Method of Remote Biometric Identification of a Person by Voice based on Wavelet Packet Transform

Oleksandr Lavrynenko[1], Bohdan Chumachenko[1], Maksym Zaliskyi[1], Serhii Chumachenko[1], and Denys Bakhtiiarov[1]

[1] *National Aviation University, 1 Lubomyr Huzar ave., Kyiv, 03058, Ukraine*

Abstract

In this research, the task of extracting speech signal recognition features for voice identification of a person in a remote mode was solved, which imposes several restrictions, namely: (1) minimum processing time of the speech signal realization, since the required recognition reliability is achieved through statistical processing of the results; (2) reduction of the dimensionality of recognition features, since the process of extracting recognition features and their classification occurs on the transmitting side of the communication channel, which in turn imposes certain factors of computing power and noise in the communication channel. After analyzing the given conditions of the voice identification system, the question arose of developing a method for extracting speech signal recognition features that would provide more informative spectral characteristics of the speech signal, which would improve the efficiency of their further classification under the influence of noise. In this paper, we consider the possibility of applying the theory of time-scale analysis to solve this problem, namely, the development of a method for extracting recognition features based on the wavelet packet transform using the orthogonal basis wavelet function of Meyer and subsequent averaging of wavelet coefficients that are in the frequency band of the corresponding wavelet packet. Experimental studies have shown the ability of the developed method to generate speech signal recognition features with a close frequency-temporal structure based on wavelet packets in the Meyer basis, namely, it was found that at a signal-to-noise ratio of 10 dB, the features obtained based on the developed method have a very acceptable result, namely, 1.6–2 times more robust to noise than the features obtained based on the traditional Fourier spectrum, where the total deviation of the root mean square error of the obtained features is unacceptable at a signal-to-noise ratio of 20 dB.

Keywords

speech signal, recognition features, wavelet transform, wavelet Meyer function, spectral analysis, voice identification, biometric authentication

1. Introduction

The development of new methods and means of ensuring information security is intended primarily to prevent threats of access to information resources by unauthorized persons. To solve this problem, it is necessary to have identifiers and create identification procedures for all users. Modern identification and authentication include various systems and methods of biometric identification [1]. The development of identification systems based on biometric measurements is associated with a whole range of advantages: such systems are more reliable because biometric indicators are more difficult to fake; modern microprocessor technology makes biometric methods more convenient than conventional identification

CPITS-2024: Cybersecurity Providing in Information and Telecommunication Systems, February 28, 2024, Kyiv, Ukraine
EMAIL: oleksandrlavrynenko@gmail.com (O. Lavrynenko); bohdan.chumachenko@npp.nau.edu.ua (B. Chumachenko); maksym.zaliskyi@npp.nau.edu.ua (M. Zaliskyi); serhii.chumachenko@npp.nau.edu.ua (S. Chumachenko); bakhtiiaroff@tks.nau.edu.ua (D. Bakhtiiarov)
ORCID: 0000-0002-7738-161X (O. Lavrynenko); 0000-0002-0354-2206 (B. Chumachenko); 0000-0002-1535-4384 (M. Zaliskyi); 0009-0003-8755-5286 (S. Chumachenko); 0000-0003-3298-4641 (D. Bakhtiiarov)

methods; and, finally, they are much easier to automate measurements [2–6].

One of the most common biometric characteristics of a person is his or her voice, which has a set of individual characteristics that are relatively easy to measure (for example, the frequency spectrum of the speech signal). The advantages of voice identification also include ease of application and use, and the fairly low cost of devices used for identification (e. g., microphones) [7].

Voice identification capabilities cover a very wide range of tasks, which distinguishes them from other biometric systems. First of all, voice identification has been widely used for a long time in various systems for differentiating access to physical objects and information resources. Its new application in remote voice identification systems, where a person is identified through a telecommunications channel, seems promising. For example, in mobile communications, voice can be used to manage services, and the introduction of voice identification helps protect against fraud [8].

Voice identification is of particular importance in the investigation of crimes, in particular in the field of computer information, and in the formation of the evidence base for such an investigation. In these cases, it is often necessary to identify an unknown voice recording. Voice identification is an important practical task when searching for a suspect based on a voice recording in telecommunication channels. Determining such characteristics of the speaker's voice as gender, age, nationality, dialect, and emotional coloring of speech is also important in the field of forensics and anti-terrorism. The identification results are important in conducting phonoscopic examinations, and in carrying out expert forensic research based on the theory of forensic identification [9].

Voice identification in real-world environments faces the following serious challenges. Firstly, such identification is subject to all kinds of hardware distortions and noise caused by the peculiarities of equipment and devices for recording, processing, and storing information. Secondly, external acoustic noise inevitably superimposes the speech signal, which can significantly distort individual informative characteristics. Given this, identification systems that have demonstrated fairly high efficiency in laboratory conditions may show much lower reliability when analyzing speech information with external noise. Finally, in several tasks, identification has to be performed in very difficult conditions of overlapping voices of several speakers, in particular, with similar acoustic characteristics. It should be noted that there have been virtually no studies of voice identification capabilities for this most difficult case [10].

Voice identification involves a set of technical, algorithmic, and mathematical methods that cover all stages, from voice recording to voice data classification. The discussed difficulties and shortcomings lead to the conclusion that further development of voice identification systems requires the development of new approaches aimed at processing large arrays of experimental speech signals, their effective analysis, and reliable classification. This indicates the relevance of research on the creation of new mathematical methods for processing, analyzing, and classifying voice data that would ensure the reliability and accuracy of person identification [11].

Traditionally, the methods that provide the required level of classification reliability under given conditions are of practical interest for speech signal recognition. Until recently, the dominant approach to the construction of biometric voice identification devices was not to impose restrictions on the processing time of the speech signal, since the required recognition reliability was achieved by statistical processing of the results obtained, as well as by increasing the dimensionality of the recognition features, and as a rule, the process of extracting recognition features and their classification took place on the transmitting side of the communication channel.

However, in the case of remote voice identification in modern mobile radio communication systems, it is difficult to ensure these conditions, since the identification of a person is carried out on the receiving side, and this, in turn, imposes certain factors of computing power and the influence of noise in the communication channel. An additional requirement is often the need to make a classification decision in a time-sensitive environment [12].

In this case, it is necessary to move to other methods that can provide the necessary contrast of the speech signal in the formed

recognition features by the specified conditions, namely, to ensure the quality of speech signal recognition features extraction under the influence of noise in the communication channel, which in turn will allow the use of voice identification technologies in a remote mode based on modern mobile radio communication systems, which will significantly expand the scope of this type of technology. In this paper, we consider the possibility of applying the theory of time-scale analysis to solve this problem [13].

2. Literature Analysis and Problem Statement

In general, recognition is the process of assigning the object under study, in this case, a speech signal represented by a set of observations, to one of the alternative classes. The process of assigning an object to a class is based on the existing differences in some ordered set of recognition features [14]. Traditionally, these features are formed based on such parameters of the speech signal as the duration of the modulating function elements, the number of signal envelope extremes, statistical characteristics of the number of zero-level transitions, and the moments of higher orders of the spectrum shape obtained as a result of observations. Then the set of observations is represented in the form of a matrix

$$X_{pn} = \begin{bmatrix} x_{11} & x_{12} & \dots x_{1i} & \dots x_{1n} \\ x_{21} & x_{22} & \dots x_{2i} & \dots x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{p1} & x_{p2} & \dots x_{pi} & \dots x_{pn} \end{bmatrix},$$

where n is the number of observations used for recognition, and each column $X_i = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right)^T$, $i = 1,2,\dots,n$ of the matrix X_{pn} is a p-dimensional vector of observed values p of features $X_1, X_2, \dots, X_p$ that reflect the most important properties of objects for recognition. The set of features p, as a rule, is the same for all recognition classes $s_1, s_2, \dots, s_k$ [15].

Thus, we consider the task of recognizing the object under study belongs to one of a finite number of classes $s_1, s_2, \dots, s_k$, which are described by a set of features $X_1, X_2, \dots, X_p$, which is the same for all classes. Differences between classes will be manifested only in differences in the characteristics of features of different objects. Then, for any set of features $X_1, X_2, \dots, X_p$, you can set rules according to which any two classes s_1 and s_r are assigned a vector

$$D_{1r} = \begin{bmatrix} d_1^{1r} \\ . \\ . \\ d_q^{1r} \end{bmatrix},$$

which consists of q parameters called interclass distances that express the degree of difference in the characteristics of recognition features [16].

An integral part of the speech signal recognition process is the definition of a set of features $X_1, X_2, \dots, X_p$, i.e., the formation of recognition features in such a way as to ensure the required classification reliability with the minimum possible dimension p. By the considered approach to solving the problem of speech signal recognition, an important point is the choice of a method for forming recognition features. The use of approaches based on the traditional Fourier spectral-time analysis for this purpose is associated with certain difficulties. First, there are high requirements for the input speech signal stream in terms of signal-to-noise ratio. Secondly, the lack of classification reliability for multicomponent and low-stationary signals, such as speech signals, and thirdly, the need for a significant amount of implementations. The desire to overcome these limitations within the framework of traditional approaches of classical spectral signal processing leads to difficult-to-implement variants of speech signal recognition devices and solutions that are unacceptable for the conditions under consideration [17].

Thus, we formulate the research objective: to develop a method that allows the formation of contrasting recognition features for automatic remote identification of a person by voice under the conditions of restrictions on the duration of the processed realization at a signal-to-noise ratio of less than 20 dB under conditions of partial or complete a priori uncertainty about their structure.

3. Proposed Method

Currently, methods of processing and analyzing speech signals based on their wavelet transforms are widely used. The essence of these transformations is to decompose the input signal into a system of basis wavelets—functions, each of which is a shifted and scaled copy of the input (generating or mother wavelet). A characteristic property of wavelet functions (hereinafter referred to as wavelets) is the finite energy at their full localization in both frequency and time domains.

Thus, any sequence of discrete samples of the speech signal $S(t_i)$ can be represented as an ordered set of coefficients of decomposition by a system of scaling functions and wavelet functions:

$$S(t_i) = \sum_{k=1}^{2^{N-M}} V_{m,k}\varphi_{m,k}(t_i) + \sum_{m=1}^{M}\sum_{K=1}^{2^{N-m}} W_{m,k}\psi_{m,k}(t_i),$$

where M is the number of decomposition levels, $V_{m,k}$ and $W_{m,k}$ are the approximating and detailing wavelet decomposition coefficients [18].

Scaling functions and wavelet functions are defined by the theory of multiple-scale analysis:

$$\varphi_{m,k}(t) = \sqrt{2^m}\varphi(2^m t - k), \quad (1)$$

$$\psi_{m,k}(t) = \sqrt{2^m}\psi(2^m t - k). \quad (2)$$

Here, in (1) and (2) $\sqrt{2^m}$ is the normalizing factor, and $k = 0, \pm 1, \pm 2, \ldots; m \in Z$.

In practice, to quickly calculate the values of wavelet coefficients $V_{m,k}$ and $W_{m,k}$ use a sequential separation scheme called the pyramid or Mallat algorithm, which is interpreted as a sequential two-band filtering of the input speech signal using cascaded low-pass (h) and high-pass (g) filter blocks (Fig. 1) [19].

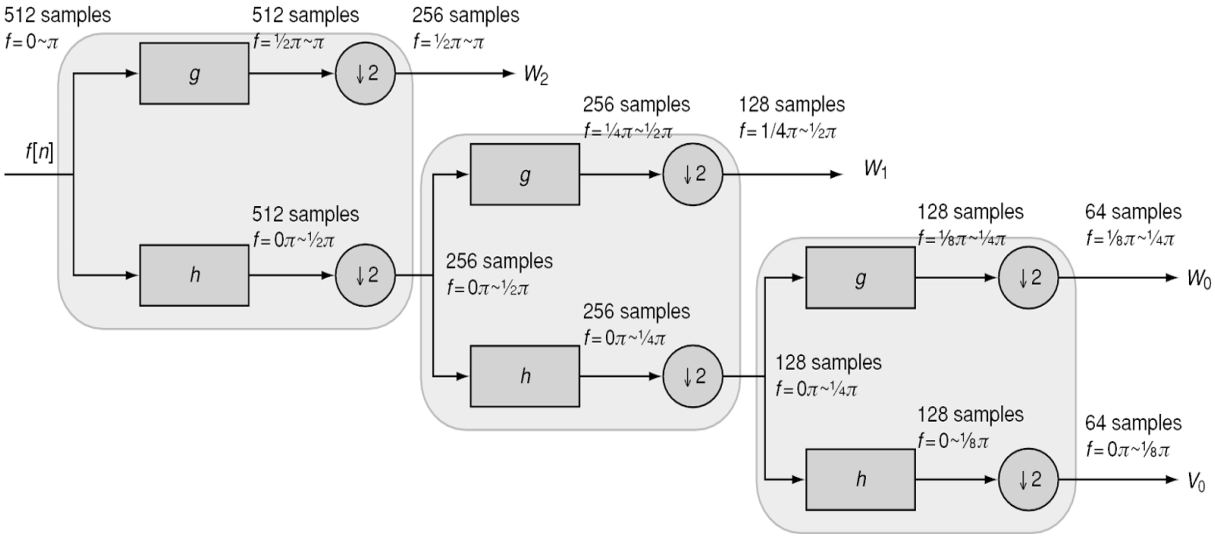


Figure 1: Scheme of signal sequence decomposition according to the Mallat algorithm

In Fig. 1, for the wavelet coefficients $V_{m,k}$ and $W_{m,k}$, the first index m corresponds to the number of the decomposition level, and the second index $k = 0,1,\ldots,2^m - 1$ corresponds to the ordinal value of the wavelet coefficient at the decomposition level m. According to the theory of multiple-scale analysis, the values of $V_{m,k}$ and $W_{m,k}$ can be obtained based on the coefficients calculated at the previous stages of speech signal decomposition:

$$V_{m,k} = \frac{1}{\sqrt{2}}\sum_n V_{m-1,n}h_{n+2k},$$

$$W_{m,k} = \frac{1}{\sqrt{2}}\sum_n V_{m-1,n}g_{n+2k},$$

where h_m and g_m are sequences that define the characteristics of filters H and G at the m level of wavelet decomposition [20].

The number of multiplication operations required to calculate all the coefficients of the discrete wavelet transform for the data set N and the length of the vectors h and g equal to L is $2LN$. The same number of operations is required to recover or calculate all the spectral components. So, to analyze a speech signal on a

wavelet basis, you need to perform $4LN$ operations. The number of complex multiplication operations for the fast Fourier transform is $N\,log_2\,N$, which is comparable to or even greater than in the case of the discrete wavelet transform [21].

The interpretation of the coefficients of the discrete wavelet transform is somewhat more complicated than the Fourier coefficients. If the analyzed speech signal is sampled at a frequency of 8 kHz and consists of 256 samples, then the top frequency of the signal is 4 kHz. Then the coefficients of the first level of decomposition (128) occupy the frequency band [2.0, 4.0] kHz. The second-level wavelet coefficients (64) are responsible for the [1.0, 2.0] kHz frequency band. They are displayed before the first level wavelet coefficients. The procedure is repeated until there is 1 wavelet coefficient and 1 scaling coefficient at level 9. The total number of coefficients is (1+1+2+4+8+16+32+64+128) = 256. That is, the number of coefficients is equal to the number of samples in the input speech signal. If the main energy of the signal was concentrated near the frequency of 1.0 kHz, then the second-level wavelet coefficients will be more informative, and the first-level wavelet coefficients can be neglected [22].

As a continuation of the development of the theory of multiple-scale analysis, it is proposed to improve the Mallat algorithm by additional processing of the high-frequency components of the pyramid of the analyzed speech signal. Thus, in the improved algorithm, recursive filtering is applied to the coefficients $W_{m,k}$. This full decomposition algorithm is called wavelet packet decomposition. The decomposition scheme based on wavelet packets is shown in Fig. 2.

For the wavelet coefficients $\zeta_m^n(i)$ (Fig. 2), the index m corresponds to the number of the decomposition level, the index n corresponds to the number of the subband at the level m, and $i = 0,1,\ldots,2^m - 1$ corresponds to the number of the wavelet coefficients at the level m. In wavelet packages, several decomposition bases are used for complete decomposition, united by the image of nesting in each other, which gave the method its name. In general, each level of the hierarchy can use its specific basis. In contrast to the Mallat algorithm, the use of wavelet packets makes it possible to take into account the subtle structure of the analyzed speech signal process in a more comprehensive way.

Indeed, the absolute values of the coefficients in the wavelet packet decomposition are smaller than those of the Mallat algorithm. Therefore, it can be argued that the approximation with wavelet packets has a much smaller error [23].

Since the wavelet basis is a complete decomposition basis, the wavelet coefficients contain individual characteristics of the input speech signal, determined by the properties of the basis functions to the same extent as the spectral components of the Fourier series. Thus, any wavelet transform, including those based on the use of wavelet packets, allows you to uniquely represent a speech signal by an ordered set of its wavelet coefficients. It is possible to assume the possibility of using them as recognition features and thus put the calculation of coefficients based on wavelet packets based on the proposed method.

The method of forming speech signal recognition features based on wavelet packets is defined as follows. In the wavelet spectrum formed based on wavelet packets, the power of the calculated wavelet coefficients within each subband of the decomposition is averaged. The averaged coefficients are normalized and, according to their place in the overall pyramid of wavelet packets from left to right and from top to bottom, converted into a vector of recognition features. Thus, specific values of the average power of the wavelet coefficients in each subband of the decomposition will serve as the primary features of speech signal recognition. It should be noted that, in general, the features obtained in this way will be correlated, so it is advisable to apply an additional decorrelation transformation to the vector, which, by the way, will reduce the size of the secondary recognition feature space [24].

Consider the sequence of stages of the proposed method (Fig. 3).

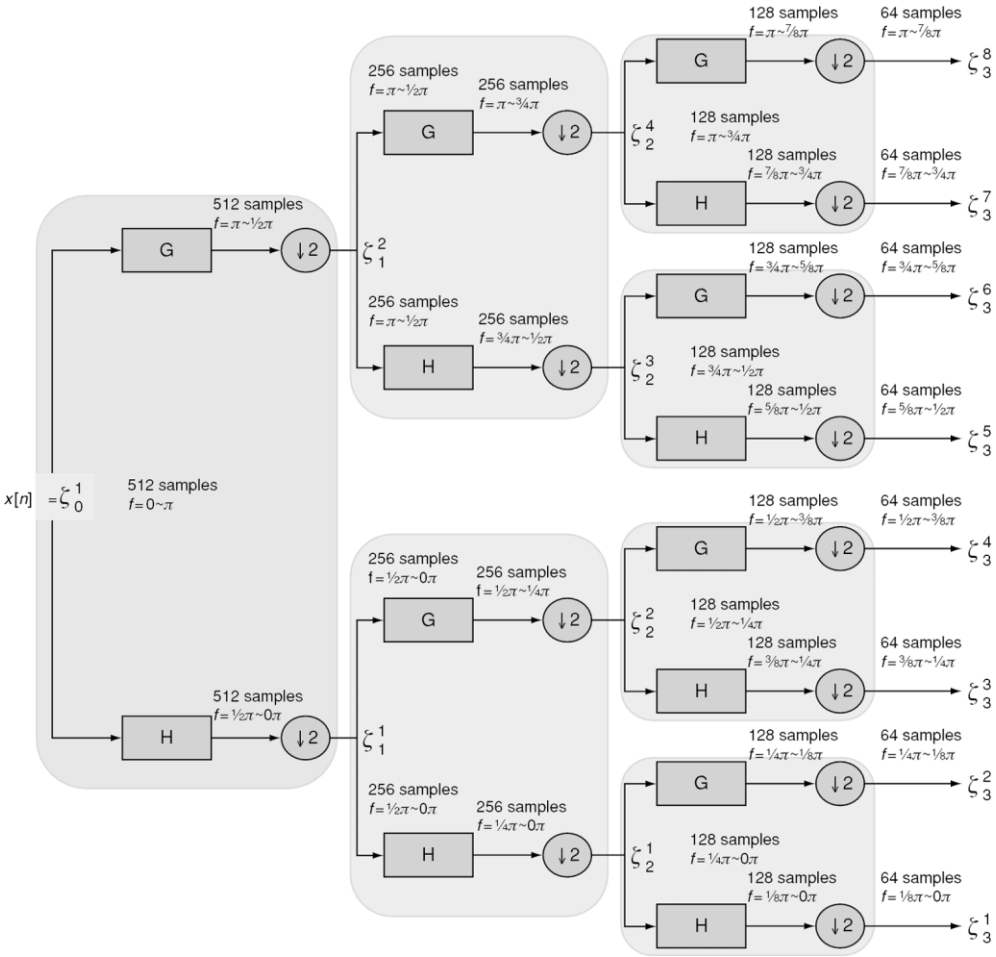


Figure 2: Signal sequence decomposition scheme based on the wavelet packet algorithm

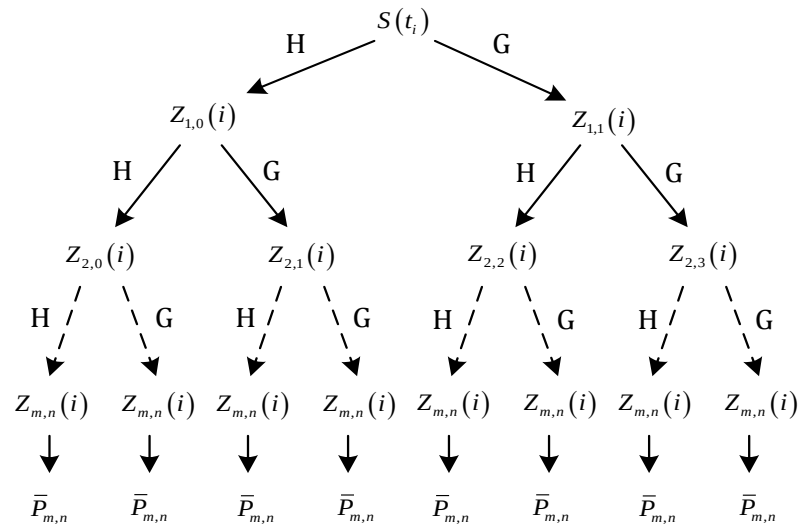


Figure 3: Scheme of speech signal recognition features selection for biometric identification of a person

Initially, the input sequence of discrete samples of the speech signal $S(t_i)$ with length N, a multiple of power 2, at $i = 0,1,2,\dots,(N-1)$ is decomposed into $K \le log_2(N)$ levels as a result of applying the wavelet packet algorithm. At the first level, the input array $S(t_i)$ is decomposed into two sets $Z_{1,0}(i)$ and $Z_{1,1}(i)$ by convolution $S(t_i)$ with sequences $\{h\}$ and $\{g\}$, which are determined by the characteristics of low H and high G frequency filters. At the 2nd level, the considered convolution procedures are repeated with each of the obtained subsets $Z_{1,0}(i)$ and $Z_{1,1}(i)$. The process of full decomposition, called wavelet packetization, involves k steps similar to the first one [25]. The analytically considered procedures can be represented in general by the following expressions:

$$Z_{m,2n}(i) = \sum_{t=0}^{N-1} Z_{m-1,n}(i) h_{m,n}(i),$$

$$Z_{m,2n+1}(i) = \sum_{t=0}^{N-1} Z_{m-1,n}(i) g_{m,n}(i),$$

where is $1 \le m \le K$, and $0 \le n \le (2^{m-1} - 1)$. At the first level of decomposition, the samples of the speech signal $S(t_i)$ are used as $Z_{0,0}(i)$. The values of the elements of the sequences $\{h\}$ and $\{g\}$ depend on the choice of the type of scaling function $\varphi(x)$ and wavelet function $\psi(x)$ and, according to (1) and (2), are calculated as follows:

$$h_{m,n}(i) = 2^{-m/2} \varphi(2^{-m} i - n),$$
$$g_{m,n}(i) = 2^{-m/2} \psi(2^{-m} i - n).$$

As a result of the transformations performed during the decomposition, the sequence of samples of the speech signal $S(t_i)$ is decomposed into $R = 2 \cdot 2^K - 1$ sequences (including the input one) of length $N/2^m$, each of which represents one of the frequency subbands of the input speech signal [26].

Different realizations of speech signals will have different energy distributions over frequency subbands since their Fourier spectra will also be different. If you calculate the average power of the wavelet coefficients in each subband, the set of values obtained will reflect the wavelet content of the speech signal subbands, similar to the frequency representation. Moreover, the transition to the average power will allow the use of relatively short input realizations for recognition, which is an important point in the operation of rapid analysis systems. The bandwidth of the frequencies falling into each of the subbands will narrow with an increase in the number of the decomposition level, which follows from the wavelet packet scheme (Fig. 2). The average powers of the wavelet coefficients in each subband, which are used as speech recognition features, are calculated according to the following expression:

$$\bar{P}_{m,n} = \frac{\sum_{i=n \cdot N/2^m}^{((n+1) \cdot \frac{N}{2^m})-1} (Z_{m,n}(i))^2}{N/2^m}. \qquad (3)$$

To eliminate the sensitivity of the features to changes in the average power of the speech signal realization, the values of $\bar{P}_{m,n}$ obtained by (3) are normalized relative to the average power $\bar{P}_{0,0}$ of the input speech signal realization $S(t_i)$ [27].

Finally, the feature vector $Y = \{y_r\}_R$, consisting of an ordered sequence of averaged powers of wavelet coefficients, is formed by sequentially recording for all m and n the calculated normalized values of $\bar{P}_{m,n}$ from left to right and from top to bottom. The number of the feature r is determined according to the expression $r = 2^m - 1 + n$ and corresponds to the ordinal number of the component element of the vector $Y = \{y_r\}_R$.

An important point in implementing the method is the choice of the scaling function $\varphi(x)$ and the wavelet $\psi(x)$. First, the size of the time-frequency window should be taken into account. Second, the smoothness and symmetry of the underlying wavelet. Third, determine (set) the order of approximation. Correct selection of the wavelet basis for the speech signal significantly reduces the number of non-zero wavelet coefficients $Z_{m,n}(i)$, which significantly reduces the size of the recognition features and makes them much more informative [28].

4. Results and Discussion

Practical experiments were conducted to investigate the contrast of the speech recognition feature vectors formed based on the proposed method. In particular, Figs. 4–5 show the feature vectors of the speech signal calculated in different wavelet decomposition bases.

Thus, in the first case (Fig. 4), a wavelet package based on the Haar basis was used to obtain wavelet coefficients, which provides a relatively coarse approximation of the speech signal, which accordingly affects the informativeness of the recognition features. In the second case (Fig. 5), the speech signal recognition features are calculated based on a smoother Meyer function, which makes the features more informative.

A comparative analysis of the results in Figs. 4–5 shows that when choosing a smoother basis function, the number of y_r values close to zero in the feature vector $Y = \{y_r\}_R$ increases and the informativeness of the decomposition increases, unlike the Haar function, where we get less informative recognition features. Thus, the use of basic wavelet functions consistently in terms of smoothness with the studied speech signal allows us to reduce the size of recognition features and increase their informativeness.

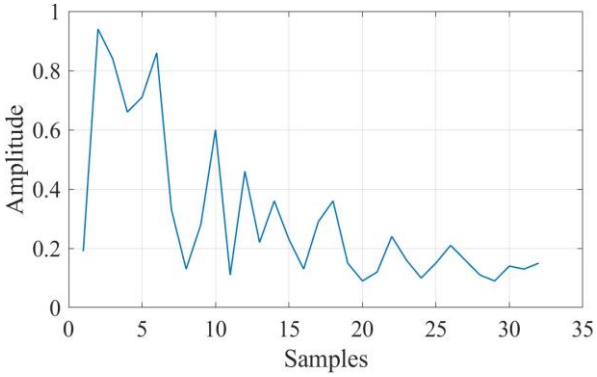


Figure 4: Components of the speech recognition feature vector based on the Haar basis

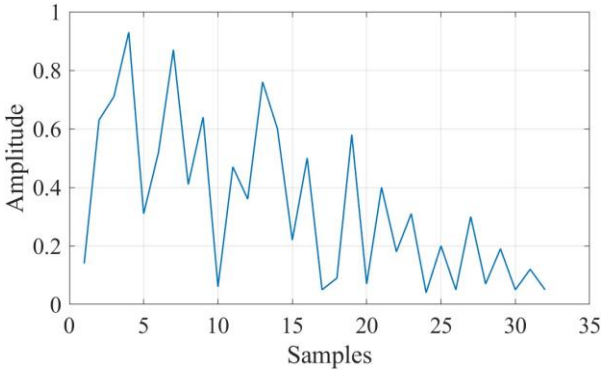


Figure 5: Components of the speech recognition feature vector based on the Meyer basis

To confirm the hypothesis that it is expedient to build speech signal recognition systems based on wavelet packets using the values of $Y = \{y_r\}_R$ obtained by expression (3) as recognition features, we studied the developed method of forming recognition features in comparison with the approach proposed in [15], which is based on the spectral components of the classical harmonic Fourier transform (Fig. 6).

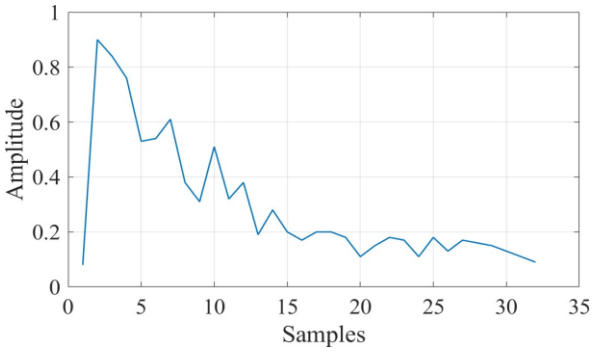


Figure 6: Components of the Fourier-based speech recognition feature vector

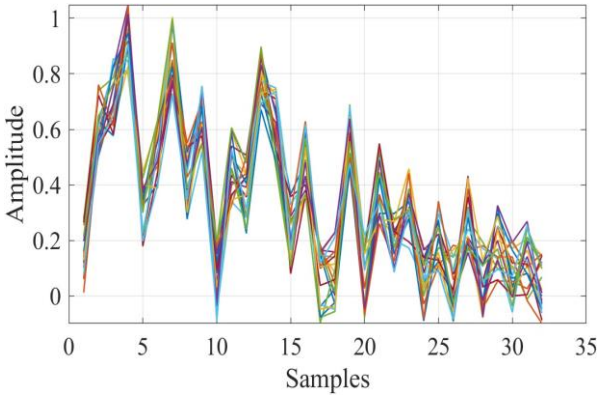
The experiment used realizations of speech signals with a duration of $N = 512$ samples, and the decomposition was performed at $m = 5$ levels of decomposition. This approach allowed us to obtain a feature vector $Y = \{y_r\}_R$ of length $R = 32$, where 16 wavelet coefficients were averaged in each subband. As for the recognition features based on the Fourier transform, the spectrum was divided into 32 bands of 16 coefficients each [29].

To illustrate more clearly the effectiveness of the proposed method (Fig. 7), an experiment was conducted using pre-recorded 30 audio recordings with the same semantic constructions by two different speakers, i.e., the words were pronounced by the speakers: "1", "2", "3", "4", "5" every 30 times. The average value of Root Mean Square Errors (RMSE) will serve as an objective indicator of the effectiveness of the developed method

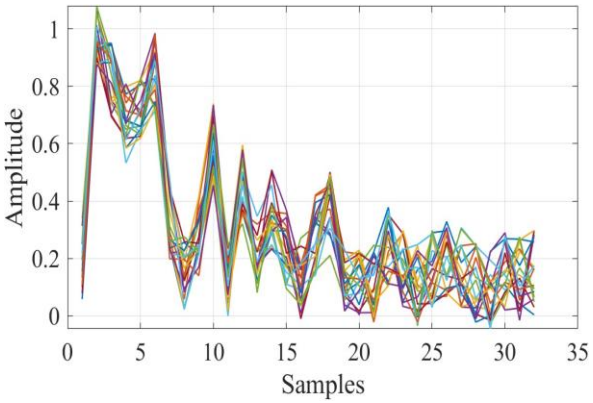
$$\sigma = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{\sum_{t=1}^{n}\left(Y(t)-\hat{Y}(t)\right)^2}{\sum_{t=1}^{n}Y(t)^2}} \to min,$$

for all 30 realizations for each speaker, so the result that shows the lowest RMSE error is the best.

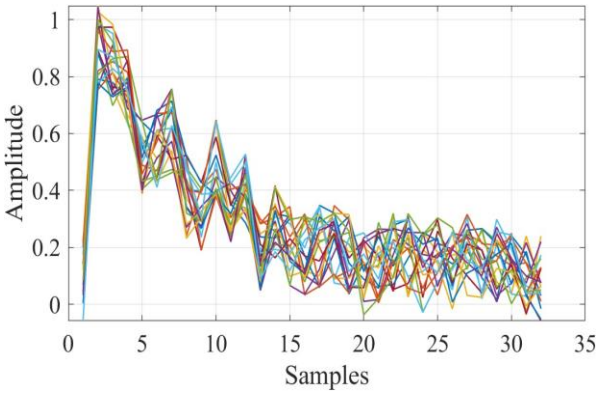
RMSE is one of many metrics that are used to evaluate model performance. To calculate RMSE, square the number of detected errors and find the average value [30].



a



b



c

Figure 7: Thirty implementations of speech signal recognition features using bases: a) Meyer, b) Haar, c) Fourier

The results of the pairwise comparison of the features of the test speech signals obtained using the Haar and Meyer wavelet-based methods and the Fourier spectral coefficient-based method are presented in Table 1.

This experimental study is needed to compute an objective measure of the inter-class distance RMSE of recognition features, i.e., the scatter of features when comparing different realizations of speech signals [31].

Table 1

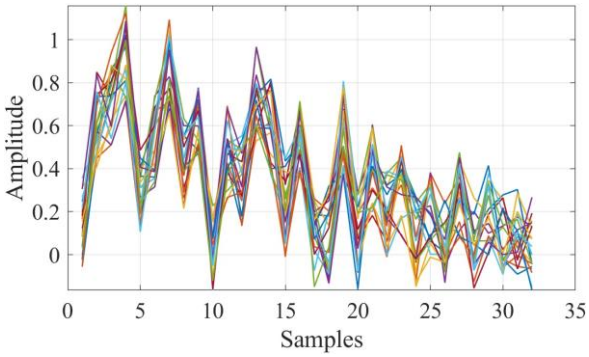
Comparative analysis of the existing and proposed methods

Phrases	Harr, σ	Meyer, σ	Fourier, σ
"1"	0.116	0.053	0.219
"2"	0.153	0.075	0.244
"3"	0.143	0.069	0.231
"4"	0.178	0.081	0.276
"5"	0.162	0.067	0.248

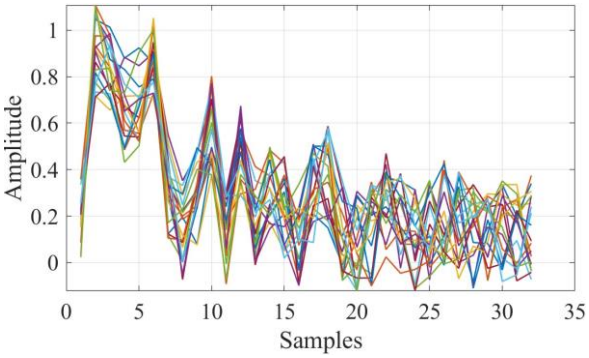
The analysis of the obtained results shows that the contrast of the recognition features of test speech signals generated based on the developed method without the influence of noise is on average 3.8 times higher than that of the method using the Fourier spectral coefficients.

To investigate the effect of noise on the robustness of feature vectors formed based on wavelet packets for the Meyer, Haar basis and based on the Fourier energy spectrum, several experiments were conducted with the addition of white noise with a signal-to-noise ratio of

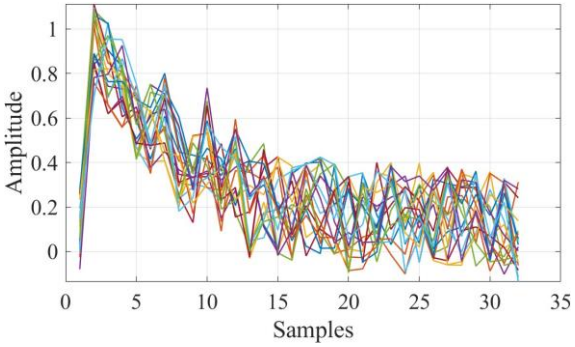
10 dB to the speech signal (noise power was measured in the analysis band) [32]. Fig. 8 shows all three feature vectors with the same noise power.



a



b



c

Figure 8: Thirty realizations of speech recognition features obtained at a signal-to-noise ratio of 10 dB based on bases: (a) Meyer, (b) Haar, and (c) Fourier

Table 2 shows the results of a comparative analysis of the stability of speech signal recognition features obtained from wavelet packets in the Meyer basis and the Fourier energy spectrum. At different signal-to-noise ratios of 10, 20, and 30 dB, the total deviation of the obtained features σ from their reference values was calculated. Then the values were normalized relative to the maximum.

Table 2

Comparative analysis of the existing and proposed methods under the influence of noise of different power

Phrases	Meyer+ noise level of 10 dB, σ	Meyer+ noise level of 20 dB, σ	Meyer + noise level of 30 dB, σ	Fourier + noise level of 10 dB, σ	Fourier + noise level of 20 dB, σ	Fourier + noise level of 30 dB, σ
"1"	0.183	0.119	0.074	0.352	0.281	0.239
"2"	0.231	0.158	0.095	0.382	0.314	0.254
"3"	0.227	0.149	0.087	0.381	0.325	0.261
"4"	0.246	0.161	0.097	0.403	0.347	0.286
"5"	0.214	0.122	0.076	0.367	0.302	0.258

Thus, it was possible to establish that at a signal-to-noise ratio of 10 dB, the features obtained based on the developed method have a very acceptable result, namely, a 1.6-2-fold increase in stability compared to the features obtained based on the traditional Fourier spectrum, where already at a signal-to-noise ratio of 20 dB the total deviation of the obtained features σ is unacceptable.

5. Conclusions and Future Research

In this research, the task of extracting speech signal recognition features for voice identification of a person in a remote mode was solved, which imposes several restrictions, namely: (1) minimum processing time of the speech signal realization, since the

required recognition reliability is achieved by statistical processing of the obtained results; (2) reduction of the dimensionality of recognition features, since the process of extracting recognition features and their classification occurs on the transmitting side of the communication channel, which in turn imposes certain factors of computing power and the influence of noise in the communication channel.

The studies have shown the ability of the developed method to form recognition features based on wavelet packets on the Meyer basis. The most important indicator of the effectiveness of the experiment is the increase in the contrast of recognition features, i.e., the increase in the interclass distance in the formed feature system for speech signals with a similar frequency-temporal structure. Even a visual analysis of the obtained values $Y = \{y_r\}_R$ (Figs. 7–8) reveals significant differences in the structure of the feature vectors formed by relatively short implementations, which proves the potential use of the presented method for speech signal recognition in rapid analysis systems. Since the recognition features are distributed according to normal law, the subsequent procedure for deciding whether speech signal realizations belong to a particular class is greatly simplified.

After analyzing the given conditions of the voice identification system, the question arose of developing a method for extracting speech signal recognition features that would provide more informative spectral characteristics of the speech signal, which would improve the efficiency of their further classification under the influence of noise.

This paper considers the possibility of applying the theory of scale-time analysis to solve this problem, namely, the development of a method for extracting recognition features based on the wavelet packet transform using the orthogonal basis wavelet Meyer function and subsequent averaging of wavelet coefficients that are in the frequency band of the corresponding wavelet packet. Experimental studies have shown the ability of the developed method to generate speech signal recognition features with a close frequency-temporal structure based on wavelet packets in the Meyer basis, namely, it was found that at a signal-to-noise ratio of 10 dB, the features obtained based on the developed method have a very acceptable result, namely, 1.6–2 times more robust to noise than the features obtained based on the traditional Fourier spectrum, where the total deviation of the root mean square error of the obtained features is unacceptable at a signal-to-noise ratio of 20 dB.

Also, the analysis of the results shows that the contrast of the recognition features of test speech signals generated based on the developed method without the influence of noise is on average 3.8 times higher than that of the method using Fourier spectral coefficients.

The authors see the further direction of research in identifying the potential capabilities of the developed method of speech signal recognition in person identification under very difficult conditions of overlapping voices of several speakers, in particular with similar acoustic characteristics, as well as in selecting and justifying the criterion for implementing recognition procedures. It should be noted that there have been virtually no studies of voice identification capabilities for this most difficult case.

References

[1] J. Anand Babu et al., Secure Data Retrieval System using Biometric Identification, International Conference on Data Science and Information System (ICDSIS) (2022) 1–4. doi: 10.1109/ICDSIS55133.2022.9915968.

[2] O. Romanovskyi, et al., Prototyping Methodology of End-to-End Speech Analytics Software, in: 4th International Workshop on Modern Machine Learning Technologies and Data Science, vol. 3312 (2022) 76–86.

[3] I. Iosifov, et al., Transferability Evaluation of Speech Emotion Recognition Between Different Languages, Advances in Computer Science for Engineering and Education 134 (2022) 413–426. doi: 10.1007/978-3-031-04812-8_35

[4] O. Iosifova, et al., Analysis of Automatic Speech Recognition Methods, in: Workshop on Cybersecurity Providing in Information and Telecommunication Systems, vol. 2923 (2021) 252–257.

[5] O. Romanovskyi, et al., Automated Pipeline for Training Dataset Creation from Unlabeled Audios for Automatic Speech Recognition, Advances in Computer Science for Engineering and Education IV, vol. 83 (2021) 25–36. doi: 10.1007/978-3-030-80472-5_3

[6] O. Iosifova, et al., Techniques Comparison for Natural Language Processing, in: 2nd International Workshop on Modern Machine Learning Technologies and Data Science, vol. 2631, no. I (2020) 57–67.

[7] H. Monday et al., Shared Weighted Continuous Wavelet Capsule Network for Electrocardiogram Biometric Identification, 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (2021) 419–425. doi: 10.1109/ICCWAMTIP53232.2021.9674078.

[8] L. Zhu, et al., An Efficient and Privacy-Preserving Biometric Identification Scheme in Cloud Computing, IEEE Access 6 (2018) 19025–19033. doi: 10.1109/ACCESS.2018.2819166.

[9] J. Upadhyay et al., Biometric Identification using Gait Analysis by Deep Learning, Pune Section International Conference (PuneCon) (2020) 152–156. doi: 10.1109/PuneCon50868.2020.9362402.

[10] C. Liu, et al., An Efficient Biometric Identification in Cloud Computing with Enhanced Privacy Security, IEEE Access 7 (2019) 105363–105375. doi: 10.1109/ACCESS.2019.2931881.

[11] O. Attallah, Multi-tasks Biometric System for Personal Identification, International Conference on Computational Science and Engineering (CSE) and International Conference on Embedded and Ubiquitous Computing (EUC) (2019) 110–114. doi: 10.1109/CSE/EUC.2019.00030.

[12] M. Aliaskar et al., Human Voice Identification Based on the Detection of Fundamental Harmonics, 7th International Energy Conference (ENERGYCON) (2022) 1–4. doi: 10.1109/energycon53164.2022.9830471.

[13] R. Kethireddy, et al., Mel-Weighted Single Frequency Filtering Spectrogram for Dialect Identification, IEEE Access 8 (2020) 174871–174879. doi: 10.1109/ACCESS.2020.3020506.

[14] Y. Dong, X. Yang, Affect-Salient Event Sequences Modelling for Continuous Speech Emotion Recognition Using Connectionist Temporal Classification, 5th International Conference on Signal and Image Processing (ICSIP) (2020) 773–778. doi: 10.1109/ICSIP49896.2020.9339383.

[15] R. Hidayat, A. Winursito, Analysis of Amplitude Threshold on Speech Recognition System, International Seminar on Application for Technology of Information and Communication (iSemantic) (2020) 449–453. doi: 10.1109/iSemantic50169.2020.9234214.

[16] Z. Qing, W. Zhong, W. Peng, Research on Speech Emotion Recognition Technology Based on Machine Learning, 7th International Conference on Information Science and Control Engineering (ICISCE) (2020) 1220–1223. doi: 10.1109/ICISCE50968.2020.00247.

[17] B. Kashyap, et al., Machine Learning-Based Scoring System to Predict the Risk and Severity of Ataxic Speech Using Different Speech Tasks, IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2023) 4839–4850. doi: 10.1109/TNSRE.2023.3334718.

[18] H. Park, Y. Chung, J.-H. Kim, Deep Neural Networks-based Classification Methodologies of Speech, Audio and Music, and its Integration for Audio Metadata Tagging, J. Web Eng. 22(1) (2023) 1–26. doi: 10.13052/jwe1540-9589.2211.

[19] O. Lavrynenko, et al., Method of Semantic Coding of Speech Signals based on Empirical Wavelet Transform, 4th International Conference on Advanced Information and Communication Technologies (AICT) (2021) 18–22. doi: 10.1109/AICT52120.2021.9628985.

[20] A. Dutt, P. Gader, Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks, Transactions on Audio, Speech, and Language Proces. 31 (2023) 2043–2054. doi: 10.1109/TASLP.2023.3277291.

[21] C. Zhang, et al., Research on Extracting Algorithm of Speech Eigenvalue Based on Wavelet Packet Transform and Gammatone Filter, 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (2019) 165–169. doi: 10.1109/ITNEC.2019.8729292.

[22] O. Lavrynenko, et al., A Method for Extracting the Semantic Features of Speech Signal Recognition Based on Empirical Wavelet Transform, Radioelectron. Comput. Syst. 107(3) (2023) 101–124. doi: 10.32620/reks.2023.3.09.

[23] G. Frusque, O. Fink, Learnable Wavelet Packet Transform for Data-Adapted Spectrograms, International Conference on Acoustics, Speech and Signal Processing (2022) 3119–3123. doi: 10.1109/ICASSP43922.2022.9747491.

[24] B. Zhao, et al., A Spectrum Adaptive Segmentation Empirical Wavelet Transform for Noisy and Nonstationary Signal Processing, IEEE Access 9 (2021) 106375–106386. doi: 10.1109/ACCESS.2021.3099500.

[25] R. Odarchenko, et al., Empirical Wavelet Transform in Speech Signal Compression Problems, 8th International Conference on Problems of Infocommunications, Science and Technology (2021) 599–602. doi: 10.1109/PICST54195.2021.9772156.

[26] T. Zhang, et al., Multiple Vowels Repair Based on Pitch Extraction and Line Spectrum Pair Feature for Voice Disorder, J. Biomedical Health Inform. 24(7) (2020) 1940–1951. doi: 10.1109/JBHI.2020.2978103.

[27] F. Costa, et al., Wavelet-Based Harmonic Magnitude Measurement in the Presence of Interharmonics, Transactions on Power Delivery 38(3) (2023) 2072–2087. doi: 10.1109/TPWRD.2022.3233583.

[28] X. Zheng, Y. Tang, J. Zhou, A Framework of Adaptive Multiscale Wavelet Decomposition for Signals on Undirected Graphs, Transactions on Signal Proces. 67(7) (2019) 1696–1711. doi: 10.1109/TSP.2019.2896246.

[29] B. Wang, J. Saniie, Massive Ultrasonic Data Compression Using Wavelet Packet Transformation Optimized by Convolutional Autoencoders, Transact. Neural Netw. Learn. Syst. 34(3) (2023) 1395–1405. doi: 10.1109/TNNLS.2021.3105367.

[30] O. Lavrynenko, et al., Remote Voice User Verification System for Access to IoT Services Based on 5G Technologies, 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (2023) 1042–1048. doi: 10.1109/IDAACS58523.2023.10348955.

[31] O. Veselska, et al., A Wavelet-Based Steganographic Method for Text Hiding in an Audio Signal, Sensors 22(15) (2022) 1–25. doi: 10.3390/s22155832.

[32] V. Kuzmin, et al., Empirical Data Approximation Using Three-Dimensional Two-Segmented Regression, 3rd KhPI Week on Advanced Technology (2022) 1–6. doi: 10.1109/KhPIWeek57572.2022.9916335.