block. 5. According to the peculiarities of organs of articulation, a person does a short-term pause before a voice command pronunciation usually of 200 ms or more (1600 samples that is 20 fragments of 80 samples, with a sampling frequency of 8000 Hz). This area of a pause doesn't contain speech and corresponds to silence with a background noise.

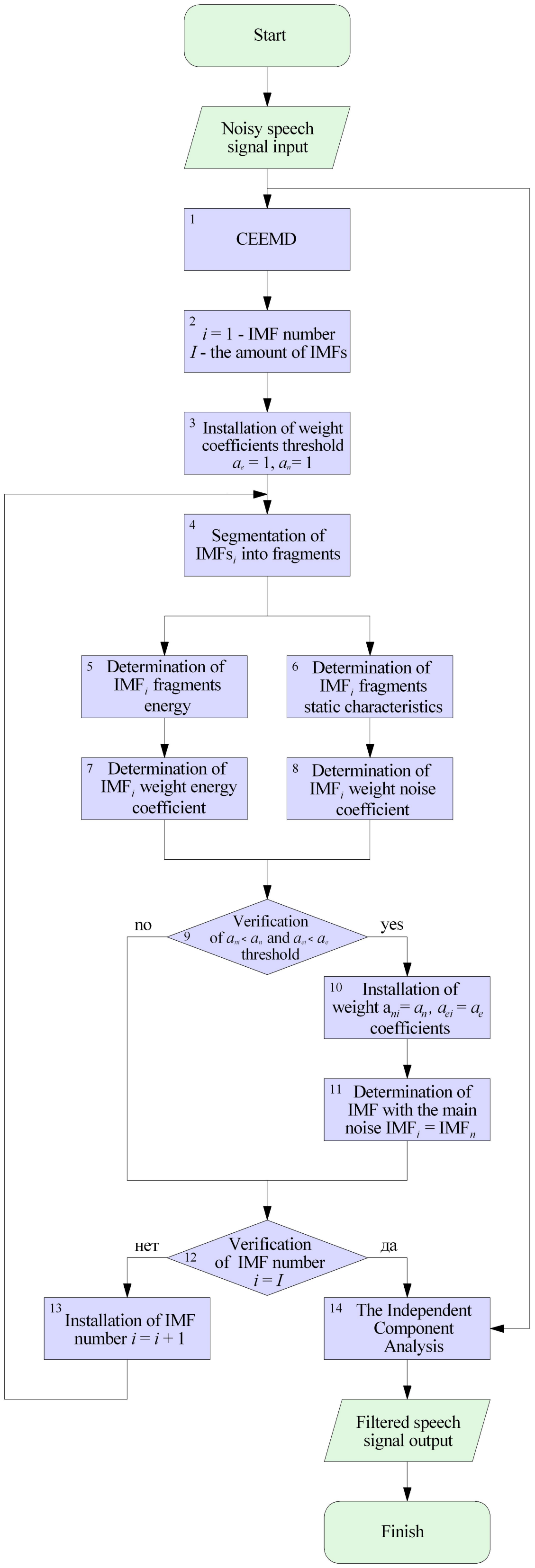

Start
Noisy speech signal input
1 CEEMD
2 $i = 1$ - IMF number
I - the amount of IMFs
3 Installation of weight coefficients threshold
$a_e = 1$, $a_n = 1$
4 Segmentation of IMFs$_i$ into fragments
5 Determination of IMF$_i$ fragments energy
6 Determination of IMF$_i$ fragments static characteristics
7 Determination of IMF$_i$ weight energy coefficient
8 Determination of IMF$_i$ weight noise coefficient
no
yes
9 Verification of $a_{ni} < a_n$ and $a_{ei} < a_e$ threshold
10 Installation of weight $a_{ni} = a_n$, $a_{ei} = a_e$ coefficients
11 Determination of IMF with the main noise IMF$_i$ = IMF$_n$
нет
да
12 Verification of IMF number $i = I$
13 Installation of IMF number $i = i + 1$
14 The Independent Component Analysis
Filtered speech signal output
Finish

Fig. 1. Speech signals filtering algorithm using the CEEMD and the ICA methods

The calculation of IMF and its fragments energy is carried out according to the formulas:

$$E_1 = \sum_{n=1}^{N} \left[IMF_1(n)\right]^2 \qquad (7),$$

$$e_{1,s+1} = \sum_{n=1}^{N} \left[IMF_{1,s+1}(n)\right]^2 \qquad (8).$$

Block 6. According to the requirement that during the first 200 ms a speech signal contains only background noise, it is possible to calculate static characteristics of noise: a mathematical deviation μ and a standard deviation σ of the first 1600 samples (20 fragments) by the formulas:

$$\mu = \frac{1}{1600} \sum_{n=1}^{1600} \left[IMF_1(n)\right]^2 \qquad (9),$$

$$\sigma = \sqrt{\frac{1}{1600} \sum_{n=1}^{1600} \left[IMF_1(n) - \mu\right]^2} \qquad (10).$$

Block 7. Determination of IMF with the main noise is carried out with the use of weight energy coefficient which is determined by the following formula:

$$a_{e1} = \frac{E_1 - e_1}{E_1} \qquad (11),$$

where $e_1 = \frac{\sum_{s=1}^{20} e_{1,s+1}}{20}$ is a mean value of energy of the first 20 fragments of IMF1 signal, E_1 is energy value of IMF1, a_{e1} is weight power coefficient of IMF1 status determination.

If the coefficient a_{e1} approaches to the absolute minimum value, the corresponding IMF1 is considered to be a mode with the main noise. And vice versa, if the coefficient a_{e1} approaches to one, the corresponding IMF1 is considered to be a mode containing a useful signal.

Block 8. Determination of IMF with the main noise is carried out with the use of weight noise coefficient. For each sample of IMF1 the one-dimensional Mahalanobis distance is calculated:

$$r_1(n) = \frac{\left|IMF_1(n) - \mu\right|}{\sigma} \qquad (12),$$

where $r_1(n)$ value allows to define if the sample is noise according to the requirement $r_1(n) > 3$. A value equal to 3 is received empirically. After the status of each sample is defined, it is necessary to define the status of each fragment.

The simple rule is used for this purpose that is if 60% of samples have the noise status in a fragment, such fragment is noise. This rule is used for the reason that the organs of articulation can't be reconstructed quickly and alternate speech and a pause during a half of a fragment (10 ms).

After the determination of fragments status the weight noise coefficient of IMF1 is defined:

$$a_{n1} = \frac{S - S_n}{S} \quad (13),$$

where S is the total amount of IMF1 fragments, S_n is the amount of IMF1 noise fragments.

If the coefficient a_{n1} approaches to the absolute minimum value, the corresponding IMF1 is considered to be a mode with the main noise. And vice versa if the coefficient a_{n1} approaches to one, the corresponding IMF1 is considered to be a mode containing a useful signal.

Blocks 9, 10, 11. After the determination of IMF1 weight coefficients the comparison with threshold values $a_{n1} < a_n$, $a_{e1} < a_e$ (block 9) is carried out. In case the requirement is satisfied, new values of threshold coefficients equal to the current values $a_n = a_{n1}$, $a_e = a_{e1}$ are established (block 10), and the analyzed IMF is determined as a mode with the main noise $IMF_i = IMF_n$ (block 11).

Blocks 12, 13. After the completing of IMF1 analysis the comparison of IMF number with their final amount ($i = I$) is carried out (block 12). In case the requirement isn't satisfied, the following IMF number $i = i + 1$ is established and the transition to block 4 for the analysis of the following IMF (block 13) is carried out.

Block 14. After the analysis of all IMFs and IMF with the main noise definition the direct filtering is carried out by means of the ICA method which led to the allocation of the filtered speech signal and noise components.

The Independent Component Analysis is the method of processing of statistical data allowing allocating the independent components possessing statistical independence and non-Gaussian distribution. The Independent Component Analysis is described by the following mathematical apparatus:

a set of observed vectors is X matrix, a vector of a noisy speech signal and IMF with the main noise which are linear combinations of independent components is Y matrix (in our case it is a pure speech signal and noise). The model of independent components can be expressed as follows:

$$X = W \cdot Y \quad (14),$$

where W is a matrix of scales for the transition from space Y to space X.

The purpose of Independent Component Method consists in the determination of W^{-1} matrix by means of which it will be possible to determine independent components Y matrix by the formula:

$$Y = W^{-1} \cdot X \quad (15).$$

IV. A PILOT STUDY OF SPEECH SIGNALS FILTERING ALGORITHM

The research was conducted on the base of software programs package MATLAB used for technical and mathematical tasks solution. Source data for the research are: the noisy speech signal representing the phrase "Henna Dripped on a Suit" lasting 1800 ms, registered without background noise, the frequency of sampling is 8000 Hz, word length of quantization is 16 bits. The CEEMD device settings are: the level of the root mean square deviation of the added white noise is 0,01, the number decomposition results is 50, the number of iterations is 50.

The phrase "Henna Dripped on a Suit" represents a nonlinear and non-stationary speech signal consisting of disparate in scale (amplitude and time) sounds [13] (Fig. 2).

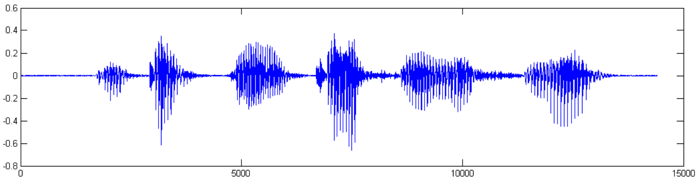


Fig. 2. A speech signal of the phrase "Henna Dripped on a Suit"

The criterion of the efficiency assessment of the offered filtering algorithm is the output value of SNR being the base of the residual noise existence estimation:

$$SNR_{OUTPUT}(dB) = 20\log(\frac{A_{OUTPUTsignal}}{A_{OUTPUTnoise}}) \quad (16),$$

where A_{signal}, A_{noise} is the root mean square value of a signal and noise amplitude ($A = \sqrt{\frac{1}{n}(a_1^2 + a_2^2 + ... + a_n^2)}$) respectively.

The results of the research will be estimated in comparison with known filtering methods program realization of which is available in open access:

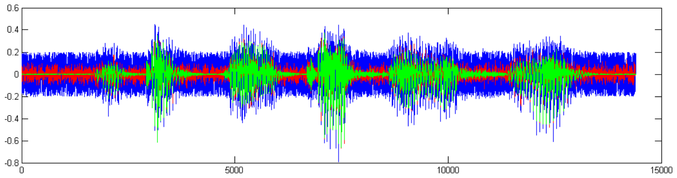
- the method based on discrete cosine transformation (Discrete Cosine Transform, DCT) with soft thresholding (SDCT);
- the method based on two-stage speech enhancement (Two-Stage Speech Enhancement, TSSE);
- the method based on hard and soft thresholding (Hard and Soft Thresholding, HST);
- the method based on the weighed noise subtraction and blind signal separation (Weighted Noise Subtraction and Blind Signal Separation, WNS+BSS).

The most characteristic noise making negative impact on the VCS functional capability is non-stationary background noise. It worsens the legibility of speech signals and can lead to a great difference between coming to the VCS noisy speech signals and standards received during the system training by pure speech signals. The great difference is the main reason of voice commands incorrect recognition. The white noise with various amplitude root mean square values was used for the algorithm pilot study. The generated white noise was imposed on a pure speech signal programmatically, forming noisy speech signals with various signal-to-noise ratios: from -5 to 15 dB.

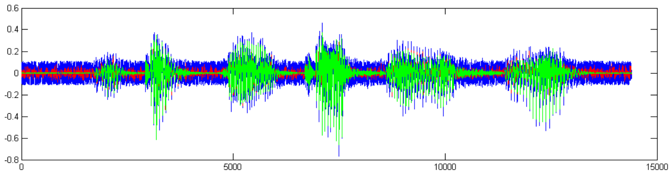
The comparison of filtering results by means of known methods and the offered algorithm are presented in the Table I. In Fig. 3 the results of the algorithm work are also presented in the form of pure, noisy and filtered speech signals oscillograms.

TABLE I. FILTERING RESULTS BY MEANS OF KNOWN METHODS AND THE OFFERED ALGORITHM

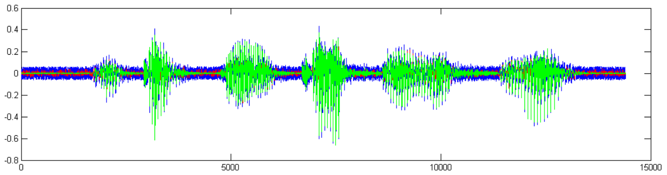
Input SNR, dB	**Output SNR, dB**				
	SDCT	*TSSE*	*HST*	*WNS+BSS*	*The offered algorithm*
- 5	8,14	7,58	8,25	8,37	9,23
0	10,56	10,34	11,23	11,08	12,9
5	14,27	13,41	14,94	14,23	16,24
10	15,8	16,1	17,7	16,5	18,12
15	23,4	26,56	26,24	21,45	28,76



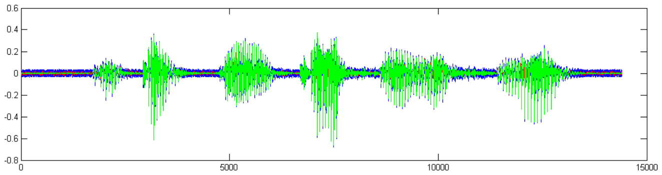
a. The result of a noisy speech signal filtering, output SNR is 9,93 dB at - 5 dB input value



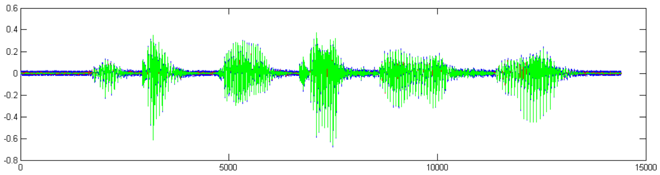
b. The result of a noisy speech signal filtering, output SNR is 12,9 dB at 0 dB input value



c. The result of a noisy speech signal filtering, output SNR is 16,24 dB at 5 dB input value



d. The result of a noisy speech signal filtering, output SNR is 18,14 dB at 10 dB input value



e. The result of a noisy speech signal filtering, output SNR is 28,76 dB at 15 dB input value

Fig. 3. The results of the filtering algorithm work: green color is an initial pure speech signal, blue color is a noisy speech signal, and red color is the filtered speech signal

According to the results presented in the Table I and in Fig. 3 it follows that the offered filtering algorithm is more effective than known methods for all range of values of input SNR which is reflected in Fig. 4.

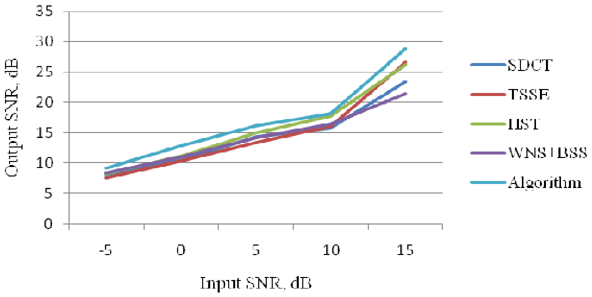


Fig. 4. Filtering results by means of known methods and the offered algorithm

The offered algorithm was investigated on speech signals contaminated with white noise. The further research of the offered algorithm on speech signals contaminated with brown and white noise is actual.

CONCLUSION

In this article the filtering algorithm of noise-robust speech signals processing for VCS on the basis of the CEEMD and the ICA methods is presented. A noisy speech signal is adaptively decomposed into a finite number of IMF. Further IMF with the main noise is determined. The ICA method is used for the allocation of a speech signal from a mixture of signal and noise. The results of a pilot study show that the offered filtering algorithm provides the minimum residual noise even at small input values of SNR (-5 dB and 0 dB).

REFERENCES

[1] Boll S. “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans Acoust Speech Signal Process. vol. 27 (2), 1979, pp. 113 - 120

[2] Berstein A., Shallom I. “A hypothesized Wiener filtering approach to noisy speech recognition”, Acoustics, Speech, and Signal Processing.

ICASSP-91, International Conference on. vol. 2, 1991, pp. 913 - 916, 14 - 17 April 1991.

[3] Furui S. “Cepstral analysis technique for automatic speaker verification”, IEEE Trans Acoust Speech Signal Process. vol. 29 (2), 1991, pp. 254 - 272.

[4] Hermansky H., Morgan N. “RASTA processing of speech”, IEEE Trans. Speech and Audio Proc., vol. 2 (4), pp. 578 - 589, October 1994.

[5] Kuo-Hau Wu, Chia-Ping Chen, Bing-Feng Yeh. “Noise-robust speech feature processing with empirical mode decomposition”, EURASIP Journal on Audio, Speech, and Music Processing, 2011, 9 p.

[6] Jingjiao Li, Dong An, Jiao Wang, Chaoqun Rong. “Speech Endpoint Detection in Noisy Environment Based on the Ensemble Empirical Mode Decomposition”, Advanced Engineering Forum. vol. 2 - 3, 2012, pp. 135 - 139.

[7] Huang N.E., Zheng Shen, Steven R.L. “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis”, Proceedings of the Royal Society of London A, vol. 454, 1998, pp. 903 - 995.

[8] Zhaohua Wu, Huang N.E., “Ensemble empirical mode decomposition: A noise - assisted data analysis method”, Advances in Adaptive Data Analysis, vol. 1 (1), 2009, pp. 1 - 41.

[9] Kuzmin A.V., Tychkov A.Yu, Alimuradov A.K. The development of effective noise biomedical signals processing method. International Journal of Applied Engineering Research. Volume 10, Number 3 (2015) pp. 8527 - 8531.

[10] Alimuradov A.K. “Speech signals filtering using the ensemble empirical mode decomposition method and the intrinsic mode functions energy assessment”, International Journal of Applied Engineering Research, Volume 10, Number 2 (2015) pp. 3175 - 3185.

[11] Tychkov A.Yu., Churakov P.P. “Processing photofluorographic images by means of decomposition into empirical modes”, Measurement Techniques, vol. 53 (10), 2011, pp.1125 - 1129.

[12] Yeh, J.-R., Shieh, J.-S., Huang N.E. “Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method”, Advances in Adaptive Data Analysis, vol. 2 (2), 2010, pp. 135 - 156.

[13] Alimuradov A.K., Churakov P.P., Tychkov A.Yu. “Choice of an optimum set of informative parameters of speech signals for voice control systems”, Measurement. Monitoring. Management. Control., vol. 1 (3), 2013, pp. 16 - 20.