UDC 004.896

ALGORITHMS FOR SPEAKER RECOGNITION BASED ON WAVELET TRANSFORM AND MACHINE LEARNING IN SPEECH SIGNAL ANALYSIS

[1+]*Urinboev J.K,* [2]*Mirzayeva G.R,* [1]*Aripova S.O.*

jahongir8010@gmail.com
[1] Digital Technologies and Artificial Intelligence Development Research Institute, Tashkent, Uzbekistan.
[2] TUIT named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

Annotation: In this work, features were isolated from the speech signal using Haar, Daubechies, and Morlet wavelet transforms, which were subsequently used for speaker recognition in SVM and LSTM models. The research was conducted on the basis of a dataset of 100 speakers. The obtained results made it possible to analyze the effectiveness between wavelet types and machine learning models.

Keywords: speech signal processing, wavelet transform, Haar wavelet, Daubechies wavelet, Morlet wavelet, SVM, LSTM, speaker recognition.

1 Introduction

Speech identification is one of the current areas of biometric authentication and is widely used in security systems, banking, and personal identification systems. While traditional methods are based on spectral and acoustic analysis, modern approaches include the use of wavelet transforms and machine learning models for in-depth study of the signal [1].

Wavelet transforms allow analyzing the signal over time and frequency. Haar, Daubechies (db4), and Morlet wavelets are capable of detecting various smoothness and frequency changes and are used to extract important features from speech signals, such as energy, entropy, amplitude, and statistical parameters. These features are subsequently fed into machine learning models, including SVM and LSTM models.

While the SVM model is a classifier that allows linear and nonlinear separation in high-dimensional space, the LSTM, as a deep learning network, has the property of preserving relationships over time. In this work, the process of feature separation using wavelet transforms and their effectiveness in speaker recognition in SVM and LSTM models are analyzed [2].

2 Methods

2.1 Problem Statement

Speech recognition technologies are currently of great importance in the field of security and biometric authentication. In this research, the following problems are studied:

• Determine which wavelet substitution is most suitable for extracting effective symbols from a speech signal.

• Comparison of the effectiveness of the SVM and LSTM models in speech recognition.

• Assessment of the influence of various wavelet transforms on the results of model operation.

During the research, based on the Haar, Daubechies, and Morlet wavelet transforms, features are isolated and their results are analyzed. Also, these methods are used in conjunction with various classifiers, and the best combination is determined [3].

2.2 Data Collection and Preprocessing

A dataset of 100 speakers was used for the research. This dataset contains several speech recordings for each speaker. The stages of speech signal preprocessing were carried out as follows:

• Noise removal: Excess background noise in the data was removed using spectral subtraction and low-pass filters.

• Normalization: The signal amplitudes were brought to the same range, which helped reduce the differences between different speakers.

• Segmentation: Each speech recording was divided into short segments, each of which was processed within a specified length (for example, 20-30 ms).

• Voice Activity Detection (VAD): Only the real speech part has been removed, and the silence or unnecessary background parts have been removed.

2.3 Identification of features.

The following features were isolated from the speech signal using Haar, Daubechies, and Morlet wavelet transforms:

• Energy
• Entropy
• Maximum and minimum amplitudes
• Average frequency
• Statistical parameters

The Haar wavelet is one of the simplest and most understandable, introduced by Alfred Haar in 1909. This wavelet is orthogonal and has a filter with compact application. The main advantage of the Haar wavelet is its fast and efficient operation. It represents a square signal and is ideal for processing fast signals. The Haar wavelet is used to detect sharp changes, but it is weak in terms of smoothness. This wavelet is mainly used for detecting abrupt changes in the speech signal, data compression, and voice recognition in biometric systems [4].

Daubechies wavelets were developed by Ingrid Daubechies in 1988 and are smoother and more accurate than the Haar wavelet. This wavelet family includes models of different levels: db1, db2, ..., dbN (where N is the wavelength level). The main feature of Daubechies wavelets is their high frequency accuracy and the ability to represent smooth functions. They are especially widely used for speech signal processing, voice-based biometric systems, and analysis of medical signals (e.g., EEG and ECG signals). All dbN wavelets are smoother and have a higher degree of differentiation. Although these wavelets are convenient for in-depth signal analysis and compression, the computational complexity is high [5].

The Morlet wavelet, on the other hand, is formed by a combination of a sinusoid and a Gaussian function and is mainly used for frequency analysis. This wavelet is ideal for continuous wavelet substitution and has good accuracy in time and frequency domains. A Morlet wavelet is a complex wavelet with real and imaginary parts, which works effectively when separating different frequencies. The main areas of application of this wavelet are the separation of phonemes and sounds in speech signals, data analysis (for example, ECG and EEG signals), and signal processing in radar and sonar systems. The Morlet wavelet is especially used in situations requiring periodic frequency analysis. However, it is not orthogonal and does not provide full flexibility in the time domain [6].

1. Table. Wavlet comparison table

Wavelet	Features	Benefits	Disadvantages
Hair	Square orthogonal wavelet	Quick, simple, effective	No smoothness, weak detailing
Daubechies	Family of smooth and orthogonal wavelets	High resolution, strong compression, and signal separation	High computational complexity

Morlet	Sinusoidal and Gaussian combination	Ideal for frequency analysis	Not orthogonal, not fully flexible in time domain

Energy. The sum of the squares of the coefficients of the speech signal, obtained as a result of wavelet transform, represents the signal energy. Energy is used to determine the power and intensity of the signal [7].

$$E = \sum_{i=1}^{N} |C_i|^2 \tag{1}$$

Entropy. Entropy measures the degree of randomness of a signal and determines the degree of order of a signal from the perspective of information theory. The high-entropy signal is more random and uncertain, while the low-entropy signal is more ordered [8].

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{2}$$

where p_i is the relative probability of wavelet coefficients.

Maximum and Minimum Amplitudes. Maximum and minimum amplitudes are determined by finding the largest and smallest values among the wavelet coefficients of the speech signal [7].

$$A_{max} = max\left(|C_i|\right) \tag{3}$$

$$A_{min} = min\left(|C_i|\right) \tag{4}$$

These features help to assess the intensity and voltage level of the signal. For example, you can measure the difference between words spoken aloud and those spoken softly.

Average Frequency. The average frequency represents the value of the signal that dominates in the frequency spectrum. This will help determine at which frequencies the signal contains more energy [8].

$$f_{\text{avg}} = \frac{\sum_{i=1}^{N} f_i |C_i|}{\sum_{i=1}^{N} |C_i|} \tag{5}$$

where f_i is the corresponding frequency of each coefficient.

Statistical parameters. The following statistical parameters are calculated based on the Wavelet coefficient distribution:

- **Mean** (μ): represents the overall trend of the signal.

$$\mu = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{6}$$

- **Dispersion** (σ^2): measures the degree of signal propagation.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left(C_i - \mu\right)^2 \tag{7}$$

- **RMS** : used to estimate signal energy.

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^{N} C_i^2} \tag{8}$$

- **Skewness**: determines whether the signal is symmetrical or asymmetrical.

$$\text{Skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(C_i - \mu)^3}{\sigma^3} \tag{9}$$

- **Kurtosis**: Evaluates how the signal distribution ends are distributed.

$$\text{Kutrosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(C_i - \mu)^4}{\sigma^4} \tag{10}$$

These parameters are used to analyze the overall dynamics, shape, and variability of the signal [9].

With the help of energy, it is possible to analyze the loudness of the speech signal, the level of stress, and the differences in different segments. In a speech signal, entropy is used to determine the degree of voice expression, the presence of emotions (such as joy, anger, or laziness), or noise. In a speech signal, the average frequency is used for voice intonation, phonetic features of language, and individual personality recognition [10].

2.4 Speaker recognition using the SVM model

Feature extraction. Speech signal analyzed using three different wavelet transforms:

- **Haar Wavelet Transform.** Discrete wave analysis of the signal is performed using Haar wavelets. Choose level for Wavelet decomposition:

$$C_j, D_j = \text{wavedec}(y, 'haar', L = 4)$$

where C_j - low-frequency coefficients, D_j - high-frequency detailing coefficients [11]. To calculate the Haar wavelet symbols, the sum of energies at each level is taken:

$$F_{haar} = \sum_{j=1}^{L}\sum_{k}\left|C_j[k]\right|^2 \tag{11}$$

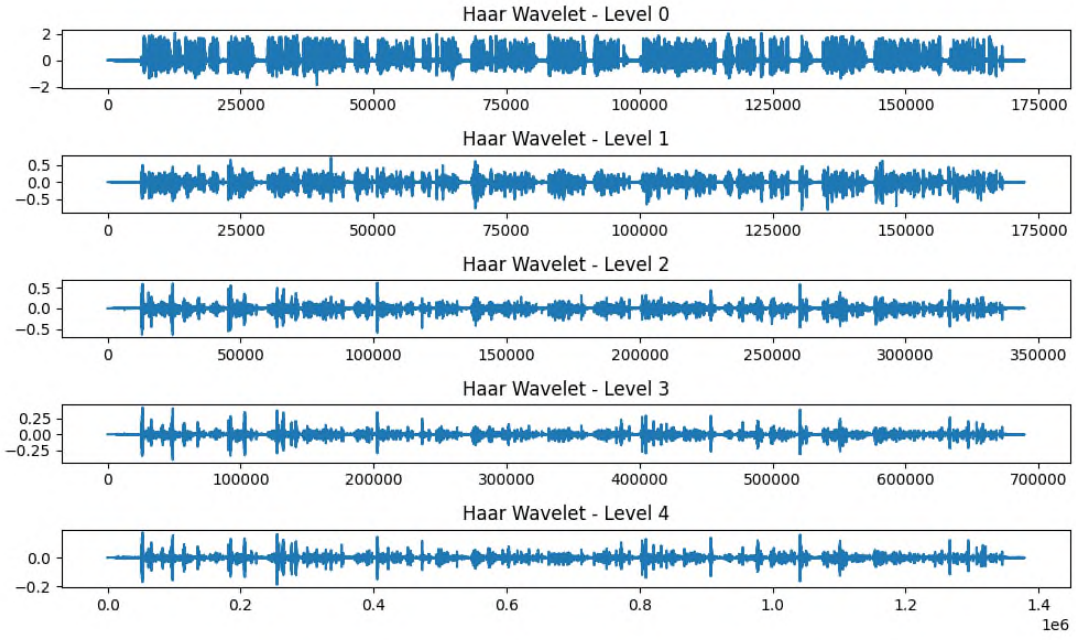


Fig 1. Multi-level decomposition of speech signals using Haar Wavelet

- **Daubechies wavelet transform.** Daubechies 4 (db4) wavelet transform will also be performed as above:

$$C_j^{'}, D_j^{'} = \text{wavedec}\left(y, 'db4', L = 4\right)$$

where the average value for each component is taken [2]:

$$F_{db} = \left[\frac{1}{N}\sum_k C_j^{'}[k]\right] \tag{12}$$

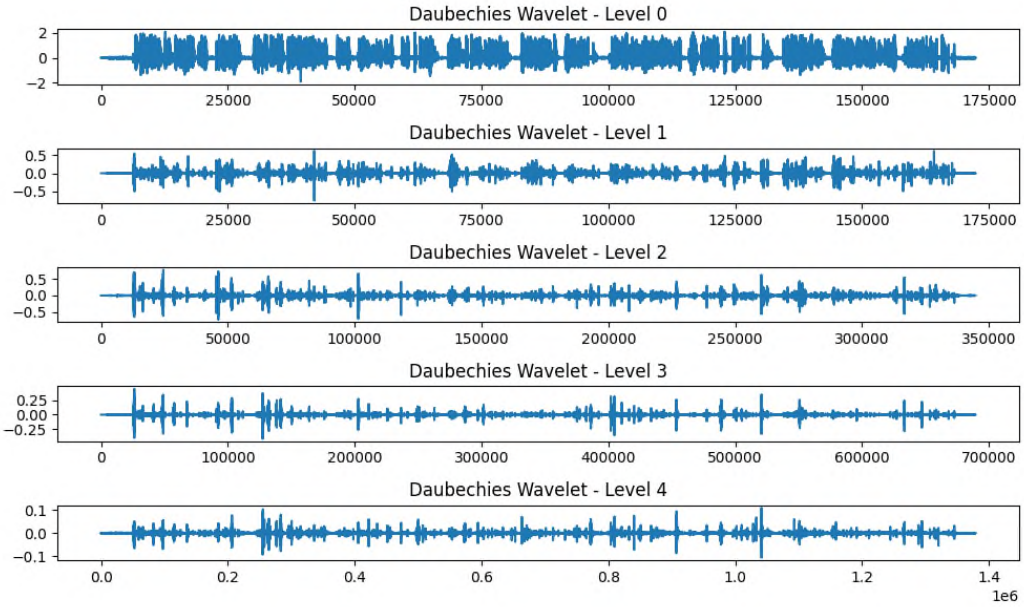


Fig 2. Multi-level decomposition of speech signal using Daubechies Wavelet

- **Morlet Wavelet Transform (CWT).** CWT is calculated using the Morlet wavelet as follows:

$$C(a,b) = \int_{-\infty}^{\infty} y(t)\psi^*\left(\frac{t-b}{a}\right)dt \tag{13}$$

where a - scale parameter, b - shift parameter, ψ^* -Morlet wavelet function [13].

Features are extracted using mean absolute values:

$$F_{morlet} = \left[\frac{1}{M}\sum_m \left|C(a_m,b)\right|\right] \tag{14}$$

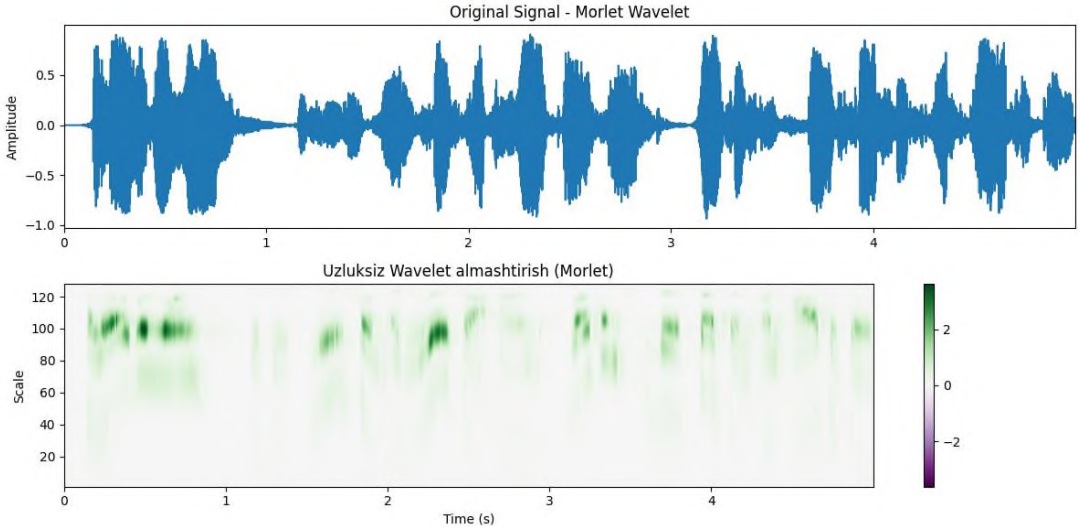


Fig 3. Frequency representation of a speech signal using Morlet wavelet transform

Finite feature vector. The features obtained through three different wavelet transformations are combined into a single vector:

$$F = \left[F_{haar}, F_{db}, F_{morlet}\right] \tag{15}$$

Data preparation and standardization. The collected feature matrix F is X, and the corresponding personal identifiers are y. We divide these into train-test sets through stratified splitting:

$$X_{train}, X_{test}, y_{train}, y_{test} = \text{StratifiedShuffleSplit}\left(X, y\right)$$

Then standardization is carried out:

$$X' = \frac{X - \mu}{\sigma} \tag{16}$$

where μ is the mean, σ is the variance.

Recognition. The SVM model solves the following optimization problem:

$$\min_{\text{w},b} \frac{1}{2}\left\|\text{w}\right\|^2 + C\sum_{i=1}^{N} max\left(0, 1 - y_i\left(\text{w} \cdot X_i + b\right)\right) \tag{17}$$

where w - normal vector of the hyperplane, b - offset, C - regularization parameter, y_i - classification label, X_i - feature vector [14].

The SVM model is optimized via GridSearchCV, where C, gamma, and kernel parameters are selected:

$$\left(C, \gamma, kernel\right) \in \left\{\left(0.1, 1, 10\right)\right\} \times \left\{'scale', 'auto' \times \left\{'rbf', 'poly'\right\}\right.$$

Accuracy is evaluated according to the final best model:

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{all predictions}}$$

2.5 Speaker recognition using the LSTM model

Feature extraction. Speech signal $y\left(t\right)$ Analyzed using Haar and Daubechies (db4) wavelet transforms.

Discrete wavelet transform (DWT). Discrete wavelet transform is performed as follows [15]:

$$C_j, D_j = \text{wavedec}\left(y, \psi, L = 5\right)$$

where ψ is the selected wavelet (Haar or Daubechies). $L = 5L$ - degree of decomposition. C_j - low-frequency (approximation) component. D_j - high-frequency (detailed) components.

The following statistical features are extracted from each component:

- Average value (energy) Let's rewrite expression (6) as follows:

$$F_{mean} = \frac{1}{N}\sum_k C_j\left[k\right]$$

- Dispersion (standard deviation) (signal propagation rate) Let's express expression (7) in the following form:

$$F_{std} = \sqrt{\frac{1}{N}\sum_k \left(C_j\left[k\right] - F_{mean}\right)^2}$$

In the end, according to the Haar and Daubechies wavelets for each file:

$$F = \left[F_{mean,haar}, F_{std,\ haar}, F_{mean,db4}, F_{std,db4}\right] \tag{18}$$

a feature vector is generated [16].

Data Standardization. Each feature is normalized by expression (16). Also, the number of speakers is N , which are indexed using Label Encoder as follows: $y \in \{0,1,\ldots,N-1\}$

Preparation of data for the LSTM model. Since LSTM is a sequence-to-sequence model, it should be in the form of symbols (num_samples, timesteps, features) [17]. Therefore, each example X is reduced to the form (1, n_features) as a one-dimensional sequence:

$$X_{train} = reshape\left(X_{train}, \left(N, 1, features\right)\right)$$

Mathematical description of the LSTM model. LSTM works at each time as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, X_t\right] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, X_t\right] + b_i\right)$$
$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, X_t\right] + b_o\right)$$
$$\tilde{C}_t = tanh\left(W_C \cdot \left[h_{t-1}, X_t\right] + b_C\right)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$
$$h_t = o_t \odot \tanh\left(C_t\right)$$

where f_t - level of forgetting, i_t - level of receiving new information, o_t - level of output, C_t - memory vector, h_t - hidden state, W_f, W_i, W_o, W_C - LSTM weights, b_f, b_i, b_o, b_C bias.

LSTM Architecture.

1. The 128-dimensional LSTM layer (return_sequences=True) works with 128 LSTM blocks, since return_sequences=True, hidden states are returned at all time steps. The result is transmitted to the next LSTM layer [18].

$$h_t^{(1)} = LSTM\left(X_t, 128\right) \tag{19}$$

2. Dropout (0.3) - preventing overfitting is a random zeroing of 30% of neurons. This prevents excessive adaptation of the model.

3. The 64-dimensional LSTM layer works with 64 LSTM blocks, since return_sequences=False, only the hidden state of the last time step is obtained.

$$h_t^{(2)} = LSTM\left(h_t^{(1)}, 64\right) \tag{20}$$

4. Again Dropout (0.3) - again prevent overfitting. 30% of neurons are randomly switched to zero.

5. The Softmax activation function is used to identify the Softmax output layer speaker. The probability distribution for each class will be displayed:

$$y_{pred} = softmax\left(W_0 \cdot h_t^{(2)} + b_0\right) \tag{21}$$

6. The categorical cross-entropy loss function is used to optimize the model:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) \tag{22}$$

where y_i is the true sign, $\hat{y}_i$ is the predicted probable sign.

Model training and evaluation. Model Adam is trained with an optimizer for 50 epochs [19]. In the test, the highest probability with $argmax$ is selected:

$$y_{pred} = \arg\max_k P(y = k \mid X_{test}) \tag{23}$$

The accuracy of the model is evaluated with accuracy:

$$accuracy = \frac{\sum_{i=1}^{m} 1\left(\hat{y}_i = y_i\right)}{m} \tag{24}$$

where $1(\hat{y}_i = y_i)$ is an indicator function that takes a value of 1 if the actual and predicted values are equal, and 0 if not.

3 Results

Energy, entropy, maximum and minimum amplitudes, average frequency, and statistical parameters isolated from speech signals using Haar, Daubechies, and Morlet wavelets are important features for speech analysis and recognition.

The Haar wavelet is primarily used for detecting sharp changes in a signal, while the Daubechies wavelet is suitable for extracting smooth and stable features. The Morlet wavelet is used for in-depth analysis of the frequency spectrum.

These symbols can be used to recognize a person through speech, extract words, identify emotions, and create speech-based biometric systems.[20]

According to the results of the experiment:

- Haar wavelet transform showed the best results with LSTM.
- Morlet wavelet transform achieved high accuracy with SVM.
- The Daubechies wavelet transform yielded average results with both models.

Table 2. Compare between models

Feature	SVM	LSTM
Computing resources	Requires less	Requires high
Learning ability	Can distinguish between linear and nonlinear boundaries	Understands time-dependence well
Accuracy	94%	92.5%
Training time	Relatively fast	Slower
Working with data balance	It gives good results when the features are well separated.	Can better learn even inaccurate information
Optimal parameters	C=10, gamma='scale', kernel='rbf'	128-64 LSTM layers, human optimizer, softmax output
Used wavelets	Haar, Daubechies (db4), Morlet	Haar, Daubechies (db4)

These results showed that the Haar wavelet transform works well with SVM, while LSTM shows better results for time-dependent symbols.

4 Conclusion

In this research, the effectiveness of the Haar, Daubechies, and Morlet wavelet transforms in isolating features from a speech signal was evaluated. The results show that the SVM model works quickly and effectively on a small set of symbols, but a decrease in accuracy is observed in complex speech signals. The LSTM model provides high accuracy in the analysis of long speech segments, since it takes into account temporal dependencies.

In the future, it is possible to increase the accuracy of these methods by implementing them in real-time systems and adding other deep learning models. Also, testing the stability of the method in low-resource languages and noisy environments remains one of the important directions.

References

[1] Heidari M., Shateyi S. Wavelet support vector machine and multi-layer perceptron neural network with continuous wavelet transform for fault diagnosis of gearboxes // Journal of Vibroengineering. - 2017. - Vol. 19, No. 1. - Б. 107-110.
[2] Elfouly F.H., Mahmoud M.I., Dessouky M.I.M., Deyab S. Comparison between Haar and Daubechies wavelet transforms on FPGA technology // International Journal of Computing. - 2007. - Vol. 6, No. 3. - P. 23-29.
[3] Nazimov A.I., Pavlov A.N., Khramov A.E. Method based on adaptive wavelet transform for recognizing characteristic oscillatory models // Journal of Communication and Electronics Technologies. - 2013. - Vol. 58, No. 8. - P. 790-795.
[4] Kankar P.K., Sharma S.C., Harsha S.P. Fault diagnosis of ball bearings using continuous wavelet transform and artificial neural network // Procedia Engineering. - 2010. - Vol. 2, No. 1. - P. 2917-2926.
[5] Rafiee J., Rafiee M.A., Tse P.W., et al. A novel technique for selecting the mother wavelet function using an intelligent fault diagnosis system // Expert Systems with Applications. - 2011. - Vol. 38, No. 5. - P. 6351-6361.
[6] Kankar P.K., Sharma S.C., Harsha S.P. Rolling element bearing fault diagnosis using wavelet transform and support vector machine // Applied Soft Computing. - 2011. - Vol. 11, No. 2. - P. 2300-2312.
[7] Saravanan N., Ramachandran K.I. Incipient gearbox fault diagnosis using discrete wavelet transform (DWT) for feature extraction and classification using artificial neural network (ANN) // Expert Systems with Applications. - 2010. - Vol. 37, No. 6. - Б. 131-143.
[8] Nikolaou N.G., Antoniadis I.A. Rolling element bearing fault diagnosis using wavelet packets // NDT & E International. - 2002. - Vol. 35, No. 3. - Б. 197-205.
[9] Prabhakar S., Mohanty A.R., Sekhar A.S. Application of discrete wavelet transform for detection of ball bearing race faults // Tribology International. - 2002. - Vol. 35, No. 12. - P. 793-800.
[10] Purushotham V., Narayanan S., Prasad S. Multi-fault diagnosis of rolling bearing elements using wavelet analysis and hidden Markov model based fault recognition // NDT & E International. - 2005. - Vol. 38, No. 8. - Б. 654-664.
[11] Chen Z., Li C., Sanchez R.V., et al. Gearbox fault identification and classification with convolutional neural networks // Shock and Vibration. - 2015. - Vol. 2015. - Article 390134.
[12] Zhao M., Jia L., Lin J., et al. Rolling bearing fault diagnosis based on wavelet packet transform and hybrid support vector machines // Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. - 2011. - Vol. 225, No. 8. - P. 269-271.
[13] Jedliński Ł., Jonak J. Application of the wavelet transform and probabilistic neural networks in IC engine fault diagnostics // Eksploatacja i Niezawodnosc - Maintenance and Reliability. - 2015. - Vol. 17, No. 1. - P. 64-69.
[14] Kumar R., Kumar H. Bearing fault diagnosis using wavelet transform and artificial neural network algorithm // International Journal of Mechanical Engineering and Robotics Research. - 2014. - Vol. 3, No. 1. - P. 57-66.
[15] Abbasion S., Rafsanjani A., Farshidianfar A., et al. Rolling element bearings multifault classification based on the wavelet denoising and support vector machine // Mechanical Systems and Signal Processing. - 2007. - Vol. 21, No. 7. - P. 2933-294.
[16] Carlos B., Emily M.P. Speech emotion recognition using deep neural networks // IEEE Transactions on Affective Computing. - 2021. - No. 12. - P. 567-578.
[17] Ozerov A.V., Gusev A.A., Ivanov V.V. Methods and algorithms for processing speech signals in automatic speech recognition systems // Bulletin of MGTU named after N. E. Bauman. - 2022. - No 4. - C. 98-112.
[18] Deshpande M., Holambe R. Speaker identification using admissible wavelet packet-based decomposition // International Journal of Electrical and Computer Engineering. - 2010. - Vol. 4, No. 1. - P. 83-86.

NUTQ SIGNALLARINI TAHLIL QILISHDA VEYVLET ALMASHTIRISH VA MASHINAVIY O‘QITISH ASOSIDA SHAXSNI TANIB OLISH ALGORITMLARI

O‘rinboyev J.K, Mirzayeva G.R, Aripova S.O.

jahongir8010@gmail.com
Raqamli texnologiyalar va sun’iy intellektni rivojlantirish ilmiy-tadqiqot instituti, Toshkent, O‘zbekistan.

Annotatsiya: Mazkur ishda nutq signalidan Haar, Daubechies va Morlet veyvlet almashtirishlari yordamida belgilar ajratib olinib, keyinchalik SVM hamda LSTM modellarida diktorni tanib olish uchun ishlatilgan. Tadqiqot 100 ta diktordan iborat dataset asosida amalga oshirildi. Olingan natijalar veyvlet turlari va mashinali o‘qitish modellar orasidagi samaradorlikni tahlil qilish imkonini berdi.

Keywords: nutq signallarini qayta ishlash, veyvlet almashtirishlari, Haar veyvlet, Daubechies veyvlet, Morlet veyvlet, SVM, LSTM, diktorni tanib olish.

АЛГОРИТМЫ РАСПОЗНАВАНИЯ ЛИЧНОСТИ НА ОСНОВЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ И МАШИННОГО ОБУЧЕНИЯ ПРИ АНАЛИЗЕ РЕЧЕВЫХ СИГНАЛОВ

Уринбоев Ж.К, Мирзаева Г.Р, Арипова С.О.

jahongir8010@gmail.com
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта, Ташкент, Узбекистан.

Аннотация: В данной работе из речевого сигнала были извлечены признаки с помощью вейвлет-преобразований Хаара, Добеши и Морле, которые затем были использованы для распознавания диктора в моделях SVM и LSTM. Исследование проводилось на основе набора данных, состоящего из 100 дикторов. Полученные результаты позволили проанализировать эффективность различных типов вейвлетов и моделей машинного обучения.

Ключевые слова: обработка речевых сигналов, вейвлет-преобразования, вейвлет Haar, вейвлет Daubechies, вейвлет Morlet, SVM, LSTM, распознавание диктора.