UDC 004.934.1'1:519.767 **doi: 10.32620/reks.2023.3.09**

Oleksandr LAVRYNENKO, Denys BAKHTIIAROV, Vitalii KURUSHKIN, Serhii ZAVHORODNII, Veniamin ANTONOV, Petro STANKO

National Aviation University, Kyiv, Ukraine

A METHOD FOR EXTRACTING THE SEMANTIC FEATURES OF SPEECH SIGNAL RECOGNITION BASED ON EMPIRICAL WAVELET TRANSFORM

The subject *of this study is methods for improving the efficiency of semantic coding of speech signals.* ***The purpose*** *of this study is to develop a method for improving the efficiency of semantic coding of speech signals. Coding efficiency refers to the reduction of the information transmission rate with a given probability of error-free recognition of semantic features of speech signals, which will significantly reduce the required source bandwidth, thereby increasing the communication channel bandwidth. To achieve this goal, it is necessary to solve the following scientific* ***tasks:*** *(1) to investigate a known method for improving the efficiency of semantic coding of speech signals based on mel-frequency cepstral coefficients; (2) to substantiate the effectiveness of using the adaptive empirical wavelet transform in the tasks of multiple-scale analysis and semantic coding of speech signals; (3) to develop a method of semantic coding of speech signals based on adaptive empirical wavelet transform with further application of Hilbert spectral analysis and optimal thresholding; and (4) to perform an objective quantitative assessment of the increase in the efficiency of the developed method of semantic coding of speech signals in contrast to the existing method. The following scientific* ***results*** *were obtained during the study**:** a method of semantic coding of speech signals based on empirical wavelet transform is developed for the first time, which differs from existing methods by constructing a set of adaptive bandpass Meyer wavelet filters with further application of Hilbert spectral analysis to find the instantaneous amplitudes and frequencies of the functions of internal empirical modes, which will allow the identification of semantic features of speech signals and increase the efficiency of their coding; for the first time, it is proposed to use the method of adaptive empirical wavelet transform in the tasks of multiple-scale analysis and semantic coding of speech signals, which will increase the efficiency of spectral analysis by decomposing the high-frequency speech oscillation into its low-frequency components, namely internal empirical modes; the method of semantic coding of speech signals based on mel-frequency cepstral coefficients was further developed, but using the basic principles of adaptive spectral analysis with the help of empirical wavelet transform, which increases the efficiency of this method.* ***Conclusions:*** *We developed a method for semantic coding of speech signals based on empirical wavelet transform, which reduces the encoding rate from 320 to 192 bps and the required bandwidth from 40 to 24 Hz with a probability of error-free recognition of approximately 0.96 (96%) and a signal-to-noise ratio of 48 dB, according to which its efficiency is increased by 1.6 times as compared to the existing method. We developed an algorithm for semantic coding of speech signals based on empirical wavelet transform and its software implementation in the MATLAB R2022b programing language.*

Keywords: *semantic features of speech signals; mel-frequency cepstral coefficients; adaptive spectral analysis; empirical wavelet transform; adaptive wavelet-filters Meyer; functions of internal empirical modes; Hilbert spectral analysis; optimal threshold processing.*

Introduction

Today, the problem of semantic encoding of speech signals is gaining relevance because of the active development of technologies such as speech recognition and synthesis, voice control of technical objects, low-speed encoding of speech information, and voice translation from foreign languages. The functionality of systems that use such technologies depends on the efficiency of the encoding. Considering the growing trend of remote interaction between people and robotic equipment using these technologies, the main problem in telecommunication systems is to increase the bandwidth of the channel for transmitting semantic speech data. This is achieved through efficient coding of the data. Therefore, the key question is what speed is required for the semantic characteristics of speech signals to be encoded with a specific probability of error-free recognition. In this research, we will try to answer this question because it is an urgent scientific and technical task [1].

Linguistic communication begins when an abstract message appears in the speaker's brain. The process of speech production converts the message into an acoustic speech waveform. The information contained in this message is represented in the acoustic waveform in a complex manner. The initial message is first converted into sequences of nerve impulses that control the articulatory apparatus (i.e., the movements of the tongue, lips, vocal cords, etc.). Under the influence of these nerve impulses, the articulatory system is set in motion. This produces an acoustic speech waveform that conveys information about the original message [2].

The semantics of speech refers to the typed formant patterns of the spectrum of the speech signal, which correspond to a certain phoneme of the studied language. Since the semantic information of a speech signal is in the frequency domain, the study of spectral transformations is important for confirming the effectiveness of semantic speech coding and quantifying the semantic information contained in speech signals [3].

A message transmitted with a speech signal is discrete, i.e., it can be represented as a sequence of characters from a finite number of characters. The characters that make up a speech signal are called phonemes. The smallest elementary unit of language is considered a phoneme, which is defined as a distinct sound that serves to distinguish semantic units of language. Each language has its own set of phonemes N, usually from 30 to 50. For example, there are N=38 phonemes in Ukrainian. Thus, a speech message can be represented as a discrete sequence of phonemes. To give an objective quantitative assessment of a speech message, the concept of the amount of information in a speech message is introduced. Thus, we obtain a reasonable statement that if the phonemes of the corresponding language are chosen under the condition of equal probability of their occurrence, i.e., P = 1/N, where N is the number of all possible phonemes, then the average amount of information per phoneme of the Ukrainian language will be $H = \log_2 N = \log_2 38=$ $=5.2$ bit/phoneme. The value calculated in this way is called entropy, which determines the average amount of information per character of a discrete message. In this case, we deal exclusively with phonemes of speech messages. Physical restrictions on the movement of the elements of the articulatory apparatus allow a person to pronounce an average of 80-130 words per min or about 10 phonemes per s. Assuming an average speech rate of $W = n/\tau = 10/1 = 10$ phonemes/s, the transmission rate of phonemic information of a speech message in Ukrainian will be $C = W\times H = 10\times 5.2 = 52$ bit/s. In other words, at a normal rate of speech, the written equivalent of a speech message in various world languages is approximately 60 bit/s. This figure roughly characterizes the information content of a language that appears in its linguistic structure. The obtained value is in good agreement with the results of psychoacoustic experiments, which have established that a person is able to process information coming through the auditory channel at a speed of up to 50 bit/s. However, this assessment does not consider factors such as the speaker’s personality and emotional state, speed of speech, and voice volume [4].

However, if we turn directly to the acoustic characteristics of speech, the information picture will be different. The instantaneous spectrum of the speech signal covers a frequency band of approximately 300 to 3400 Hz, and the dynamic range of amplitudes is approximately 48 dB. Sound vibrations are characterized not only by amplitude-time and frequency-time parameters but also by phase relations. If all this is taken into account, then recording the full set of sound features contained in one word spoken in one second in the form used in computational mathematics requires several tens of thousands of binary characters. Thus, speech signals have a huge information redundancy [5]. Here is a reasoned justification for this statement. As you know, the phonetic and acoustic information of a speech message is in the frequency band from 300 Hz to 3400 Hz and with a dynamic range D of at least 48 dB. Hence, using the counting theory for continuous signals with finite spectra, we obtain the sampling rate $F_s = 2f_{max} = 2\times 3400 = 6800$ Hz. Considering the condition that $F_s \geq 2f_{max}$, in practice, the sampling rate often takes the value $F_s = 8000$ Hz. Then the total number of samples υ for a signal with a duration T will be $\upsilon = F_s T = 8000\times 1 = 8000$ samples, where $T = 1$ s, and the transmission rate of samples υ of a speech message with a duration of τ s will be $W_\upsilon = \upsilon/\tau = 8000/1 = 8000$ samples/s, where $\tau=1$ s, that is, this value obviously represents the dimensionality of the space corresponding to the signal base. Since the dynamic range of speech D should be at least 48 dB, the number of bits allocated for one sample of the speech message will be at least $k = 8$ bit/sample corresponding to the number of possible quantization levels $L_\upsilon = 256$ and dynamic range $D = 48.2$ dB. In this case, the rate of speech information transmission C_υ (acoustic characteristics of speech) will be $C_\upsilon = W_\upsilon \times k = 8000\times 8 = 64000$ bit/s [6].

Thus, the rate of transmission of phonemic information (linguistic semantics) of a speech message is $C = 60$ bit/s, and the information transmission rate of the acoustic characteristics of speech is $C_\upsilon = 64000$ bit/s, which shows redundancy by more than 1000 times. This proves that the semantic component of the speech signal is encoded in the acoustic oscillation in a very inefficient way, but extracting and recoding it by the optimal method is a rather non-trivial task, the solution of which is the focus of this research. The fact is that for semantic recognition, we do not need the acoustic characteristics of speech, i.e., we can recognize the phonemes of the speech message, which in turn will significantly reduce the redundancy of speech as well as the amount of data transmitted through the communication channel [7].

The problem of scientific research is that having the ability to determine and measure the amount of phonemic and acoustic information contained in speech signals according to the above material, today there is no final theoretical substantiation of the problem associated with the semantic coding of speech, namely, proving the possibility of quantitative measurement of the semantics hidden in the deep patterns of the speech signal. This is largely due to the fact that the speech signal is inherently a non-stationary and nonlinear process. Therefore, the study of such functions for deep semantic components (instantaneous frequencies and amplitudes) is problematic because the existing methods of semantic speech coding based on spectral analysis, such as the Fourier transform, wavelet transform, and cosine transform, use a priori basis functions at all iterations of the decomposition. This does not allow the optimality of coding to be proved under this condition because the error introduced by the basis itself will accumulate in the amplitude-frequency formant pattern characteristic of this spectral transform. The optimality of the semantic coding of speech signals and the determination of the quantitative measure of semantic information are possible only if the adaptability of the basis function to the studied signal is observed at each iteration of the spectral decomposition into a certain basis series with the subsequent determination of the instantaneous frequency and amplitude of the formant pattern of the speech signal spectrum [8].

The modern methods of semantic coding of speech signals do not adhere to the formulated statement; therefore, it was first proposed to use the method of adaptive empirical wavelet transform with subsequent Hilbert spectral analysis and optimal threshold processing to determine the semantic features of speech signals and their informational quantitative measurement. The developed method of semantic coding of speech signals based on empirical wavelet transform with further application of Hilbert spectral analysis and optimal thresholding fully complies with the conditions of adaptability, due to which the optimality of this method will be theoretically proved and the gain in terms of increasing the efficiency of semantic coding in contrast to existing methods will be obtained [9].

The purpose, tasks, object, subject, and methods of the research, scientific novelty, and practical significance of the results are described below.

1. Statement of the purpose of research

Purpose and tasks of the research. The purpose of this research is to develop a method for improving the efficiency of semantic coding of speech signals.

To achieve this goal, it is necessary to solve the following scientific **task:**

– to investigate a known method for improving the efficiency of semantic coding of speech signals based on mel-frequency cepstral coefficients;

– to substantiate the effectiveness of using the adaptive empirical wavelet transform in multiple-scale analysis and semantic coding of speech signals;

– develop a method for semantic coding of speech signals based on adaptive empirical wavelet transform with further application of Hilbert spectral analysis and optimal thresholding;

– to conduct an objective quantitative assessment of the increase in the efficiency of the developed method of semantic coding of speech signals in contrast to the existing method.

The object of this research is the processes of semantic coding of speech signals.

The subject of this research is methods for improving the efficiency of semantic coding of speech signals.

Research methods. The research is based on the following modern methods:

– spectral analysis (empirical wavelet transform, construction of adaptive Meyer wavelet filters, finding the function of internal empirical modes, cepstral analysis, Hilbert transform to find semantic features of speech signals);

– digital signal processing (Fourier spectrum segmentation, processing with a bank of triangular mel-frequency filters, logarithmization of Fourier spectrum energy, thresholding of wavelet coefficients to find semantic features of speech signals);

– theory of electrical communication (estimation of compression ratio, bit rate, signal-to-noise ratio and peak signal-to-noise ratio of semantic features of speech signals for quantitative measurement of coding efficiency);

– information and coding theory (estimation of the amount of information, source entropy, coding efficiency, redundancy factor, and coding speed of semantic features of speech signals for quantitative measurement of coding efficiency);

– probability theory and mathematical statistics (estimation of correlation coefficient, mathematical expectation, variance, root mean square error and the probability of error-free recognition of semantic features of speech signals for quantitative measurement of coding efficiency).

The scientific novelty of the obtained results is as follows:

– a method of semantic coding of speech signals based on empirical wavelet transform is developed for the first time, which differs from existing methods by constructing a set of adaptive bandpass Meyer wavelet filters with the subsequent application of Hilbert spectral analysis to find instantaneous amplitudes and frequencies of functions of internal empirical modes, which will allow the determination of the semantic features of speech signals and increase the efficiency of their coding;

– for the first time, we propose the use of the adaptive empirical wavelet transform method in the tasks of multiple-scale analysis and semantic coding of speech signals, which will increase the efficiency of spectral analysis by decomposing the high-frequency speech oscillation into its low-frequency components, namely, internal empirical modes;

– the method of semantic coding of speech signals based on mel-frequency cepstral coefficients was further developed using the basic principles of adaptive spectral analysis with the help of empirical wavelet transform, which increases the efficiency of this method.

The practical significance of the results obtained is as follows:

– a method of semantic coding of speech signals based on empirical wavelet transform is developed, which allows the reduction of the coding rate from 320 to 192 bit/s and the required bandwidth from 40 to 24 Hz with a probability of error-free recognition of approximately 0.96 (96%) and a signal-to-noise ratio of 48 dB, according to which its efficiency increases by 1.6 times in contrast to the existing method;

– an algorithm for semantic coding of speech signals based on empirical wavelet transform and its software implementation in the MATLAB R2022b programming language was developed.

The results obtained in this study can be used to build systems for remote interaction between people and robotic equipment using speech technologies, such as speech recognition and synthesis, voice control of technical objects, low-speed encoding of speech information, and voice translation from foreign language. The results of the research have been implemented in the scientific and technical activities of the Educational and Scientific-Production Complex "Information and Communication Systems" and the educational and scientific process of the Department of Telecommunication and Radio Electronic Systems of the Faculty of Aero Navigation, Electronics and Telecommunications of the National Aviation University, as confirmed by the relevant implementation acts.

Below, we mathematically formalize the above statements of scientific research in a specific comparison with the existing method of semantic coding of speech signals based on mel-frequency cepstral coefficients.

2. Problem statement

During writing this scientific article, we investigated a well-known method for improving the efficiency of semantic coding of speech signals based on mel-frequency cepstral coefficients [10-12], which involves finding the average values of the coefficients of the discrete cosine transform

$$c[n] = \sum_{m=0}^{N_f-1} E[m]\cos\left(\frac{\pi n\left(m+\frac{1}{2}\right)}{N_f}\right),$$

$$n = 0,\ \ldots,N_f - 1,$$

prologarithmized energy of the spectrum

$$E[m] = \ln\left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\right),$$

$$m = 0,\ \ldots,N_f - 1,$$

discrete Fourier transform

$$X[k] = \sum_{n=0}^{N-1} x[n]w[n]e^{\frac{-2\pi j}{N}kn},\quad k = 0,\ \ldots,N-1,$$

processed using a triangular filter

$$H_m[k] = \begin{cases} 0, & k < f[m-1]; \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])}, & f[m-1] \le k < f[m]; \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])}, & f[m] \le k \le f[m+1]; \\ 0, & k > f[m+1], \end{cases}$$

where,

$$f[m] = \left(\frac{N_f}{Fs}\right)M^{-1}\left(M(F_{min}) + m\frac{M(F_{max} - F_{min})}{N_f + 1}\right)$$

in mel-scale $M = 1127.01048\ln(1 + F/700)$.

The problem is that the presented method of semantic coding of speech signals based on mel-frequency cepstral coefficients does not meet the adaptability

$$\bigcup_{n=1}^{N}\Lambda_n=[0,\pi],$$

where $\Lambda_n=\left[\omega_{n-1},\omega_n\right]$ are the segments of the Fourier spectrum $\left[0,\pi\right]$ of the speech signal under study, which is divided into N contiguous segments with boundaries ω_n (where $\omega_0=0$ and $\omega_N=\pi$) [13-15].

Coding efficiency refers to the reduction of the information transmission rate with a given probability of error-free recognition of semantic features of speech signals, which significantly reduces the required source bandwidth, thereby increasing the communication channel bandwidth.

Let us devise the main scientific hypothesis of this research, which is that it is possible to increase the efficiency of semantic coding of speech signals using an adaptive empirical wavelet transform with the subsequent application of Hilbert spectral analysis and optimal thresholding.

3. Materials and methods of research

In this study, we propose the application of a modern method of empirical wavelet transform based on the construction of a family of adaptive wavelet functions to improve the efficiency of spectral analysis of speech signals and further semantic coding.

If we take the features of the Fourier frequency spectrum as the basis, then the task is equivalent to building a set of bandpass wavelet filters. One of the ways to achieve adaptability is to consider that compact wavelet filter media directly depends on where the semantic information we need is located in the speech signal spectrum, i.e., larger amplitudes of the Fourier spectrum carry more important information for function recovery, and hence for qualitative assessment of the semantic component of the speech signal, and small amplitudes are less important. Indeed, the properties of the internal empirical mode function stated by N. Huang [16] are equivalent to the statement that the spectrum of this function has a compact carrier and is centered around a certain frequency (depending on the signal). For the sake of clarity of the theoretical presentation of the essence of this method, we will consider only real periodic signals (their spectrum is symmetrical with respect to frequency $\omega=0$), and therefore easier to build an evidence base. However, the following considerations can be easily applied to speech signals, which we will do in the future by building different wavelet filters on positive and negative frequencies, respectively. During this study, we will consider the normalized Fourier spectrum, which has 2π periodicity, to comply with Shannon’s criteria and limit the frequency $\omega\in\left[0,\pi\right]$.

Let us start with the assumption that the Fourier frequency spectrum $\left[0,\pi\right]$ is divided into N adjacent segments (later we will discuss how to get such a division). Let's define ω_n as the boundaries between each segment (where $\omega_0=0$ and $\omega_N=\pi$), as shown in Fig. 1.

Each segment of the spectrum is designated as $\Lambda_n=\left[\omega_{n-1},\omega_n\right]$, where $\omega_n=\left(\Omega_n+\Omega_{n+1}\right)/2$, and Ω_n are local maxima in the frequency spectrum that characterize the semantic features of speech signals. Thus it is obvious that $\bigcup_{n=1}^{N}\Lambda_n=\left[0,\pi\right]$. At the center of each ω_n, we define the transition phase (blue rectangular areas in Fig. 1) T_n with a width of $2\tau_n$. Empirical wavelet functions are defined as bandpass filters for each spectrum segment Λ_n. To do this, we use an idea that is used in the construction of Littwood-Paley and Meyer wavelet functions. Then $\forall n>0$, we define the empirical scaling function and empirical wavelet functions using equations (1) and (2), respectively.

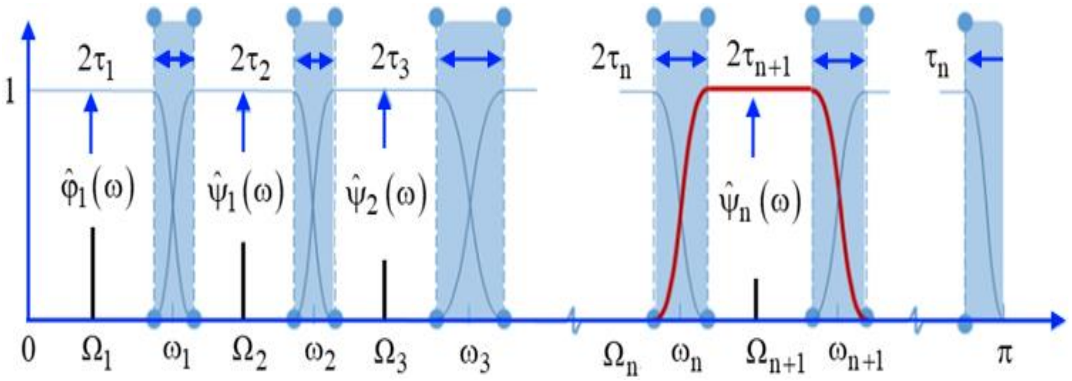


Fig. 1. Fourier spectrum division using adaptive low-pass $\hat{\varphi}_1(\omega)$ and bandpass $\hat{\psi}_n(\omega)$ Meyer filters

$$\hat{\varphi}_n(\omega)=\begin{cases}1, & \text{if } |\omega|\le\omega_n-\tau_n; \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}\left(|\omega|-\omega_n+\tau_n\right)\right)\right], & \\ \quad \text{if } \omega_n-\tau_n\le|\omega|\le\omega_n+\tau_n; & \\ 0, & \text{otherwise};\end{cases} \tag{1}$$

and

$$\hat{\psi}_n(\omega)=\begin{cases}1, & \text{if } \omega_n+\tau_n\le|\omega|\le\omega_{n+1}-\tau_{n+1}; \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_{n+1}}\left(|\omega|-\omega_{n+1}+\tau_{n+1}\right)\right)\right], & \\ \quad \text{if } \omega_{n+1}-\tau_{n+1}\le|\omega|\le\omega_{n+1}+\tau_{n+1}; & \\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}\left(|\omega|-\omega_n+\tau_n\right)\right)\right], & \\ \quad \text{if } \omega_n-\tau_n\le|\omega|\le\omega_n+\tau_n; & \\ 0, & \text{otherwise}.\end{cases} \tag{2}$$

Function $\beta(x)$ is an arbitrarily chosen $C^k([0,1])$ function is such that

$$\beta(x)=\begin{cases}0 & \text{if } x\le 0; \\ 1 & \text{if } x\ge 1;\end{cases} \text{ and } \beta(x)+\beta(1-x)=1.$$

$$\forall x\in[0,1].$$

Different functions follow these properties, the most commonly used in the literature is the following function

$$\beta(x)=x^4\left(35-84x+70x^2-20x^3\right).$$

As for the choice τ_n, then several options are possible. The easiest is to choose τ_n in proportion to ω_n: $\tau_n=\gamma\omega_n$ where $0<\gamma<1$. Thus, $\forall n>0$, equations (1) and (2) are simplified to equations (3) and (4), respectively

$$\hat{\varphi}_n(\omega)=\begin{cases}1, & \text{if } |\omega|\le(1-\gamma)\omega_n; \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}\left(|\omega|-(1-\gamma)\omega_n\right)\right)\right], & \\ \quad \text{if } (1-\gamma)\omega_n\le|\omega|\le(1+\gamma)\omega_n; & \\ 0, & \text{otherwise}\end{cases} \tag{3}$$

and

$$\hat{\psi}_n(\omega)=\begin{cases}1, & \text{if } (1+\gamma)\omega_n\le|\omega|\le(1-\gamma)\omega_{n+1}; \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_{n+1}}\left(|\omega|-(1-\gamma)\omega_{n+1}\right)\right)\right], & \\ \quad \text{if } (1-\gamma)\omega_{n+1}\le|\omega|\le(1+\gamma)\omega_{n+1}; & \\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}\left(|\omega|-(1-\gamma)\omega_n\right)\right)\right], & \\ \quad \text{if } (1-\gamma)\omega_n\le|\omega|\le(1+\gamma)\omega_n; & \\ 0, & \text{otherwise}.\end{cases} \tag{4}$$

An example of an empirical scaling function $\hat{\varphi}_n$ for $\nu_n=1$, $\gamma=0.5$ and the empirical wavelet function $\hat{\psi}_n$ for $\nu_n=1$, $\nu_{n+1}=2.5$, $\gamma=0.2$ in the frequency domain is shown in Fig. 2 [17].

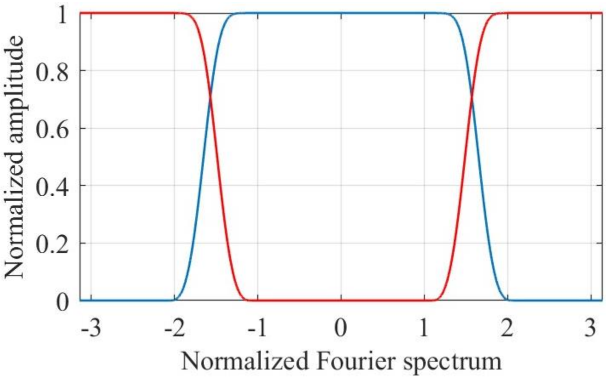


Fig. 2. Blue line: Fourier transform of the scaling function for $\nu_n=1$, $\gamma=0.5$. Red line: Fourier transform of the wavelet function for $\nu_n=1$, $\nu_{n+1}=2.5$, $\gamma=0.2$

Qualitative segmentation of the Fourier spectrum of the speech signal is a primary task in the process of semantic coding based on the empirical wavelet transform, since this step ensures the adaptability of the proposed method to the analyzed signal, which allows for a better study of the frequency spectrum. In view of the above, we aim to divide the Fourier spectrum into different segments that correspond to the functions of internal empirical modes and are centered around a certain frequency and a compact medium.

At this stage, we assume that the segment number N is known and set (below we will propose a method for estimating the optimal number of segments). This implies that only $N+1$ limits are needed, but we already have 0 and π, at our disposal, i.e. 2 limits, so we must find $N-1$ additional limits. To find such limits, you must first identify local maxima in the frequency spectrum and sort them in descending order (0 and π

are not taken into account). Let's assume that the algorithm has found M maxima.

Two cases may arise:

1) $M \geq N$: the algorithm has found enough maxima to determine the desired number of segments, then we keep only the first $N-1$ maxima;

2) M<N: the signal has fewer internal empirical modes than expected, then we keep the number of all detected maxima and reset N it to the appropriate value.

Now, having the set of found maxima, as well as 0 and π, we define the limits ω_n of each segment as the center between two consecutive maxima.

If it is possible to determine the optimal number of empirical modes N (frequency spectrum segments) for simple signals by experimentation, we usually deal with complex signals, such as speech signals, where a priori information about the modes of the studied signal is always unavailable. In such cases, it should be possible to automatically estimate the required number of mods. In general, this task is quite complex. Below, we present a simple method of assessment N. However, to ensure greater efficiency of the method, it is necessary to conduct an in-depth analysis of this issue [18].

The following statement shows that with an appropriate choice of parameter γ, a dense frame structure can be obtained.

Statement 1. If $\gamma < \min_n \left(\frac{\omega_{n+1}-\omega_n}{\omega_{n+1}+\omega_n}\right)$, then the set $\left\{\varphi_1(t), \{\psi_n(t)\}_{n=1}^{N}\right\}$ is a dense frame structure $L^2(\Re)$.

Proof. We stick to the idea of building a wavelet Meyer function.

Set $\left\{\varphi_1(t), \{\psi_n(t)\}_{n=1}^{N}\right\}$ is a dense frame structure if

$$\sum_{k=-\infty}^{+\infty}\left(\left|\hat{\varphi}_1(\omega+2k\pi)\right|^2+\sum_{n=1}^{N}\left|\hat{\psi}_n(\omega+2k\pi)\right|^2\right)=1.$$

According to periodicity 2π, it is sufficient to focus on interval [0,2π].

Following the previous definitions, we can write the following expression

$$[0,2\pi]=\bigcup_{n=1}^{N}\Lambda_n \cup \bigcup_{n=1}^{N}\Lambda_{\sigma(n)},$$

where $\Lambda_{\sigma(n)}$ is a copy of Λ_n but centered on $2\pi-\nu_n$ instead of ν_n. First, it is easy to see from expressions (5) and (6), i.e., that for

$$\omega \in \left(\frac{\bigcup_{n+1}^{N}\Lambda_n}{\bigcup_{n+1}^{N}T_n}\right)\cup\left(\frac{\bigcup_{n=1}^{N}\Lambda_{\sigma(n)}}{\bigcup_{n+1}^{N}T_{\sigma(n)}}\right), \tag{5}$$

we have

$$\left|\hat{\varphi}_1(\omega)\right|^2+\left|\hat{\varphi}_1(\omega-2\pi)\right|^2+ \\ +\sum_{n=1}^{N}\left(\left|\hat{\psi}_n(\omega)\right|^2+\left|\hat{\psi}_n(\omega-2\pi)\right|^2\right)=1. \tag{6}$$

Then, it remains to look at the transition areas. Because of the properties of β, this result also holds for Tₙ, if consecutive Tₙ do not overlap:

$$\tau_n+\tau_{n+1}<\omega_{n+1}-\omega_n, \\ \Leftrightarrow \gamma\omega_n+\gamma\omega_{n+1}<\omega_{n+1}-\omega_n, \\ \Leftrightarrow \gamma<\frac{\omega_{n+1}-\omega_n}{\omega_{n+1}+\omega_n}. \tag{7}$$

Condition (7) must be satisfied for all n, as well as for the smallest Tₙ, which is equivalent; therefore, we obtain the desired result if

$$\gamma<\min_n\left(\frac{\omega_{n+1}-\omega_n}{\omega_{n+1}+\omega_n}\right).$$

Fig. 3 shows an example of a bank of empirical wavelet filters based on the set $\omega_n \in \{0,\ 1.21,\ 2.02,\ 2.58,\ \pi\}$ of $\gamma = 0.05$ (according to theory $\gamma < 0.057$) [19].

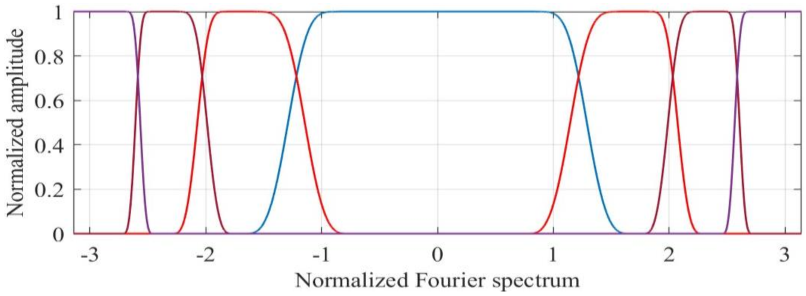


Fig. 3. An example of Fourier spectrum division by a bank of empirical wavelet filters

From the above explanation, we now know how to construct a set of frames of empirical wavelet functions of a dense structure. Now we can define the empirical wavelet transform (EWT), $W_f^{\varepsilon}(n,t)$, in the same way as for the classical wavelet transform. Then, the detailed coefficients are given by scalar products with empirical wavelet functions:

$$W_f^{\varepsilon}(n,t) = \langle f, \psi_n \rangle = \int f(\tau)\overline{\psi_n(\tau - t)}d\tau =$$
$$= \left(\hat{f}(\omega)\overline{\hat{\psi}_n(\omega)}\right)^{\vee},$$

and the approximation coefficients (denoted as follows are $W_f^{\varepsilon}(0,t)$) by the scalar product with a scaling function:

$$W_f^{\varepsilon}(0,t) = \langle f, \varphi_1 \rangle = \int f(\tau)\overline{\varphi_1(\tau - t)}d\tau =$$
$$= \left(\hat{f}(\omega)\overline{\hat{\varphi}_1(\omega)}\right)^{\vee},$$

where $\hat{\psi}_n(\omega)$ and $\hat{\varphi}_1(\omega)$ are defined by equations 11 and 10, respectively. The reconstruction (inverse EWT) of the original speech signal $f(t)$ by the wavelet coefficients of detail and approximation is given by the following expression

$$f(t) = W_f^{\varepsilon}(0,t)\cdot\varphi_1(t) + \sum_{n=1}^{N} W_f^{\varepsilon}(n,t)\cdot\psi_n(t) =$$
$$= \left(\hat{W}_f^{\varepsilon}(0,\omega)\cdot\hat{\varphi}_1(\omega) + \sum_{n=1}^{N}\hat{W}_f^{\varepsilon}(n,\omega)\cdot\hat{\psi}_n(\omega)\right)^{\vee}.$$

The above statements prove the effectiveness of using the empirical wavelet transform in the tasks of spectral analysis of speech signals, which will increase the efficiency of their semantic coding by maintaining adaptability to the studied signal. Next, we will proceed to the development of a method for semantic coding of speech signals based on an adaptive empirical wavelet transform with the subsequent application of Hilbert spectral analysis. According to the developed method (see Fig. 4), a speech signal is fed to its input, the frequency range of which is very limited and is located in the range from 300 to 3400 Hz. It follows from this fact that by modeling a bandpass filter, it is possible to discard frequency components that are outside this range and accordingly do not carry a semantic load.

As you know, speech signals are non-stationary signals of complex shape, the parameters and characteristics of which usually change rapidly over time. The established approach to speech signal processing uses short-term analysis.

In other words, the signal is divided into time frames of a fixed size, in which the signal parameters do not change.

To obtain a set of semantic features of the same length, the speech signal must be split into equal frames, and then the transform is performed, assuming that the signal in such a segment is approximately stationary (see Fig. 5).

For a speech signal, the frame size is usually selected within 10-20 ms. For a more accurate representation of the signal, an overlap equal to half the frame length is made between the frames.

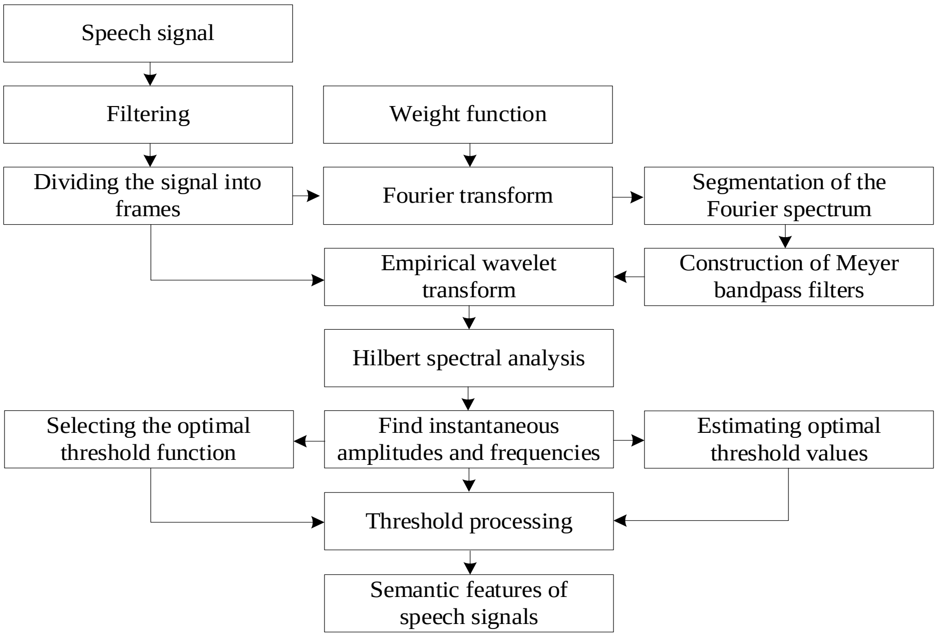

Speech signal
Filtering
Weight function
Dividing the signal into frames
Fourier transform
Segmentation of the Fourier spectrum
Empirical wavelet transform
Construction of Meyer bandpass filters
Hilbert spectral analysis
Selecting the optimal threshold function
Find instantaneous amplitudes and frequencies
Estimating optimal threshold values
Threshold processing
Semantic features of speech signals

Fig. 4. A method for semantic coding of speech signals based on empirical wavelet transform and Hilbert spectral analysis with subsequent optimal thresholding

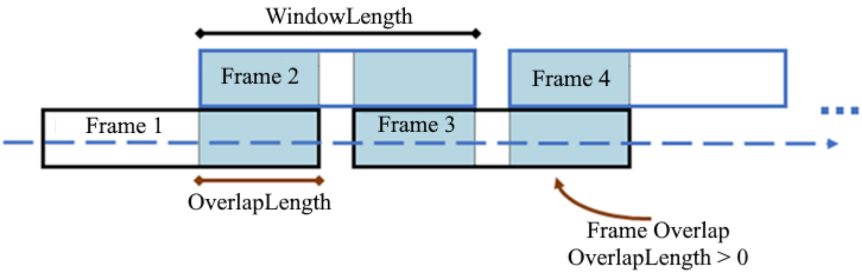


Fig. 5. Dividing the speech signal into frames

Frame overlap is used to prevent the loss of semantic information about the signal at the frame boundary. The smaller is the overlap, the smaller will be the dimensionality of the feature set characteristic of a given part of the speech signal. Then, a semantic component extraction algorithm is applied to each frame. Based on the above, the speech signal that has undergone pre-processing is divided into K frame of N samples, which intersect by 1/2 frame lengths. The input of the discrete Fourier transform unit is a sequence of samples of the speech signal section (K-st frame) studied at this iteration $x_0, \ldots, x_{N-1}$. A weight function is applied to this sequence, and a discrete Fourier transform is then applied. The weighting function is used to reduce distortions in the Fourier analysis caused by the finite sample size. In practice, the Hamming window is often used as a weighting function, which has the following form:

$$w[n] = 0.53836 - 0.46164 \cdot \cos\left(2\pi \frac{n}{N-1}\right),$$
$$n = 0, \ldots, N-1,$$

where, N is the length of the window expressed in samples.

The discrete Fourier transform of the weighted speech signal can then be written in the form of the following formula:

$$X[k] = \sum_{n=0}^{N-1} x[n] w[n] e^{\frac{-2\pi j}{N} kn}, \quad k = 0, \ldots, N-1.$$

Index values k correspond to the frequencies:

$$f[k] = \frac{F_s}{N} k, \quad k = 0, \ldots, N/2,$$

where, F_s is the sampling rate of the speech signal.

We adhere to the idea that the most important information for assessing the semantics of speech is stored in the maximum amplitudes (maxima) of the Fourier spectrum of the original signal (corresponding to the center of each of the N Fourier segments), which significantly exceed other existing maxima in the spectrum. Let us define the set M of the found maxima of the Fourier spectrum amplitudes by $\{M_i\}_{k=1}^{M}$. Suppose that this set is sorted in the descending order of values $(M_1 \geq M_2 \geq \ldots M_M)$ and normalized according to $[0;1]$. In this case, the above idea is equivalent to preserving all amplitude maxima of the frequency spectrum that exceed a certain value of the difference between the larger and smaller maxima. This can be formalized as follows: all amplitude maxima of the Fourier spectrum that are greater than a given threshold of $M_M + \alpha(M_1 - M_M)$, where α corresponds to the relative ratio of amplitudes, should be preserved. The task is to choose a value of α, that would lead to a compromise between the too frequent detection of so-called "false maxima" that do not carry important information and a qualitative division of the Fourier spectrum into segments that correspond to empirical modes of the speech signal. As a result, the threshold value directly affects the number of detected maxima and bands of Fourier spectrum segmentation and the number of modes into which the studied speech signal is decomposed. Following this formalism, the empirical mode f_k, defined by N. Huang, is the final sum of N+1 internal mode functions $f_k(t)$ with amplitude $F_k(t)$ and frequency $\phi_k(t)$ modulations, which can be written as follows

$$f_k(t) = F_k(t) \cos(\phi_k(t)),$$

where $F_k(t), \phi_k(t) > 0 \ \forall t$, and are such that

$$f(t) = \sum_{k=0}^{N} f_k(t),$$

is determined using the formulas

$$f_0(t) = W_f^{\varepsilon}(0,t) \cdot \varphi_1(t),$$
$$f_k(t) = W_f^{\varepsilon}(k,t) \cdot \psi_k(t).$$

As mentioned above, EWT is a tool for the time-frequency analysis of non-stationary and nonlinear signals, which are speech signals. One way to express the non-stationary nature of speech data is to determine the instantaneous frequency and amplitude of the signal under study. The Hilbert transform of signal $x(t)$ is given by the following expression

$$y(t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau,$$

where P is the principal Cauchy value of the singular integral.

Using the Hilbert transform of signal x(t) an analytical signal can be obtained

$$z(t) = x(t) + iy(t) = a(t)e^{i\theta(t)},$$

where $i = (-1)^{1/2}$.

Then a(t) can be expressed as

$$a(t) = \sqrt{\left(x^2 + y^2\right)}, \qquad (8)$$

where a(t) is the instantaneous amplitude.

The instantaneous phase function can be expressed as follows

$$\theta(t) = \arctan\frac{y}{x}.$$

The instantaneous frequency is determined by the expression

$$\omega(t) = \frac{d\theta}{dt}. \qquad (9)$$

By applying the Hilbert transform to the individual components of the internal empirical modes, the original data can be expressed as the following equation

$$x(t) = \mathrm{Re}\left\{\sum_{j=1}^{n} a_j(t)\exp\left[i\int \omega_j(t)dt\right]\right\}. \qquad (10)$$

Equation (10) defines the real part of the amplitude (8) and frequency (9) of each component of the internal empirical modes as a function of time. The analysis of signals in the time-frequency domain can be expressed as a Hilbert energy spectrum or a Hilbert amplitude spectrum, which are defined as the distribution of energy density and the distribution of amplitude density in time-frequency space, respectively [20].

The Hilbert energy density spectrum is defined as

$$S_{i,j} = H\left(t_i, \omega_j\right) = \frac{1}{\Delta t \cdot \Delta\omega} H\left[\sum_{k=1}^{n} a_k^2(t)\right].$$

The resolution of the Hilbert spectrum is given by intervals of equal size $\Delta t \cdot \Delta\omega$. Each interval represents a value of $a^2(t)$ at a given time and frequency. This transform has the property of energy compactness: more energy corresponds to less information.

At the next stage, thresholding of the Hilbert spectrum plays a crucial role in rejecting spectral coefficients that do not carry semantic information of the speech signal. Thus, we obtain a very small set of semantic features that, when encoded, successfully replaces thousands of samples of the speech signal that correspond exactly to the semantic form of the speech signal.

In practice, four threshold functions are widely used:

1) threshold function $T_H\left(\tilde{d}, \lambda\right)$ of the type:

$$T_H\left(\tilde{d}, \lambda\right) = \begin{cases} 0, & \left|\tilde{d}\right| \le \lambda; \\ \tilde{d}, & \left|\tilde{d}\right| > \lambda; \end{cases} \qquad (11)$$

2) threshold function $T_S\left(\tilde{d}, \lambda\right)$ of the type:

$$T_S\left(\tilde{d}, \lambda\right) = \begin{cases} 0, \ \left|\tilde{d}\right| \le \lambda; \\ \mathrm{sign}\left(\tilde{d}\right) \times \left[\left|\tilde{d}\right| - \lambda\right], \ \left|\tilde{d}\right| > \lambda; \end{cases} \qquad (12)$$

where λ is the threshold value, $\tilde{d}$ is the processed decomposition coefficient. The graphs of functions (11), (12) are shown in Fig. 6 for $\lambda = 0.5$ (1 is the graph of the linear function, 2 is the graph of function (11), 3 is the function (12)). Let us note the characteristic features of these functions:

1. By reducing the amplitude of the decomposition coefficient by the value of λ in function $T_S\left(\tilde{d}, \lambda\right)$ it is possible to smooth out the contrast elements of the processed signal, especially at large values of λ.

2. The presence of a gap in function $T_H\left(\tilde{d}, \lambda\right)$ in the environment of λ can cause oscillations (Gibbs effect) in the processed signal.

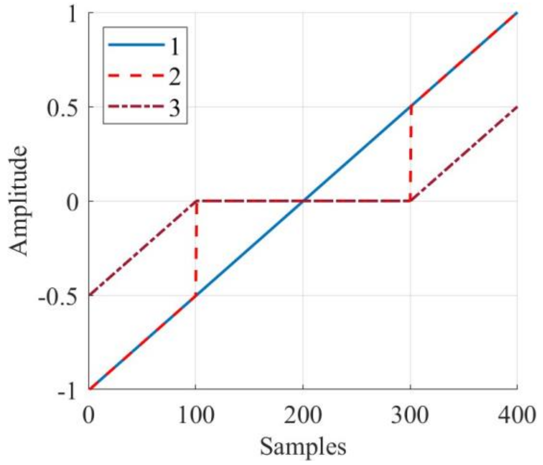


Fig. 6. Graphs of threshold functions (11), (12)

To overcome these shortcomings, two-parameter functions have been proposed, which will now be considered.

3. Threshold function $T_{SS}\left(\tilde{d},\lambda_1,\lambda_2\right)$ of the type:

$$T_{SS}\left(\tilde{d},\lambda_1,\lambda_2\right)=\begin{cases} 0, \left|\tilde{d}\right|\le\lambda_1; \\ \operatorname{sign}\left(\tilde{d}\right)\times\left[\dfrac{\lambda_2\left(\left|\tilde{d}\right|-\lambda_1\right)}{\lambda_2-\lambda_1}\right], \\ \lambda_1<\left|\tilde{d}\right|\le\lambda_2; \\ \tilde{d}, \left|\tilde{d}\right|>\lambda_2; \end{cases} \qquad (13)$$

which already includes two thresholds λ_1, λ_2. The graph of this function (at $\lambda_1=0.5$, $\lambda_2=0.75$) is shown in Fig. 7 (curve 2).

4. Threshold function $T_Z\left(\tilde{d},\lambda_1,\lambda_2\right)$, is defined by the expression:

$$T_Z\left(\tilde{d},\lambda_1,\lambda_2\right)=\begin{cases} 0, \left|\tilde{d}\right|\le\lambda_1; \\ \dfrac{\tilde{d}}{e-1}\times\left[e^{\frac{\left|\tilde{d}\right|-\lambda_1}{\lambda_2-\lambda_1}}-1\right], \\ \lambda_1<\left|\tilde{d}\right|\le\lambda_2; \\ \tilde{d}, \left|\tilde{d}\right|>\lambda_2; \end{cases} \qquad (14)$$

Fig. 7 shows the graphs of function (13) (curve 2) and function (14) (curve 3), constructed at $\lambda_1=0.5$, $\lambda_2=0.75$, as well as the graph of the linear function (curve 1). It can be seen that at interval $\left[\lambda_1,\lambda_2\right]$ the function (14) differs from the straight line (which is present in function (13)). This fact illustrates the reduction of the negative effect of oscillation (Gibbs effect) and smoother approximation within the threshold values, which makes function (14) the best in the threshold processing of speech signals. This statement is formalized below.

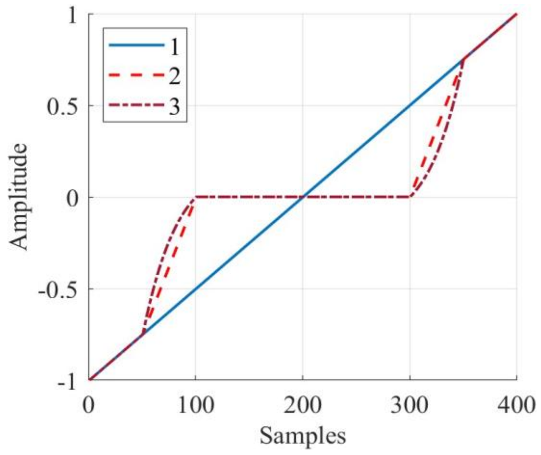


Fig. 7. Graphs of threshold functions (13), (14)

The optimal threshold function is selected according to the following algorithm [21].

The input data are formed as

$$\tilde{f}_i=f_i+\eta_i,\ i=1,2,\ldots,N,$$

where $\tilde{f}_i$ is the value of the speech signal function, η_i are pseudo-random numbers (noise) subject to a normal distribution with zero mean and variance σ^2. The variance value was set in terms of the relative noise level $\delta_\eta=\|\eta\|/\|f\|$, where $\|\eta\|$ and $\|f\|$ are the Euclidean norms of the corresponding vectors. The accuracy of wavelet filtering was determined by the relative error as follows:

$$\delta_f\left(T\right)=\frac{\left\|\hat{f}\left(T\right)-f\right\|}{\|f\|},$$

where vector $\hat{f}\left(T\right)$ is the result of wavelet filtering with a threshold function T. Obviously, $\delta_f\left(T\right)$ is a random variable and therefore a sample estimate of the mathematical expectation of this random variable was calculated from the sample:

$$\bar{\delta}_f\left(T\right)=\frac{1}{N_s}\times\sum_{l=1}^{N_s}\delta_f^{(l)}\left(T\right),$$

where N_s is the sample size, $\delta_f^{(l)}\left(T\right)$ is the relative filtering error of the l -th data realization $\tilde{f}^{(l)} = f + \eta^{(l)}$. Next, we find the minimum filtering error for each threshold function (11) – (14) by minimizing $\overline{\delta}_f(T)$ and relating these minimum errors to the minimum error of the threshold function (14). In practice, as an optimal two-parameter threshold function, we can accept function (14), which depends on two parameters λ_1, λ_2 and these parameters significantly affect the error of speech signal filtering. We select the optimal threshold values by evaluating parameters λ_1, λ_2, the threshold function (14), which allows us to find the optimal thresholds λ_{1opt}, λ_{2opt} based on the optimality criterion, with a minimum standard deviation, which is determined by the expression:

$$\Delta(\lambda_1,\lambda_2) = M\left[\left\|\hat{f}_{\lambda_1,\lambda_2} - f\right\|^2\right],$$

where $M[\,]$ is the operator of mathematical expectation on the density of noise distribution; f, $\hat{f}_{\lambda_1,\lambda_2}$ are vectors whose projections are equal to the "exact" and "smoothed" values of the signals (at the given threshold values λ_1, λ_2). We will show that the constructed algorithm allows us to accurately estimate the optimal value of threshold λ_{opt}, at which the standard deviation of filtering is minimal. Estimation of optimal threshold values λ_{1opt}, λ_{2opt} for the threshold function (14). Let us assume that: 1) instead of exact values f_i of the discrete speech signal, we have noisy values $\tilde{f}_i = f_i + \eta_i$, $i = 1,2,\ldots,N$, where noise η_i has zero mean $M[\eta_i] = 0$, variance σ^2 and values η_i, η_j are uncorrelated at $i \neq j$; 2) the basis functions $\{\varphi_{j,k}(t)\}$, $\{\psi_{j,k}(t)\}$ are orthonormalized, which corresponds to orthogonal wavelets (Meyer, Daubechies, Symlets, and Coiflets wavelets);

Then we define the disjoint vector $e_{\lambda_1,\lambda_2} = \tilde{f} - \hat{f}_{\lambda_1,\lambda_2}$ and introduce the following statistics:

$$\rho_W(\lambda_1,\lambda_2) = \frac{1}{\sigma^2}\left\langle e_{\lambda_1,\lambda_2}, \tilde{f}\right\rangle = = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(e_{\lambda_1,\lambda_2}\right)\times\tilde{f}_i. \quad (15)$$

As in the linear filtering algorithms, we take as estimates for λ_{1opt}, λ_{2opt}, take the values λ_{1W}, λ_{2W}, for which the statistics $\rho_W(\lambda_{1W},\lambda_{2W})$ satisfy the inequalities:

$$\vartheta_{m,\gamma/2} \le \rho_W(\lambda_{1W},\lambda_{2W}) \le \vartheta_{m,1-\gamma/2}, \quad (16)$$

where $\vartheta_{m,\gamma/2}$, $\vartheta_{m,1-\gamma/2}$ are quantiles, χ^2_m is a distribution with m degrees of freedom of levels $\gamma/2$, $1-\gamma/2$ respectively, γ is the probability of a first-order error when testing the statistical hypothesis about the optimality of the smoothing parameter (usually $\gamma = 0.05$), $m = N$ is the number of values of the filtered signal (projection of vector $\tilde{f}$) [22]. If the number of degrees of freedom $m > 30$ (in filtering tasks, this is always done), then χ^2_m is a distribution well approximated by a normal distribution with a mathematical expectation m and variance $2m$. Then, assuming the probability of a first-order error $\gamma = 0.05$, we obtain the following formulas for calculating the quantiles included in inequality (16)

$$\vartheta_{m,0.025} = m - 1.96\sqrt{2m},$$
$$\vartheta_{m,0.975} = m + 1.96\sqrt{2m}.$$

To calculate the scores λ_{1W}, λ_{2W}, threshold values λ_1, λ_2, of the function (14) is defined in the form:

$$\lambda_1(\beta) = \beta\sqrt{2\ln(N_j)},$$
$$\lambda_2(\beta, C) = \beta\times C\sqrt{2\ln(N_j)},$$

where N_j is the number of processed coefficients of the j-th level, and the multiplier $C > 1$ follows from the inequality $\lambda_2 > \lambda_1$ (see (14)). Note that the multiplier $\sqrt{2\ln(N_j)}$ makes the thresholds equidependent and ensures the asymptotic optimality of the thresholds in order at $N_j \to \infty$. Therefore, it is necessary to evaluate β_{opt}, C_{opt}, using the statistic (15), i.e., calculate the value of β_W, C_W, that satisfy inequalities:

$$\vartheta_{m,\gamma/2} \le \rho_W(\beta_W, C_W) \le \vartheta_{m,1-\gamma/2}. \quad (17)$$

Then the estimates λ_{1W}, λ_{2W} are defined by the following expressions:

$$\lambda_{1W} = \beta_W \sqrt{2\ln\left(N_j\right)},$$
$$\lambda_{2W} = \beta_W \times C_W \sqrt{2\ln\left(N_j\right)}.$$

To calculate β_W, C_W instead of solving the nonlinear equation $\rho_W(\beta, C) = m$, which includes two unknown quantities β, C, consider the problem of minimizing the functionality $F(\beta, C) = \left|\rho_W(\beta, C) - m\right|^2$. Note that a solution to this problem always exists and well-known minimization procedures can be used to find it. As β_W, C_W the following element is accepted $\left\{\beta^{(n)}, C^{(n)}\right\}$ minimizing sequences for which inequality (17) holds. It can be shown that when using orthogonal wavelets, the criterion $\rho_W(\beta, C)$ is calculated using the coefficients of wavelet decomposition:

$$\rho_W(\beta, C) = \frac{1}{\sigma^2} \sum_{j=j_0+1}^{j_0+J} \sum_k \tilde{d}_{j,k} \times \times\left(\tilde{d}_{j,k} - T\left(\tilde{d}_{j,k}, \lambda_1(\beta), \lambda_2(\beta, C)\right)\right). \quad (18)$$

This allows you to find the value of the criterion (when implementing the minimization procedure) in the space of wavelet decomposition coefficients, and then (with the found β_W, C_W and computed $\hat{d}_{j,k}$) perform the inverse wavelet transform only once and obtain smoothed function values. Let us note some properties $\rho_W(\beta, C)$, which are obtained from (18):

1) all components included in formula (18) are non-negative (can vary from 0 to $\tilde{d}_{j,k}^2$) and therefore

$$\rho_W(\beta, C) \geq 0;$$

2) at $\beta \to 0$ and $C < \infty$ fair border

$$\rho_W(\beta, C) \to 0;$$

3) at $\beta \to \infty$ and $C < \infty$ fair border

$$\rho_W(\beta, C) \to \frac{1}{\sigma^2} \sum_{j=j_0+1}^{j_0+J} \sum_k \tilde{d}_{j,k}^2 = \frac{1}{\sigma^2} \left\|\tilde{f}\right\|^2.$$

The latter equality holds for orthogonal wavelets with appropriate normalization of the basis functions. The last two properties lead to the following statement.

Statement 2. If the inequality

$$\rho_W(\infty, C) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \tilde{f}^2 > \vartheta_{m,1-\gamma/2},$$

then there are finite values β_W, C_W for which inequality (17) holds. Failure to fulfill condition (17) means that the value of $\tilde{f}_i = \eta_i$, i.e. $f_i \equiv 0$. In this case $\beta_W = \infty$ and the smoothed values are equal to 0.

The essential feature of the above algorithm for calculating θ_W is the use of noise variance σ^2. In practice, as a rule, this value is unknown, and in this case, an estimate for the standard deviation can be used σ:

$$\hat{\sigma} = \frac{\operatorname{median}\left(\left|\tilde{d}_{1,k}\right|\right)}{0.6745}, \quad (19)$$

where operator $\operatorname{median}\left(\left|\tilde{d}_{1,k}\right|\right)$ calculates the median of the absolute values of the detailing coefficients of the decomposition level $j_0 + 1$ (the sample size is equal to $N/2$). This estimate is widely used in robust regression analysis algorithms. With respect to wavelet filtering algorithms, this estimate was studied [23], where acceptable accuracy was shown, namely, for a given variance $\sigma^2 = 0.91 \times 10^{-1}$ are the values of the estimate (19) calculated from 30 realizations of length $N/2 = 1024$ were in the range of $\left[0.88 \times 10^{-1}, 0.97 \times 10^{-1}\right]$.

4. Research results

In this work, the developed method of semantic coding of speech signals based on EWT was investigated and modeled in the MATLAB software package. In particular particular, the compression ratio (CR), bit rate (BR), correlation coefficient (CC), signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR) and root mean square error (RMSE) were evaluated, as well as the probability of error-free recognition of semantic features, which are the main indicators of the effectiveness of the proposed method. The formalization of performance indicators for semantic coding of speech signals is presented below. Let there be two vectors of semantic features of a speech signal $x = \left(x_1 \ \dots \ x_L\right)$, $y = \left(y_1 \ \dots \ y_L\right)$ by length L samples, then the Pearson correlation coefficient (CC) is calculated according to the following expression

$$CC = \frac{1}{L}\sum_{i=1}^{L}\left(\frac{x_i - M_x}{S_x}\right)\left(\frac{y_i - M_y}{S_y}\right),$$

where $M_x = \frac{1}{L}\sum_{i=1}^{L} x_i$, $M_y = \frac{1}{L}\sum_{i=1}^{L} y_i$ are the mathematical expectations of the vectors x and y, $S_x = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(x_i - M_x)^2}$, $S_y = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(y_i - M_y)^2}$ are standard deviations of vectors x and y. Using the Pearson correlation coefficient, we can determine the strength of the linear relationship between two vectors of values x and y, that is, if $|CC| = 1$ there is a functional linear relationship, and if $CC = 0$ there is no linear dependence. In cases where the calculated value of the correlation coefficient lies in accordance with the condition $0 \le |CC| \le 1$, then with an acceptable error, the correlation coefficient can be qualitatively assessed in accordance with Table 1.

Table 1
Qualitative relationship of the correlation coefficient

Quantitative measure of closeness connection, $\lvert CC \rvert$	Quality characteristic bonding forces
0 – 0.1	None
0.1 – 0.3	Weak
0.3 – 0.5	Moderate
0.5 – 0.7	Noticeable
0.7 – 0.9	High
0.9 – 0.99	Very high
0.99 – 1	Functional

The root mean square error (RMSE) is calculated by the formula

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(x_i - y_i)^2}.$$

In this case, we are interested in the smallest error with the highest geometric similarity between the compared semantic features of the speech signal, i.e. $RMSE \to 0$.

The signal-to-noise ratio (SNR) and peak-to-peak signal-to-noise ratio (PSNR) were calculated according to the following formulas

$$SNR = 10\log_{10}\left(\frac{\frac{1}{L}\sum_{i=1}^{L} x_i^2}{\frac{1}{L}\sum_{i=1}^{L}(x_i - y_i)^2}\right) \text{ [dB]},$$

$$PSNR = 10\log_{10}\left(\frac{\max\left(x_i^2\right)}{\frac{1}{L}\sum_{i=1}^{L}(x_i - y_i)^2}\right) \text{ [dB]}.$$

Obviously, the greater the geometric proximity between the compared semantic features of the speech signal, the greater the SNR and PSNR, respectively, i.e. $SNR(PSNR) \to \infty$, otherwise $SNR(PSNR) \to 0$.

The compression ratio (CR) of speech data characterizes the efficiency of the semantic coding algorithm and is calculated according to the expression

$$CR = \frac{S_o}{S_c},$$

where S_o is the amount of input speech data, S_c is the amount of semantically encoded speech data. Thus, the higher is the compression ratio, the more efficient the algorithm is. It should be noted that if $CR = 1$, then the algorithm does not perform compression, i.e., the output message is equal in volume to the input message.

Bit rate (BR)

$$BR = \upsilon \times \log_2 L_\upsilon \text{ [bit/s]},$$

where υ [samples/s] is the rate of transmission of speech signal samples υ per 1 second; L_υ is the total number of quantization levels of the speech signal samples.

The probabilities of error-free recognition of semantic features are calculated according to the following statements. Suppose that the recognition probability P of frequencies and amplitudes of the harmonic distribution function

$$x(t) = A \times \sin(\omega t + \varphi),$$

is equal to 1, and the functions of the uniform distribution law

$$x(t) = \begin{cases} \frac{1}{b-a}, & x \in [a,b]; \\ 0, & x \notin [a,b]; \end{cases}$$

is equal to 0.5, which is equivalent to the complete absence of semantic features in the studied speech signal. According to this statement, the actual probability of recognizing the semantic features of speech signals will be in the range from 0.5 to 1. The theoretical criterion for finding the maximum possible probability of recognizing the semantic features of the analyzed frame is written as follows, which is based on the balance between the energy of semantic features (probability distribution of the occurrence of samples of the studied speech signal) of the speech and their number

$$P = \frac{\sqrt{\sum_{k=1}^{N} |C_{i\ldots N}|^2}}{\sqrt{\sum_{k=1}^{N} |C|^2}},\ i = 1\ldots N,$$

where C is the Hilbert energy spectrum, which characterizes the probability distribution of the occurrence of samples of the speech signal of length N. It is obvious that the greater the geometric proximity between the compared semantic features of a speech signal, the more P, i.e. $P \to 1$, otherwise $P \to 0.5$. The percentage representation of the probability of recognizing the semantic features of speech signals can then be written as follows

$$P_{\%} = P \cdot 100\ (\%),$$

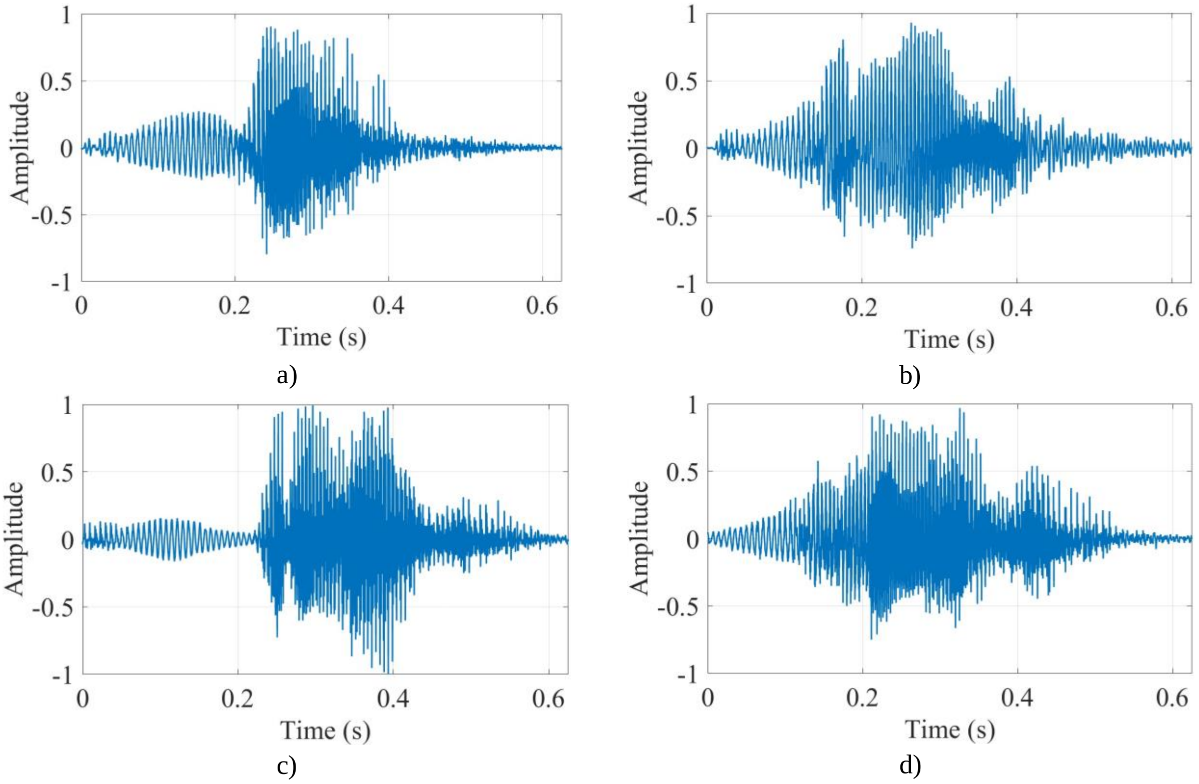
which is an absolute indicator of the semantic recognition of speech signals, which takes into account the internal probability distribution of the source of the process under study. The input digital speech signals for semantic coding are recordings of male and female voices with a sampling rate of 8 kHz and a quantization bit depth of 8 bits, which corresponds to the main digital channel of the telephone network is 64 kbit/s (Fig. 8).

Then, the adaptive basis is set by the scaling function and wavelet functions corresponding to the low-pass filter and Meyer bandpass filters for each spectrum segment. Let's build the amplitude spectra of the signal under study, where the location of the spectral peaks determines the frequency bands of the filter bank (Fig. 9).

Let us build the internal empirical modes of the studied signal using the empirical wavelet transform (Fig. 10).

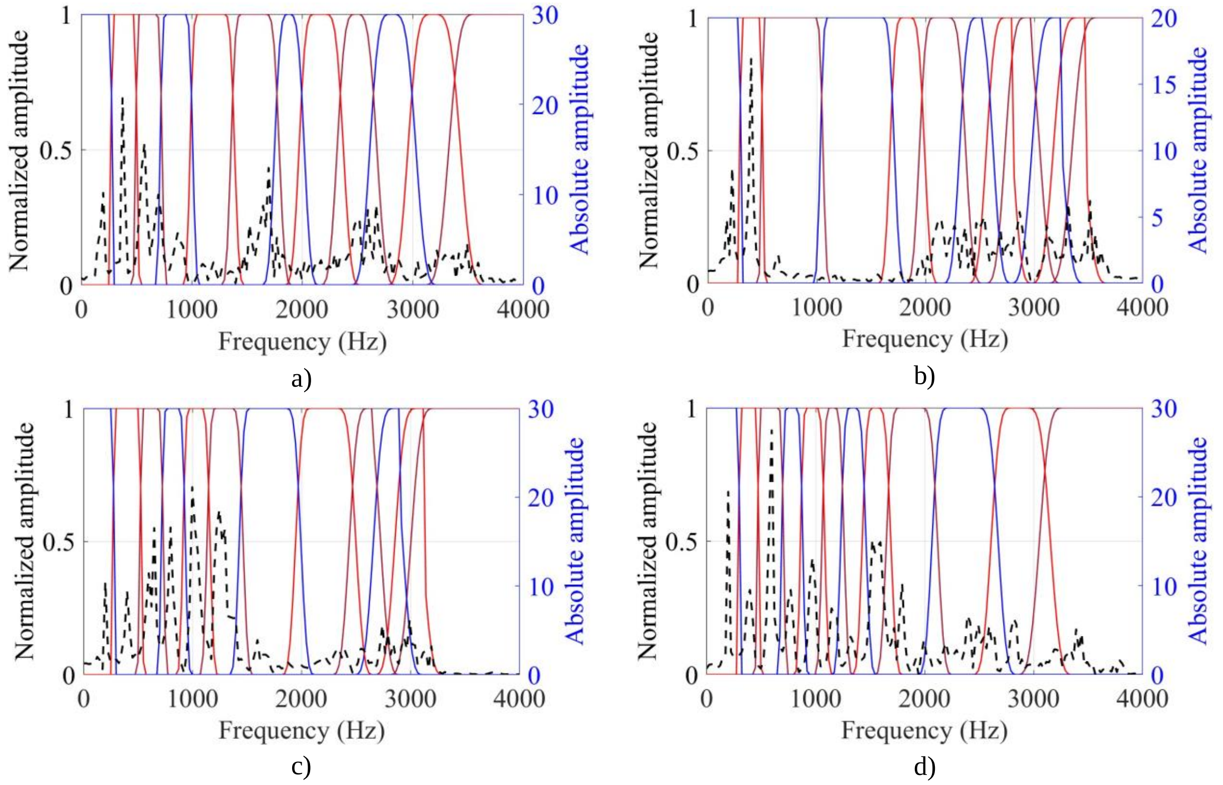
By applying the Hilbert transform to the mode functions of the empirical wavelet transform obtained because of the decomposition of the speech signal, we obtain the Hilbert energy spectrum, which depends on the instantaneous frequency and time (Fig. 11). The integral of this value over time gives the Hilbert integral spectrum, which is an analog of the Fourier spectrum.

The Hilbert transform and the empirical wavelet transform open up new possibilities for the analysis of speech signals in the detailed analysis of the frequency and time structure of their spectrum, namely the use of thresholding methods, as discussed above.



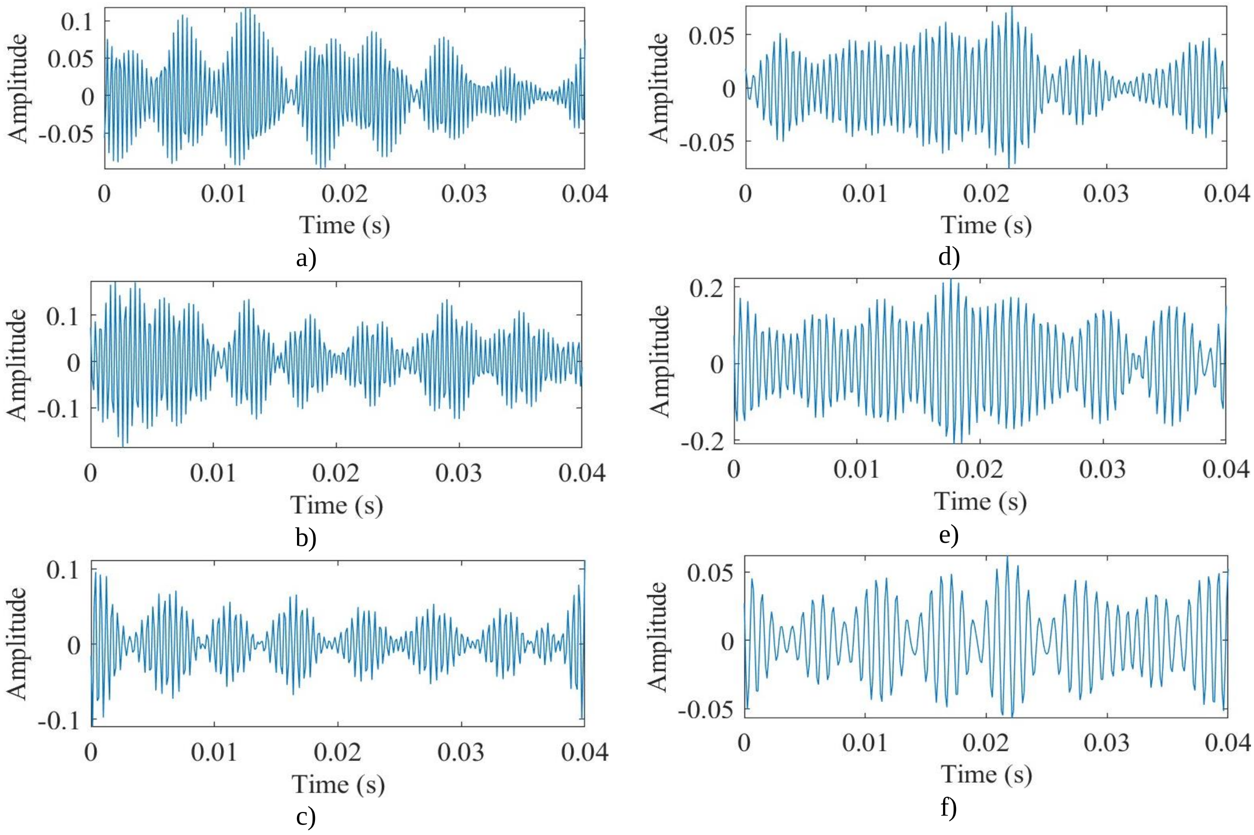
a) b) c) d)

Fig. 8. An example of voice commands: "Up" a), "Down" b), "Right" c), "Left" d)



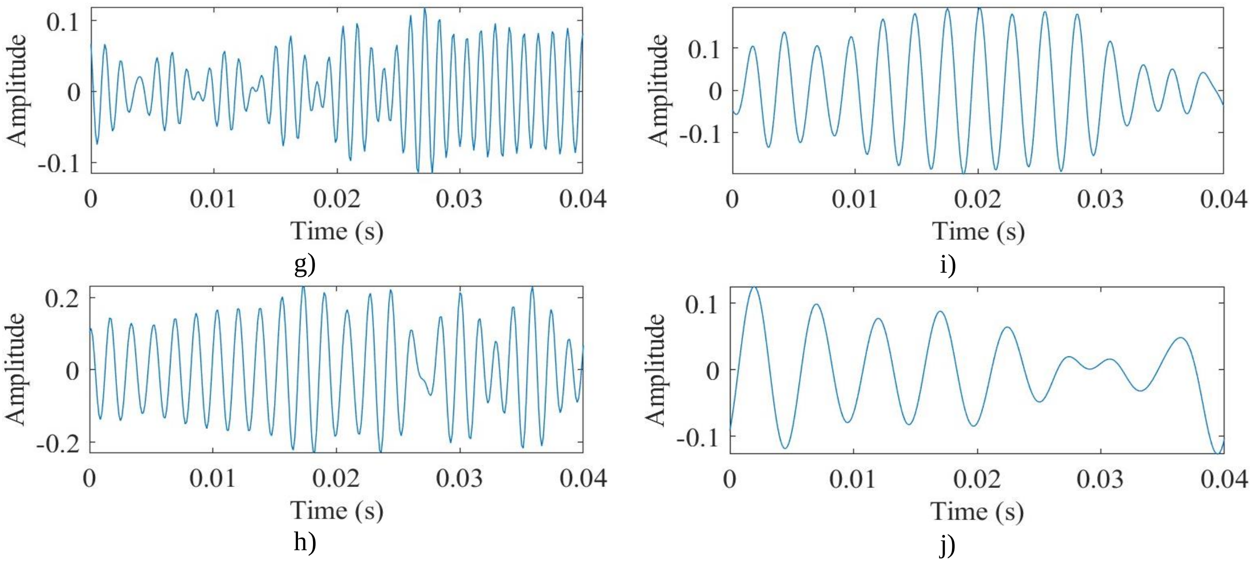
a) b) c) d)

Fig. 9. Adaptive EWT bandpass filters for voice commands:
"Up" a), "Down" b), "Right" c), "Left" d)



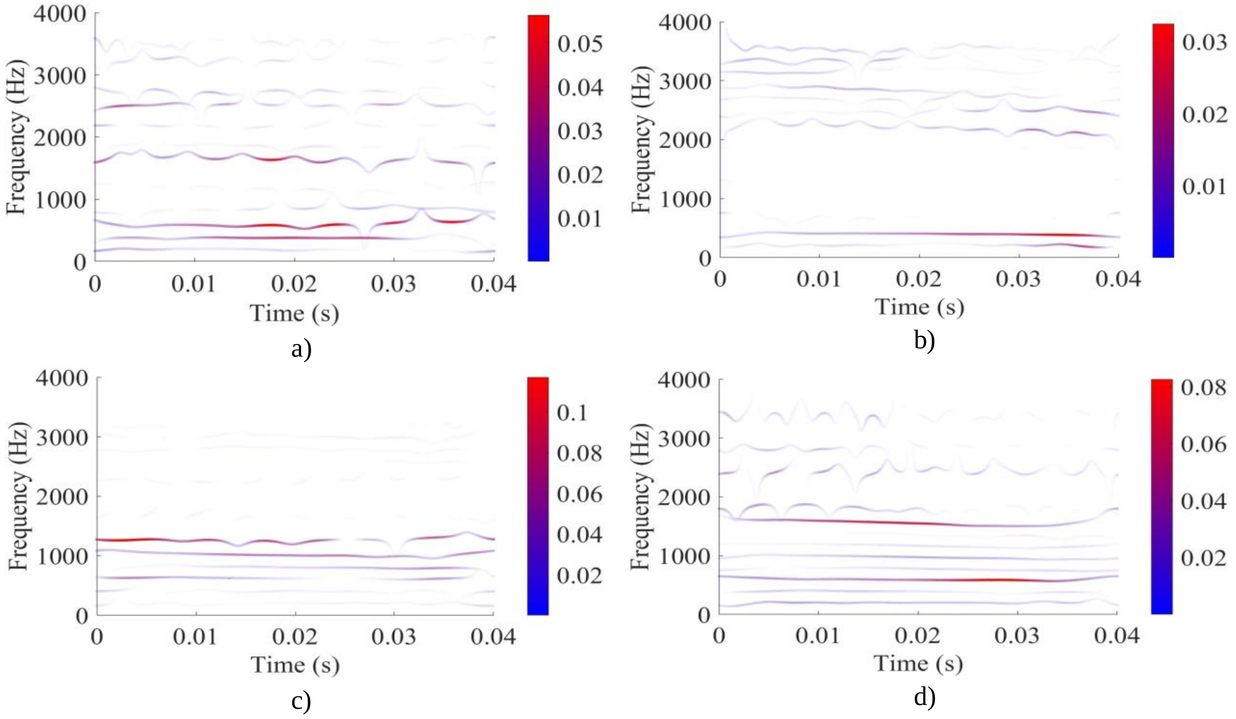
a) d) b) e) c) f)

Fig. 10. Internal empirical modes (IEMs) of the EWT of the studied signal for the voice command "Up":
IEM 3 a), IEM 4 b), IEM 5 c), IEM 6 d), IEM 7 e), IEM 8 f)



g) i) h) j)

Fig. 10. Internal empirical modes (IEMs) of the EWT of the studied signal for the voice command "Up": IEM 9 g), IEM 10 h), IEM 11 i), IEM 12 j)

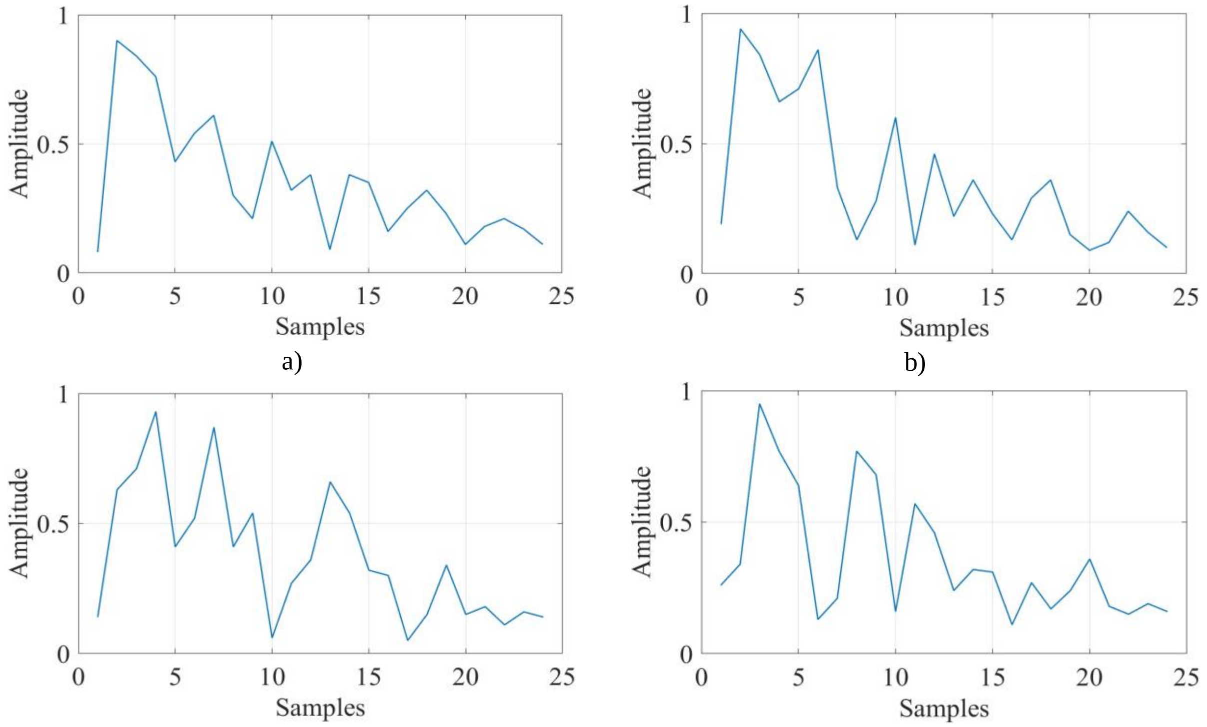


a) b) c) d)

Fig. 11. Hilbert spectral analysis of the IEMs of the EWT for voice commands: "Up" a), "Down" b), "Right" c), "Left" d)

At the next stage, thresholding of the Hilbert spectrum plays a crucial role in rejecting spectral coefficients that do not carry semantic information of the speech signal. Thus, we obtain a very small set of semantic features that, when encoded, successfully replaces thousands of samples of the speech signal that correspond exactly to the semantic form of the speech signal. Fig. 12 shows the semantic features of voice commands based on EWT and Hilbert spectral analysis after thresholding according to the proposed method.

The results of scientific and experimental studies on improving the efficiency of semantic coding of speech signals are presented in Table 2. In this experiment, we evaluated CR, BR, CC, SNR, PSNR, and RMSE for two implementations of the semantic features of voice commands found on the basis of EWT and Gilbert spectral analysis. The following results of the study (see Table 2) clearly show that the optimal solution for the given criteria of semantic coding efficiency for voice commands is: "Up" - CR = 333, BR = 192 bit/s,



a)

b)

c)

d)

Fig. 12. Semantic features based on EWT of voice commands:
"Up" a), "Down" b), "Right" c), "Left" d)

Table 2

Results of evaluating the effectiveness of the developed method of semantic coding of speech signals on the example of two implementations of voice commands

Training	Testing			
Voice commands	"Up"	"Down"	"Right"	"Left"

"Up"	CR = 333
	BR = 192 bit/s
	CC = 0.96
	SNR = 23 dB
	PSNR = 39 dB
	RMSE = 0.03
		CR = 333
		BR = 192 bit/s
		CC = 0.34
		SNR = 7 dB
		PSNR = 13 dB
		RMSE = 0.40
			CR = 333
			BR = 192 bit/s
			CC = 0.14
			SNR = 6 dB
			PSNR = 10 dB
			RMSE = 0.54
				CR = 333
				BR = 192 bit/s
				CC = 0.43
				SNR = 8 dB
				PSNR = 17 dB
				RMSE = 0.49

"Down"	CR = 333
	BR = 192 bit/s
	CC = 0.38
	SNR = 9 dB
	PSNR = 15 dB
	RMSE = 0.39
		CR = 333
		BR = 192 bit/s
		CC = 0.95
		SNR = 24 dB
		PSNR = 44 dB
		RMSE = 0.07
			CR = 333
			BR = 192 bit/s
			CC = 0.24
			SNR = 4 dB
			PSNR = 9 dB
			RMSE = 0.37
				CR = 333
				BR = 192 bit/s
				CC = 0.34
				SNR = 6 dB
				PSNR = 13 dB
				RMSE = 0.43

"Right"	CR = 333
	BR = 192 bit/s
	CC = 0.15
	SNR = 8 dB
	PSNR = 14 dB
	RMSE = 0.50
		CR = 333
		BR = 192 bit/s
		CC = 0.28
		SNR = 6 dB
		PSNR = 12 dB
		RMSE = 0.40
			CR = 333
			BR = 192 bit/s
			CC = 0.97
			SNR = 25 dB
			PSNR = 46 dB
			RMSE = 0.05
				CR = 333
				BR = 192 bit/s
				CC = 0.17
				SNR = 5 dB
				PSNR = 13 dB
				RMSE = 0.63

"Left"	CR = 333
	BR = 192 bit/s
	CC = 0.43
	SNR = 5 dB
	PSNR = 10 dB
	RMSE = 0.53
		CR = 333
		BR = 192 bit/s
		CC = 0.40
		SNR = 6 dB
		PSNR = 14 dB
		RMSE = 0.52
			CR = 333
			BR = 192 bit/s
			CC = 0.25
			SNR = 4 dB
			PSNR = 12 dB
			RMSE = 0.47
				CR = 333
				BR = 192 bit/s
				CC = 0.93
				SNR = 27 dB
				PSNR = 44 dB
				RMSE = 0.09

CC = 0.96, SNR = 23 dB, PSNR = 39 dB, RMSE = 0.03; "Down" - CR = 333, BR = 192 bit/s, CC = 0.95, SNR = 24 dB, PSNR = 44 dB, RMSE = 0. 07; "Right" - CR = 333, BR = 192 bit/s, CC = 0.97, SNR = 25 dB, PSNR = 46 dB, RMSE = 0.05; "Left" - CR = 333, BR = 192 bit/s, CC = 0.93, SNR = 27 dB, PSNR = 44 dB, and RMSE = 0.09.

It shows quite good results, preserving the semantic features of voice commands found on the basis of EWT and Gilbert spectral analysis. This enables the semantic identification of speech signals.

To assess the developed method more reliably, we need to check it for invariance to the realization of speech signals. This experiment was conducted by increasing the number of realizations of speech signals of the same semantic property. The results of the following scientific and experimental studies to evaluate the increase in the efficiency of semantic coding of speech signals using the developed method are presented in Table 3.

In this experiment, we evaluated the average values of CR, BR, CC, SNR, PSNR, and RMSE for twenty implementations of the semantic features of voice commands based on EWT and Gilbert spectral analysis.

The following results of the study (see Table 3) clearly show that with the increase in implementations, the performance indicators of semantic coding remain at a high level, where the semantic component of speech signals retains its semantic patterns, which makes this method resistant to non-stationary and nonlinear processes.

This fact is also confirmed in Fig. 13, which shows twenty realizations of the semantic features found on the basis of EWT and Gilbert spectral analysis of voice commands: "Up", "Down", "Right", "Left".

An experimental study has shown (Fig. 14, where line 1 is number of non-semantic features and line 2 is energy of semantic features) that the developed method of semantic coding of speech signals based on empirical wavelet transform reduces the coding rate from 320 to 192 bits/s and the required bandwidth from 40 to 24 Hz

Table 3

Results of evaluating the effectiveness of the developed method of semantic coding of speech signals on the example of twenty implementations of voice commands

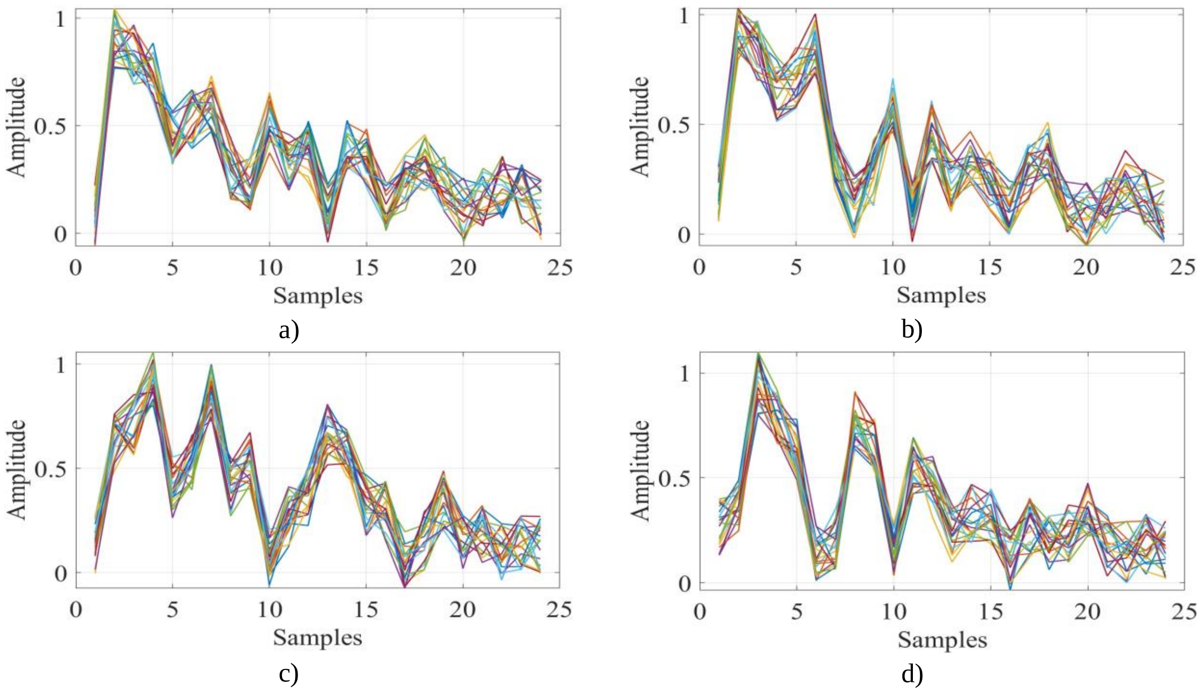
Training	Testing			
Voice commands	"Up"	"Down"	"Right"	"Left"

"Up"	CR = 333
	BR = 192 bit/s
	CC = 0.93
	SNR = 18 dB
	PSNR = 35 dB
	RMSE = 0.08
		CR = 333
		BR = 192 bit/s
		CC = 0.31
		SNR = 7 dB
		PSNR = 12 dB
		RMSE = 0.47
			CR = 333
			BR = 192 bit/s
			CC = 0.19
			SNR = 4 dB
			PSNR = 8 dB
			RMSE = 0.62
				CR = 333
				BR = 192 bit/s
				CC = 0.35
				SNR = 6 dB
				PSNR = 13 dB
				RMSE = 0.48

"Down"	CR = 333
	BR = 192 bit/s
	CC = 0.33
	SNR = 9 dB
	PSNR = 16 dB
	RMSE = 0.37
		CR = 333
		BR = 192 bit/s
		CC = 0.92
		SNR = 22 dB
		PSNR = 39 dB
		RMSE = 0.11
			CR = 333
			BR = 192 bit/s
			CC = 0.08
			SNR = 2 dB
			PSNR = 6 dB
			RMSE = 0.65
				CR = 333
				BR = 192 bit/s
				CC = 0.25
				SNR = 6 dB
				PSNR = 9 dB
				RMSE = 0.48

"Right"	CR = 333
	BR = 192 bit/s
	CC = 0.12
	SNR = 4 dB
	PSNR = 12 dB
	RMSE = 0.59
		CR = 333
		BR = 192 bit/s
		CC = 0.20
		SNR = 7 dB
		PSNR = 11 dB
		RMSE = 0.42
			CR = 333
			BR = 192 bit/s
			CC = 0.93
			SNR = 21 dB
			PSNR = 37 dB
			RMSE = 0.09
				CR = 333
				BR = 192 bit/s
				CC = 0.16
				SNR = 7 dB
				PSNR = 9 dB
				RMSE = 0.64

"Left"	CR = 333
	BR = 192 bit/s
	CC = 0.24
	SNR = 3 dB
	PSNR = 7 dB
	RMSE = 0.55
		CR = 333
		BR = 192 bit/s
		CC = 0.28
		SNR = 6 dB
		PSNR = 9 dB
		RMSE = 0.57
			CR = 333
			BR = 192 bit/s
			CC = 0.21
			SNR = 4 dB
			PSNR = 10 dB
			RMSE = 0.43
				CR = 333
				BR = 192 bit/s
				CC = 0.91
				SNR = 22 dB
				PSNR = 37 dB
				RMSE = 0.13



a)

b)

c)

d)

Fig. 13. Twenty implementations of semantic features based on EWT voice commands: "Up" a), "Down" b), "Right" c), "Left" d)

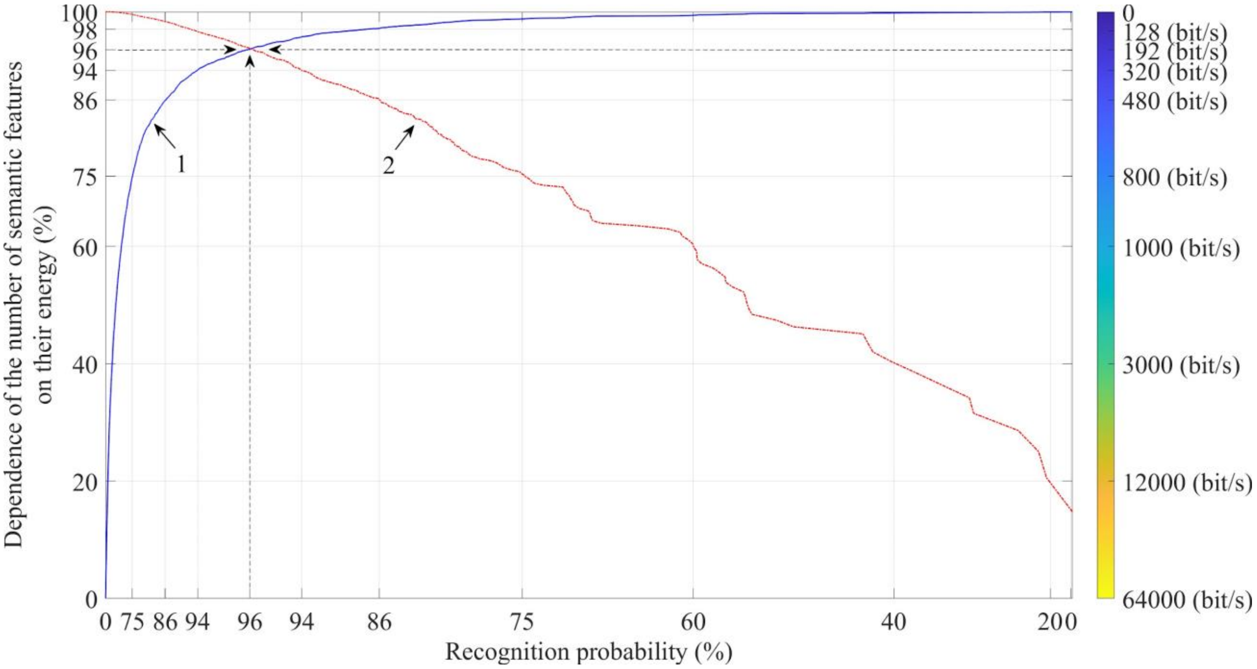


Fig. 14. Dependence of the probability of recognizing semantic features on the reduction in their energy and coding speed using the proposed method

with a probability of error-free recognition of about 0.96 (96%) and a signal-to-noise ratio of 48 dB, according to which its efficiency increases by 1.6 times in contrast to the existing method, without exceeding the boundary value of the processing and data transmission delay of 300 ms, this will allow the system to operate in real time.

5. Discussion and future research directions

In conclusion, the authors would like to make a few points and explain the main trends in this area for future research.

First, the issue of uneven semantic coding (uneven bitrate of semantic speech data transmission, i.e., recognition features of different dimensions) remains open, considering the distribution of internal probability dependencies of the message source. The relevance of solving this problem lies in the fact that we will be able to reduce the semantic data transmission rate by at least 20% additionally, i.e., reduce the coding rate from 192 to about 150 bits/s and the required bandwidth from 40 to about 30 Hz, which is very attractive from a scientific and engineering point of view and thus bring the bitrate of semantic coding even closer to the minimum possible level from the theoretical point of view, which was mentioned in the introduction of this article.

Second, solving the problem of uneven semantic encoding, we can already see a very significant additional problem that will also need to be solved, namely, what criterion to use for objective comparison (classification) of semantic data, since they will not be invariant in frequency between each implementation of recognition features, and invariance is the main property that must be observed according to the theory of pattern recognition and methods of extracting recognition features. Thus, by reducing the speed of semantic coding due to uneven processing of spectral coefficients, we lose frequency invariance, which makes it virtually impossible to classify recognition features using existing methods, i.e., uneven semantic coding and invariance of semantic recognition features are mutually opposite properties. The above problems are of primary importance for solving to obtain significant improvements in practical results in this area of research.

Conclusions

The result of this work is the solution to the actual scientific and practical task of developing and researching new effective methods of semantic coding of speech signals.

During this research, the following scientific results were obtained:

– the well-known method of semantic coding of speech signals based on mel-frequency cepstral coefficients, which does not comply with the condition of adaptability to the studied signal, is investigated, which is a significant drawback of the existing method.

– it is proposed to use the adaptive empirical wavelet transform method in the tasks of multiple-scale analysis and semantic coding of speech signals, which will increase the efficiency of spectral analysis by decomposing the high-frequency speech oscillation into its low-frequency components, namely, internal empirical modes.

– we developed a method of semantic coding of speech signals based on empirical wavelet transform, which builds sets of adaptive bandpass Meyer wavelet filters with the subsequent application of Hilbert spectral analysis to find instantaneous amplitudes and frequencies of functions of internal empirical modes, which will allow us to determine the semantic features of speech signals and increase the efficiency of their coding.

– the optimal threshold processing function is selected and its parameters of threshold values λ_1, λ_2 of wavelet filtering are estimated, which allows finding the optimal thresholds λ_{1opt}, λ_{2opt} with a minimum standard deviation $\Delta(\lambda_1,\lambda_2)$, thereby increasing the efficiency of determining the semantic features of the speech signal.

– the adaptive threshold processing of the Hilbert spectrum of the speech signal with the calculation of the optimal threshold values of wavelet filtering λ_{1opt}, λ_{2opt} was carried out, to filter out the coefficients characterizing instantaneous amplitudes and frequencies of low power.

– the method of semantic coding of speech signals based on mel-frequency cepstral coefficients, but using the basic principles of adaptive spectral analysis with the help of empirical wavelet transform, which increases the efficiency of this method by at least 1.3 times, is investigated.

– the developed method of semantic coding of speech signals based on empirical wavelet transform allows the coding rate to be reduced from 320 to 192 bit/s and the required bandwidth from 40 to 24 Hz with a probability of error-free recognition of approximately 0.96 (96%) and a signal-to-noise ratio of 48 dB, according to which its efficiency increases by 1.6 times in contrast to the existing method, without exceeding the boundary value of the processing and data transmission delay of 300 ms. This will allow the system to operate in real time.

– we developed an algorithm for semantic coding of speech signals based on empirical wavelet transform and its software implementation in MATLAB R2022b programing language.

Contributions of authors: conceptualization, methodology – **Oleksandr Lavrynenko;** formulation of tasks, analysis **– Oleksandr Lavrynenko, Denis Bakhtiiarov;** development of model, software, verification **– Vitaliy Kurushkin;** analysis of results, visualization **– Serhii Zavhorodnii;** writing – original draft preparation **– Veniamin Antonov;** writing – review and editing **– Petro Stanko.**

All the authors have read and agreed to the published version of this manuscript.

References

1. Boucheron, L. E., De Leon, P. L., & Sandoval, S. Low bit-rate speech coding through quantization of mel-frequency cepstral coefficients. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, vol. 20, no. 2, pp. 610-619. DOI: 10.1109/TASL. 2011.2162407.

2. Chai, L., Du, J., Liu, Q., & Lee, C. A Cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing dnn-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 29, no. 1, pp. 106-117. DOI: 10.1109/TASLP.2020.3036783.

3. Patel, M., Kothari, A., & Koringa, H. A novel approach for semantic segmentation of automatic road network extractions from remote sensing images by modified UNet. *Radioelectronic and Computer Systems*, 2022, no. 3, pp. 161-173. DOI: 10.32620/reks.2022.3.12.

4. Bu, S., Zhao, Y., Zhao, T., Wang, S., Han, M. Modeling speech structure to improve T-F masks for speech enhancement and recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022, vol. 30, no. 1, pp. 2705-2715. DOI: 10.1109/TASLP.2022.3196168.

5. Barannik, V., Sidchenko, S., Barannik, D., Yermachenkov, A., Savchuk, M., & Pris, M. Video images compression method based on floating positional coding with an unequal codograms length. *Radioelectronic and Computer Systems*, 2023, no. 1, pp. 134-146. DOI: 10.32620/reks.2022.1.11.

6. Ai, Y., Ling, Z., Wu, W., & Li, A. Denoising-and-dereverberation hierarchical neural vocoder for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022, vol. 30, no. 1, pp. 2036-2048. DOI: 10.1109/TASLP.2022.3182268.

7. Lee, K., & Ellis, D. P. W. Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, vol. 18, no. 6, pp. 1406-1416. DOI: 10.1109/TASL.2009.2034776.

8. Luo, M., Wang, D., Wang, X., Qiao, S., & Zhou, Y. Error-diffusion based speech feature quantization for small-footprint keyword spotting. *IEEE Signal Processing Letters*, 2022, vol. 29, no. 1, pp. 1357-1361. DOI: 10.1109/LSP.2022.3179208.

9. Karbasi, M., Zeiler, S., & Kolossa, D. Microscopic and blind prediction of speech intelligibility: theory and practice. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022, vol. 30, no. 1, pp. 2141-2155. DOI: 10.1109/TASLP.2022.3184888.

10. Milner, B., & Darch, J. Robust acoustic speech feature prediction from noisy mel-frequency cepstral coefficients. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, vol. 19, no. 2, pp. 338-347. DOI: 10.1109/TASL.2010.2047811.

11. Milner, B., & Shao, X. Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, vol. 15, no. 1, pp. 24-33. DOI: 10.1109/TASL.2006.876880.

12. Hazra, S., Ema, R., Galib, S., Kabir, S., & Adnan, N. Emotion recognition of human speech using deep learning method and MFCC features. *Radioelectronic and Computer Systems*, 2022, no. 4, pp. 161-172. DOI: 10.32620/reks.2022.4.13.

13. Zhang, Y., & Ling, Z. Extracting and predicting word-level style variations for speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 29, no. 1, pp. 1582-1593. DOI: 10.1109/TASLP.2021.3074757.

14. Tulyakova, N., & Trofymchuk, O. Adaptive myriad filter with time-varying noise- and signal-dependent parameters. *Radioelectronic and Computer Systems*, 2022, no. 2, pp. 217-238. DOI: 10.32620/reks.2022.2.17.

15. Farias, F., & Coelho, R. Blind adaptive mask to improve intelligibility of non-stationary noisy speech. *IEEE Signal Processing Letters*, 2021, vol. 28, no. 1, pp. 1170-1174. DOI: 10.1109/LSP.2021.3086405.

16. Rudenko, O., & Bezsonov, O. Adaptive identification under the maximum correntropy criterion with variable center. *Radioelectronic and Computer Systems*, 2022, no. 1, pp. 216-228. DOI: 10.32620/reks. 2022.1.17.

17. Daubechies, I., Lu, J., & Wu, H-T. Synchro-squeezed wavelet transforms: An empirical mode decomposition-like tool. *Journal of Applied and Computational Harmonic Analysis*, 2011, vol. 30, no. 2, pp. 243-261. DOI: 10.1016/j.acha.2010.08.002.

18. Gilles, J. Empirical Wavelet Transform. *IEEE Transactions on Signal Processing*, 2013, vol. 61, no. 16, pp. 3999-4010. DOI: 10.1109/TSP.2013.2265222.

19. Donoho, D. L., Javanmard, A., & Montanari, A. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 2013, vol. 59, no. 11, pp. 7434-7464. DOI: 10.1109/TIT.2013.2274513.

20. Lavrynenko, O., Konakhovych, G., & Bakhtiiarov, D. Method of voice control functions of the UAV. *Proc. IEEE 4th Int. Conf. on Methods and Systems of Navigation and Motion Control (MSNMC)*, Kyiv, Oct. 18-20, 2016, pp. 47-50. DOI: 10.1109/MSNMC. 2016.7783103.

21. Lavrynenko, O., Odarchenko, R., Konakhovych, G., Taranenko, A., Bakhtiiarov, D., & Dyka, T. Method of semantic coding of speech signals based on empirical wavelet transform. *Proc. IEEE 4th Int. Conf.*

on Advanced Information and Communication Technologies (AICT), Lviv, Sept. 21-25, 2021, pp. 18-22. DOI: 10.1109/AICT52120.2021.9628985.

22. Odarchenko, R., Lavrynenko, O., Bakhtiiarov, D., Dorozhynskyi, S., Antonov, V., & Zharova, O. Empirical wavelet transform in speech signal compression problems. *Proc. IEEE 8th Int. Conf. on Problems of Infocommunications, Science and Technology (PIC S&T),* Kharkiv, Oct. 5-7, 2021, pp. 599-602. DOI: 10.1109/PICST54195.2021.9772156.

23. Veselska, O., Lavrynenko, O., Odarchenko, R., Zaliskyi, M., Bakhtiiarov, D., Karpinski, M., & Rajba, S. A Wavelet-based steganographic method for text hiding in an audio signal. *Sensors,* 2022, vol. 22, no. 15, pp. 1-25. DOI: 10.3390/s22155832.

Received 19.05.2023, Accepted 20.09.2023

МЕТОД ВИДІЛЕННЯ СЕМАНТИЧНИХ ОЗНАК РОЗПІЗНАВАННЯ МОВНИХ СИГНАЛІВ НА ОСНОВІ ЕМПІРИЧНОГО ВЕЙВЛЕТ-ПЕРЕТВОРЕННЯ

Олександр Лавриненко, Денис Бахтіяров, Віталій Курушкін,
Сергій Завгородній, Веніамін Антонов,
Петро Станко

Предметом дослідження є методи підвищення ефективності семантичного кодування мовних сигналів. **Метою** дослідження є розроблення методу підвищення ефективності семантичного кодування мовних сигналів, де під ефективністю кодування розуміється зниження швидкості передачі інформації із заданою ймовірністю безпомилкового розпізнавання семантичних ознак мовних сигналів, що дозволить значно знизити необхідну смугу пропускання, тим самим підвищуючи пропускну здатність каналу зв'язку. Для досягнення поставленої мети необхідно вирішити наступні наукові **задачі:** дослідити відомий метод підвищення ефективності семантичного кодування мовних сигналів на основі мел-частотних кепстральних коефіцієнтів; обґрунтувати ефективність використання адаптивного емпіричного вейвлет-перетворення в задачах кратномасштабного аналізу та семантичного кодування мовних сигналів; розробити метод семантичного кодування мовних сигналів на основі адаптивного емпіричного вейвлет-перетворення з подальшим застосуванням спектрального аналізу Гільберта та оптимальної порогової обробки; провести об'єктивну кількісну оцінку підвищення ефективності розробленого методу семантичного кодування мовних сигналів на відміну від існуючого методу. Під час дослідження були одержані наступні наукові **результати:** вперше розроблено метод семантичного кодування мовних сигналів на основі емпіричного вейвлет-перетворення, який відрізняється від існуючих методів побудовою множини адаптивних смугових вейвлет-фільтрів Мейера з подальшим застосуванням спектрального аналізу Гільберта для знаходження миттєвих амплітуд і частот функцій внутрішніх емпіричних мод, що дозволить визначити семантичні ознаки мовних сигналів та підвищити ефективність їх кодування; вперше запропоновано використовувати метод адаптивного емпіричного вейвлет-перетворення в задачах кратномасштабного аналізу та семантичного кодування мовних сигналів, що дозволить підвищити ефективність спектрального аналізу за рахунок розкладання високочастотного мовного коливання на його низькочастотні складові, а саме внутрішні емпіричні моди; отримав подальший розвиток метод семантичного кодування мовних сигналів на основі мел-частотних кепстральних коефіцієнтів, але з використанням базових принципів адаптивного спектрального аналізу за допомогою емпіричного вейвлет-перетворення, що підвищує ефективність даного методу. **Висновки:** розроблено метод семантичного кодування мовних сигналів на основі емпіричного вейвлет-перетворення, що дозволяє знизити швидкість кодування від 320 до 192 біт/с та необхідну смугу пропускання від 40 до 24 Гц з ймовірністю безпомилкового розпізнавання близько 0,96 (96%) і відношенням сигнал/шум 48 дБ, згідно чого його ефективність підвищується в 1,6 рази на відміну від існуючого методу; розроблено алгоритм семантичного кодування мовних сигналів на основі емпіричного вейвлет-перетворення та його програмна реалізація мовою програмування MATLAB R2022b.

Ключові слова: семантичні ознаки мовних сигналів; мел-частотні кепстральні коефіцієнти; адаптивний спектральний аналіз; емпіричне вейвлет-перетворення; адаптивні вейвлет-фільтри Мейєра; функції внутрішніх емпіричних мод; спектральний аналіз Гільберта; оптимальна порогова обробка.

Лавриненко Олександр Юрійович – канд. техн. наук, доц. каф. телекомунікаційних та радіоелектронних систем, Національний авіаційний університет, Київ, Україна.

Бахтіяров Денис Ілшатович – канд. техн. наук, доц. каф. телекомунікаційних та радіоелектронних систем, Національний авіаційний університет, Київ, Україна.

Курушкін Віталій Євгенович – канд. техн. наук, доц. каф. телекомунікаційних та радіоелектронних систем, Національний авіаційний університет, Київ, Україна.

Завгородній Сергій Олександрович – канд. техн. наук, декан факультету аеронавігації, електроніки та телекомунікацій, Національний авіаційний університет, Київ, Україна.

Антонов Веніамін Валерійович – канд. техн. наук, доц. каф. телекомунікаційних та радіоелектронних систем, Національний авіаційний університет, Київ, Україна.

Станко Петро Олександрович – канд. техн. наук, директор студмістечка, Національний авіаційний університет, Київ, Україна.

Oleksandr Lavrynenko – Candidate of Technical Sciences, Associate Professor at the Telecommunication and Radio-electronic Systems Department, National Aviation University, Kyiv, Ukraine,
e-mail: oleksandrlavrynenko@gmail.com, ORCID: 0000-0002-7738-161X.

Denys Bakhtiiarov – Candidate of Technical Sciences, Associate Professor at the Telecommunication and Radio-electronic Systems Department, National Aviation University, Kyiv, Ukraine,
e-mail: bakhtiiaroff@tks.nau.edu.ua, ORCID: 0000-0003-3298-4641.

Vitalii Kurushkin – Candidate of Technical Sciences, Associate Professor at the Telecommunication and Radio-electronic Systems Department, National Aviation University, Kyiv, Ukraine,
e-mail: vitaliy.kurushkin@npp.nau.edu.ua.

Serhii Zavhorodnii – Candidate of Technical Sciences, Dean at the Faculty of Air Navigation, Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine,
e-mail: serhii.zavhorodnii@npp.nau.edu.ua.

Veniamin Antonov – Candidate of Technical Sciences, Associate Professor at the Telecommunication and Radio-electronic Systems Department, National Aviation University, Kyiv, Ukraine,
e-mail: veniaminas@tks.nau.edu.ua.

Petro Stanko – Candidate of Technical Sciences, Campus Director, National Aviation University, Kyiv, Ukraine,
e-mail: p_stanko@ukr.net, ORCID: 0000-0001-5794-3593.