UDC 004.934:57.087.1

DOI: 10.15587/1729-4061.2025.347451

DEVELOPMENT AND INCREASE OF NOISE IMMUNITY OF A MODEL OF BIOMETRIC IDENTIFICATION OF A SPEAKER BASED ON METAL-FREQUENCY CEPSTRAL COEFFICIENTS AND A CONVOLUTIONAL NEURAL NETWORK

Muhabbat Khizirova
Candidate of Physico-Mathematical Sciences, Associate Professor*

Katipa Chezhimbayeva
Correspondent author
Candidate of Technical Sciences, Professor*
E-mail: k.chezhimbayeva@aues.kz

Abdurazak Kassimov
Candidate of Technical Sciences,
Associate Professor, Professor-Lecturer*

Muratbek Yermekbaev
PhD, Associate Professor
Department of Telecommunications Engineering*

Assiya Iskakova
Master of Technical Sciences, PhD-Student*

Zhaina Abilkaiyr
Science Degree Master*

*Department of Telecommunications Engineering
Almaty University of Power Engineering and Telecommunications
named after Gumarbek Daukeyev
Baitursynuly str., 126/1, Almaty, Republic of Kazakhstan, 050013

This study is focused on improving the noise robustness of a biometric speaker identification system based on mel-frequency cepstral coefficients (MFCC) and a convolutional neural network (CNN). The object of analysis is the acoustic structure of the Kazakh language under clean and noisy conditions. The experimental database consisted of 16 speakers, each represented by 12 audio recordings with a duration of approximately 1 s. The speech signals were corrupted by additive pink noise with different signal-to-noise ratio (SNR) levels.

Under clean signal conditions, the CNN-based classifier achieved a high recognition accuracy of approximately 96%, as confirmed by the confusion matrix with strong diagonal dominance. When exposed to noise, the classification accuracy decreased to about 69%, demonstrating the significant impact of acoustic interference on speaker identification performance. To improve noise immunity, noise augmentation was applied during training. After retraining on the augmented dataset, the classification accuracy under noisy conditions increased to approximately 89–90%.

The heatmaps of precision, recall, and F1-score demonstrate that after robustness enhancement, most speaker classes achieve stable metric values in the range of 0.85–1.00, while the averaged performance metrics reach accuracy ≈ 0.89–0.90, confirming consistent recognition across the entire dataset. The results show that MFCC features retain discriminative speaker-specific spectral characteristics even under noise and that CNN-based classification significantly outperforms traditional approaches in terms of robustness.

The proposed MFCC–CNN approach provides high identification accuracy in clean environments and maintains competitive performance under noise after data augmentation. The obtained results confirm the practical applicability of the developed system for reliable speaker verification in acoustically unstable environments, including remote biometric authentication, access control, and intelligent communication systems

Keywords: speaker identification, voice biometrics, Kazakh speech, mel-frequency cepstral coefficients, noise

Received 01.10.2025
Received in revised form 17.11.2025
Accepted date 15.12.2025
Published date 30.12.2025

***How to Cite:** Khizirova, M., Chezhimbayeva, K., Kassimov, A., Yermekbaev, M., Iskakova, A., Abilkaiyr, Z. (2025). Development and increase of noise immunity of a model of biometric identification of a speaker based on metal-frequency cepstral coefficients and a convolutional neural network. Eastern-European Journal of Enterprise Technologies, 6 (9 (138)), 37–53. https://doi.org/10.15587/1729-4061.2025.347451*

1. Introduction

Globally, and in Kazakhstan, it is a period of rapid development of digital services, such as automated control systems, remote communication platforms, and intelligent technologies, significantly increasing the requirements for reliable identity verification. In practice, all sectors – for example, banking, education, telecommunications, medicine, defense, and the electronic circulation of government documents – require highly accurate user identification that ensures a stable level of security with minimal delays and ease of use. The biometric methods used in this study stand out for their reliability, as they rely on intrinsic physiological or behavioral characteristics of a person, rather than information that is easily forgotten or lost.

When it comes to biometric technologies, voice identification occupies a special place, and here's why. Voice is the most natural way we communicate with technology, so using it to verify identity is very simple, without any touching or effort. Furthermore, our voice is more than just sound. Voice biometrics is highly complex and depends on the structure of our vocal apparatus, the way we speak, and our habits. All of this makes our voices highly individual and difficult to counterfeit. Therefore, studying how to recognize who is speaking is a very important and interesting task, one that offers practical benefits and poses interesting questions for scientists.

Voice biometrics is a field fraught with challenges. Unlike static biometric data such as fingerprints or vein patterns, voice is a dynamic signal susceptible to significant distortion. These distortions can be caused by external noise, the quality of communication channels, the speaker's emotional state, the speaking rate, the type of microphone, the sounds being spoken, and the recording conditions. Therefore, achieving reliable speaker recognition requires in-depth research in areas such as digital signal processing, statistical modeling, machine learning, and neural network development. Ensuring noise immunity is particularly challenging, as real-world usage scenarios (e.g., recording in public transportation, outdoors, in echo-intensive rooms, or in the presence of other voices) differ radically from ideal laboratory conditions. Studying linguistic features remains an important area of research, as the acoustic and phonetic characteristics of a language directly influence the structure of extracted features. For example, the Kazakh language, with its developed vowel harmony system, specific articulation techniques, and unique sound organization, requires a tailored approach to the methods used. To improve recognition accuracy, it is critical to consider the influence of language phonological patterns on the calculation of Mel-frequency cepstral coefficients and other spectral characteristics.

In recent years, the importance of neural networks in speaker recognition has grown significantly. Deep models demonstrate superiority in revealing hidden structures in multivariate data and modeling complex relationships, which goes beyond the capabilities of classical statistical approaches. However, creating reliable and robust systems requires a thorough preliminary analysis of the behavioral characteristics of features, the impact of acoustic interference, and the development of effective methods for regularizing and adapting neural networks to the specifics of speech material. Consequently, the development and improvement of voice biometrics systems is based on a number of unresolved scientific problems, confirming the continued relevance of this topic. Given the rapid development of intelligent devices, smart home systems, voice assistants, remote services, as well as secure authentication methods and multi-factor security systems in Kazakhstan, voice biometrics is becoming a critical element of future digital ecosystems. Moreover, this scientific field forms the foundation for the development of adaptive and universal security algorithms capable of operating effectively in a wide range of acoustic conditions and taking into account the linguistic and typological features of speech signals. Therefore, research aimed at developing methods for extracting voice characteristics, analyzing the influence of noise, and constructing reliable neural models for speaker recognition is a relevant and scientifically significant task.

2. Literature review and problem statement

Research on filtering and spectral analysis of speech signals based on the segmentation method is discussed in [1]. It is established that the procedures cover the entire digital speech processing system. However, under changing recording conditions, this method is considered ineffective in terms of stability, as it must take into account the dynamics of transmission channel parameters, microphone characteristics, and background noise. To date, universal algorithms for solving these problems have not been fully explored, but adaptive preprocessing procedures do not fully address the identified limitations.

The paper [2] examines the most recent biometric method for voice signal recognition, which is determined by the information content of the extracted features and various classifier parameters. However, the problem of low feature immunity to reverberation, noise, and channel distortion remains, due to the limited ability of traditional filters to compensate for speech signal nonlinearity.

Improving the noise immunity of identification systems is discussed in [3, 4]. It has been established that even the most advanced models significantly lose accuracy with deteriorating signal-to-noise ratios. The main challenge remains the scalability of these solutions: training huge balanced databases with different types of noise. Their creation incurs the most significant costs associated with collecting, processing, and storing acoustic data. As is discussed above [3], adaptive feature optimization is the most effective method, but its effectiveness decreases under these conditions. Mel-frequency cepstral coefficients (MFCCs) used for feature extraction in speech recognition systems using a convolutional neural network (CNN) architecture are discussed in [5]. These studies demonstrate that this approach achieves the required recognition accuracy in acoustic conditions. At the same time, this method remains sensitive to noise and spectral distortions.

The evolution from traditional i-vector approaches to modern x-vector embeddings is observed in [6], where the emphasis is on creating more robust and scalable architectures for working with different recording conditions. Analytical methods provide statistical tools for testing the stability of such systems, which is critical in noisy or heterogeneous acoustic conditions [7].

In [8], a classification model based on an artificial neural network (ANN) using low-frequency MFCC features is considered. It performs well in a clean environment, but in noisy environments, its performance decreases due to the strong dependence of cepstral characteristics on the spectral profile of the signal.

In [9], significant progress was made: neural network methods were proposed that outperform traditional i-vector models in terms of robustness to speech signal variations. However, the impact of channel distortions and background noise on the quality of representations remains a challenge due to the lack of universal methods for adapting to different recording conditions. Integrating deep networks with classical spectral approaches could improve the situation, but this area has only been partially explored so far.

Hybrid approaches from [10] demonstrate that combining spectral and statistical features improves classification quality. However, some unsolved problems remain, including high computational loads. This, in turn, makes them unsuitable for real-time and mobile designs.

The fundamental principles of digital filtering described in [11] highlight the limited flexibility of traditional algorithms. This is emphasized by the need to create universal filters for the nonlinear and nonstationary nature of speech signals. This is achieved using flexible signal processing procedures for biometric tasks.

The importance of the persistence problem is demonstrated in studies of related areas. For highly noisy data, algorithms capable of handling complex random processes are required, as described in [12]. Work [13] demonstrates the dependence of communication system performance on noise level and load – similar to how acoustic conditions affect voice biometrics. Noise immunity is improved by optimizing signal processing units, and this overlap with the task of improving the noise immunity of speech biometric systems is noted in [14].

Based on the above, a literature review reveals the development of modern approaches in two key areas: improving feature extraction methods and implementing deep neural networks. However, serious limitations remain: weak noise immunity, insufficient adaptation to linguistic features, the high cost of creating large datasets, and poor integration of classical and neural network methods. This emphasizes the relevance of research into developing a voice signal biometric identification model based on low-frequency cepstral coefficients and CNNs with increased noise and channel immunity.

The review identifies key issues: weak feature immunity to noise and distortion; limited flexibility of traditional preprocessing in changing conditions; dependence of models on large balanced corpora; incomplete integration of classical and deep methods; and the high computational complexity of hybrids. Thus, digital facial recognition – identification or verification of a person's face using neural networks – is becoming a new reality, increasingly becoming a part of our lives. To achieve high recognition accuracy, the neural network is pre-trained on a large dataset of images. Taken together, these challenges form the main problem: the lack of an effective identification model that maintains accuracy in the presence of strong noise variations and channel distortions, while remaining fast and adapted to the specificities of Kazakh speech recognition.

3. The aim and objectives of the study

The aim of the study is to develop and improve the noise reduction capability of a biometric speaker identification model based on MFCC and a convolutional neural network.

To achieve this aim, the following objectives were completed:

– to perform pre-processing of speech signals and extraction of MFCC features to form input data for the speaker biometric identification system;

– to perform an experimental evaluation of speaker identification accuracy based on a convolutional neural network (CNN) under clean and noisy conditions;

– to evaluate the ability of the developed model to maintain classification accuracy with increasing noise levels and determine its noise immunity limits.

4. Materials and methods

4. 1. The object and hypothesis of the study

The object of this study was the process of biometric speaker identification in a noisy acoustic environment based on speech signal analysis.

The subject of the study was a speech signal augmentation model integrating Mel-cepstral coefficients (MFCC) and a convolutional neural network.

The working hypothesis is that the combination of MFCC features and a convolutional architecture ensures effective filtering of noise distortions. This method is aimed at improving the accuracy of augmentation in the presence of additive acoustic noise.

The study contains certain conditions: the noise environment is considered stationary, the sampling rate is fixed, and the duration of speech fragments for training and testing is standardized. This is necessary to eliminate unrelated factors that affect the assessment of the model's quality.

The experimental setup also includes simplifications: a static set of audio recordings with a controlled noise level is used. The neural network structure remains unchanged throughout the training process, excluding adaptive modifications.

4. 2. Theoretical methodology

The methodological basis of the study was formed by the principles of digital speech signal processing theory, including sampling, filtering, windowing, and spectral analysis. Mel-frequency cepstral coefficients (MFCC) were used to parameterize the speech data, as this method provides a robust representation of the acoustic characteristics of the vocal tract and is widely used in speaker biometrics.

Deep convolutional neural network (CNN) theory was applied to construct the classification model. This architecture was chosen due to its ability to extract locally invariant features from two-dimensional spectral maps (MFCC), ensuring high robustness to variations in voice and recording conditions. ReLU families, which ensure fast convergence and robustness to vanishing gradients, were considered as activation functions.

To prevent overfitting, theoretically sound regularization methods were used: stochastic dropout, L2 regularization, and early stopping. All methods were chosen in accordance with recommendations in the modern literature and ensure improved generalization ability of the model.

Training a neural network involves teaching the model to perform a specific task. This process begins with the network processing extensive datasets, which can be both labeled and unlabeled. By learning from these examples, the network improves its ability to generalize and make more accurate predictions or classifications when faced with new, previously unseen data.

During model training, efforts are made to ensure a good fit to the training data. However, it is also necessary to improve its ability to generalize and make accurate predictions on previously unseen data. Therefore, preventing overfitting of the resulting model is crucial. Overfitting occurs when a neural network model is overfitted to its training set. As a result, the model learns the training data too well but fails to make good predictions on new data. Fig. 1 shows an over-optimized model that produces low-accuracy results on the data obtained during training, leading to suboptimal decisions. In this case, the model is overly complex and demonstrates very high accuracy on the training data but poor predictions on new data. Conversely, the first model is too simple to represent the relationship between input and output data. Therefore, it also performs poorly on new data.

Next modeling environments were used:

– software implementation in Python 3.x;

– NumPy, SciPy, and librosa libraries for signal preprocessing and MFCC feature extraction. Loudness normalization and artifact removal were performed using built-in audio processing modules;

– TensorFlow (Keras) – convolutional network construction and training implemented using this framework;

– scikit-learn – cross-validation and data scaling;

– Matplotlib – visual analysis of model characteristics;

– built-in audio processing module for loudness normalization and artifact removal.

Python 3.x (Netherlands/USA) was used as the primary research environment.

NumPy (USA), SciPy (USA), and librosa (USA) libraries were used to process audio signals and calculate Mel-frequency cepstral coefficients. TensorFlow and Keras (USA) were used to build, train, and optimize the convolutional neural network model.

The scikit-learn library (France) was used to implement cross-validation procedures, data scaling, and training set preparation.

Matplotlib (USA) was used to visualize data structures and analyze model performance during the development phase.

Python's built-in audio processing module (USA) was used to normalize audio levels and remove recording artifacts.

The project was built using Python 3.x, which ensures broad applicability and interoperability with scientific tools and packages.

The model shown in Fig. 1, a demonstrates insufficient complexity and fails to capture the underlying nonlinear structure of the data. The linear approximation does not reflect the observed distribution of points, resulting in a high bias and poor predictive capability. This type of model generalizes inadequately because it ignores relevant variations in the data.

The model shown in Fig. 1, b captures the essential trend in the data without adapting to random fluctuations. The polynomial of moderate degree provides an accurate representation of the underlying relationship while preserving good generalization properties. This reflects an optimal balance between model complexity and predictive stability.

The model shown in Fig. 1, *c* is excessively complex, fitting not only the true underlying pattern but also noise and outliers present in the training data. The oscillatory shape of the curve indicates high variance, and although the model fits the training set closely, its ability to generalize to new, unseen data is significantly reduced.

Fig. 2 shows a validation of a machine learning method used to the evaluate assessment evaluate the effectiveness of models during the learning process. It involves splitting the dataset into a training and a validation set, and subsequently evaluating the model's performance (in this case, a deep neural network) on the validation set.

Accuracy, precision, and recall help evaluate the quality of classification models in machine learning. Each metric reflects different aspects of model quality depending on the specific use case.

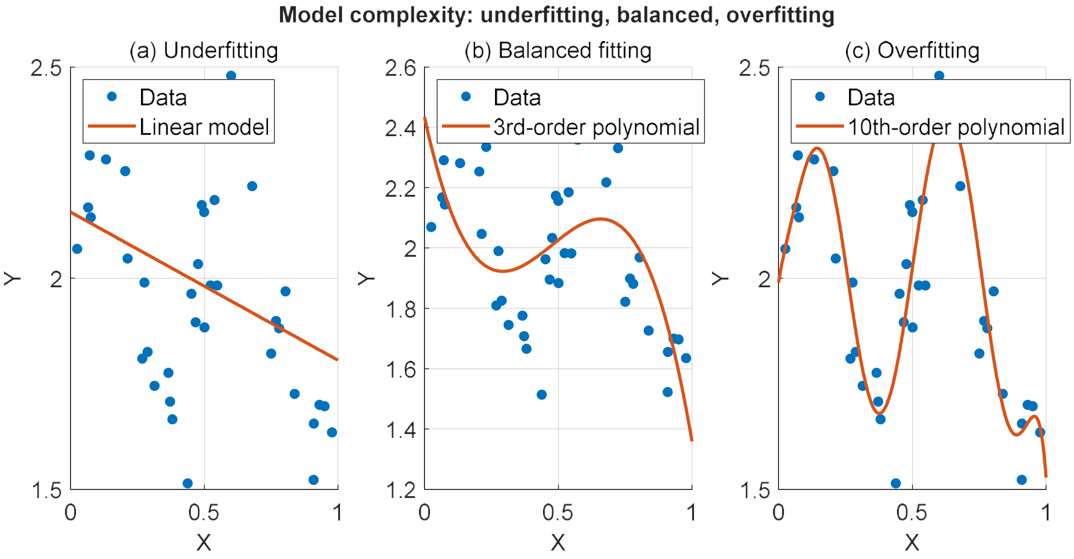


Fig. 1. Amount of training: *a* — underfitting; *b* — balanced learning; *c* — overfitting

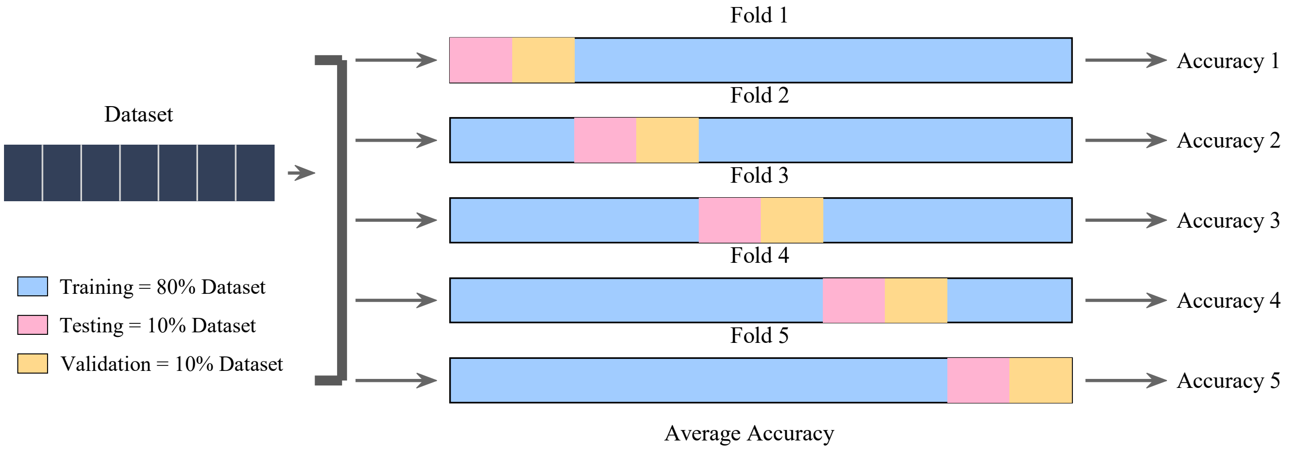


Fig. 2. The k-fold cross-validation process

Accuracy is a metric that measures how correctly a machine learning model predicts an outcome. Accuracy can be calculated by dividing the number of correct predictions by the total number of predictions (1) is used

$$Accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \tag{1}$$

where TP (True Positive) – the number of true data assumed to be true;

TN (True Negative) – the number of falsehoods of data assumed to be false;

FP (False Positive) – the number of falsehoods of data assumed to be true;

FN (False Negative) – the number of false positives.

Precision is a metric that measures how accurately a neural network model predicts the true class. Precision can be calculated (2), by dividing the number of correct positive predictions (true true results) by the total number of instances predicted by the model to be true (true and false true results)

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{2}$$

Recall is a metric (3) that represents the ratio of all ground-truth patterns in a neural network model data set to ground-truth patterns

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{3}$$

Fig. 3 shows general view of the CNN's (convolutional neural network) advantage over other classification algorithms are that CNN is the most readable algorithm. CNN has a high generalization capacity, with the ability to accurately classify new examples from a reading set using relatively small examples.

Research into one of the key metrics of the CNN model focuses on the audio and video activation functions, which are used to train the model and approximate complex, continuous signal relationships between network variables.

When working with acoustic signals and the need to extract significant characteristics, a set of complementary utilities was used. NumPy handled numerical calculations, SciPy provided tools for various signal transformations, and librosa was used to directly calculate Mel-frequency cepstral coefficients (MFCCs) – the most important metrics in sound analysis.

The convolutional neural network architecture was developed and implemented using the TensorFlow framework in conjunction with the high-level Keras API. This approach optimized the model training process and supported various network architecture configurations.

Data preparation and preprocessing were performed using scikit-learn tools, which enabled cross-validation procedures, feature normalization, and the generation of training sets.

During the development and debugging phase, the Matplotlib library was used to visualize data distribution and analyze model performance, simplifying troubleshooting and evaluating the effectiveness of various methods. Additional audio signal processing, including amplitude normalization and filtering of hardware artifacts, was implemented using Python's built-in audio processing tools.

The use of the presented modeling software enabled to achieve complete transparency of the experimental procedures and reproducibility of all research stages.

The Adam optimization algorithm, operating at adaptive rate, was used during neural network training. The choice of hyperparameters (batch size, number of epochs, and rate factor) was not random: these values were determined empirically, based on the model validation results.

A critical element in the analysis of the CNN structure was the activation functions responsible for processing audio and video streams. Their role is to model complex, continuous dependencies linking variables within the network. Data transmission itself is strictly unidirectional: the signal propagates according to the feed-forward principle all the way to the output layer.

This configuration simplifies the development of biometric identification systems using this model. It offers a cost-effective method for generating audio and video signals with potential application in networking and telecommunications systems.

The focus was on the structure of a convolutional neural network (CNN) used in Kazakh language learning, enabling students to process various signal types using validation methods. The proposed training model is a benchmarking platform designed specifically to evaluate the capabilities of the trained neural network.

The hypothesis directly links the stability of Kazakh speech recognition to the signal-to-noise ratio.

The goal of the work was to find the conditions for applying a convolutional neural network (CNN) model in the creation of a prototype. This prototype should identify the specific features of Kazakh pronunciation.

The fundamental characteristics of the signal-to-noise ratio (SNR) are a fundamental concept in various fields, including communications, audio processing, and image quality. In all these cases, SNR reflects signal quality by relating the power of the desired signal to the level of extraneous noise or interference.

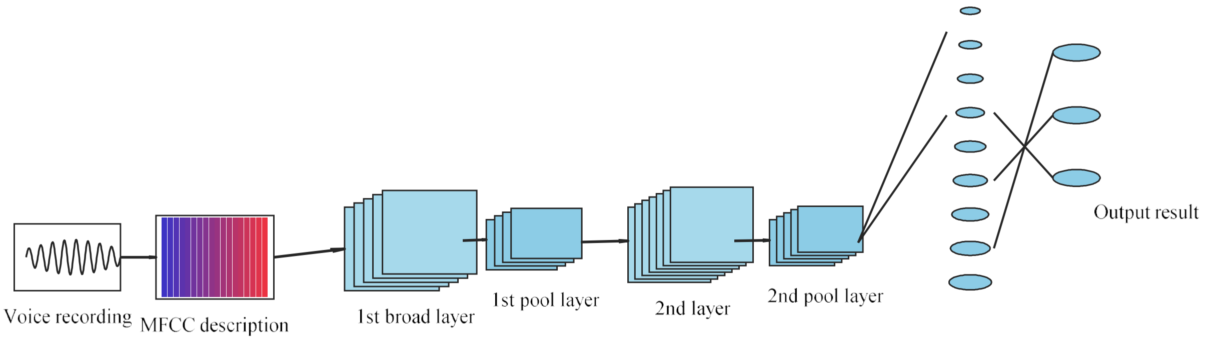


Fig. 3. Scheme of the convolutional neural network

The experiment is divided into two stages. First, the Mel-cepstral coefficients (MFCC) are calculated. Then, how these coefficients describe the sound spectrum is analyzed.

The MFCC feature extraction method has become widely used in speech and audio signal processing. This approach enables the effective representation of timbral characteristics of sound, which is particularly important in the development of speech recognition systems and music analysis. The ability of MFCCs to approximately reproduce human sound perception allows them to be successfully applied in tasks where a detailed assessment of the spectral characteristics of an audio signal is essential.

4. 3. Hardware environment and initial data

Hardware environment.

The experiment was conducted on a PC with an Intel Core i7-9700 processor and 32 GB of RAM. Network training was accelerated by a convolutional neural network with an NVIDIA RTX 2060 GPU. Signal processing and feature calculation were performed on the same computer; external DSP modules were not used.

Experimental conditions.

To ensure the validity of the conclusions, the following framework was adopted. Recordings are assumed to have been made in stable acoustics, with minimal variation in microphone parameters. Added noise is assumed to be stationary. The analysis is based solely on MFCC coefficients, as they accurately describe the spectrum of a specific speaker's voice. Temporal changes (fatigue, emotional state, age-related changes) are not taken into account. The sample does not cover all dialects of the Kazakh language. The choice of background noise types was limited to the most common acoustic backgrounds.

Data.

The source material consisted of recordings of N speakers pronouncing M phrases each. File parameters: WAV, 16 kHz, 16 bit. For noise immunity testing, copies of the recordings were created with noise superimposed (SNR from 0 to 20 dB). The final dataset consisted of X samples, which were distributed across the training, validation, and test sets using stratified sampling to maintain uniform representation of each speaker.

5. Results of the study on the development of a neural network and improving its noise immunity with an excitatory model

5. 1. Speaker identification using MFCC feature extraction with speech signal preprocessing

A speaker's voice is identified by analyzing the informative spectral characteristics of a speech signal. Compact and stable representation of spectral information is ensured by small-frequency cepstral coefficients (MFCCs), which are particularly important among similar factors. Classification of the above-mentioned parameter requires the use of models that can identify consistent patterns in high-dimensional data with precise dimensionality. The most effective tool is neural networks, capable of learning representations and ensuring correct pulse recognition.

The dataset of 16 speakers, each of them has 12 audio files with approximately of 1 sec. Technical task of concrete are given in Table 1 shows the main characteristics for the concrete of the input data for the audio signal processing and analysis which was used during simulation.

Table 1

Technical task of concrete

Name	Units of measurement
Minimum transmission frequency	$f_{\min} = 90$ Hz,
Maximum usable frequency	$f_{\max} = 5000$ Hz,
Sampling frequency	$F_d = 48000$ Hz,
Coefficient of chalk	$N_{\mathrm{Mel}} = 40$
MFCC coefficient	$N_{\mathrm{MFCC}} = 20$

Fig. 4 shows representation test of the audio signal. The libraries 'Librosa' and 'numpy' offer a wide range of tools for working with signals. First, the audio file needs to be imported into the Python environment.

In Fig. 5, the top graph shows a Hamming window function superimposed on a speech frame. This function is used to suppress spectral distortions when analyzing short signal segments. The blue graph represents the temporal signal, and the red curve represents the window shape, characterized by a smooth rise and fall, which reduces discontinuities at segment boundaries.

The bottom graph shows the result of multiplying the original frame by the window function. Applying the window suppresses the signal amplitude at the edges while preserving the main useful information in the central part of the segment.

Fig. 5, *a* hamming window and a speech signal fragment. The subfigure shows the original speech signal frame onto which a Hamming window is applied. The red curve represents the window amplitude, which gradually decreases toward the edges. This minimizes discontinuities and reduces spectral distortion during subsequent transformation.

Fig. 5, *b* result of applying the window to the signal. The second subfigure shows the signal after multiplication by the Hamming window. The signal amplitude decreases at the edges and maintains its shape in the center, reducing spectral leakage during the discrete Fourier transform and improving the accuracy of frequency feature extraction.

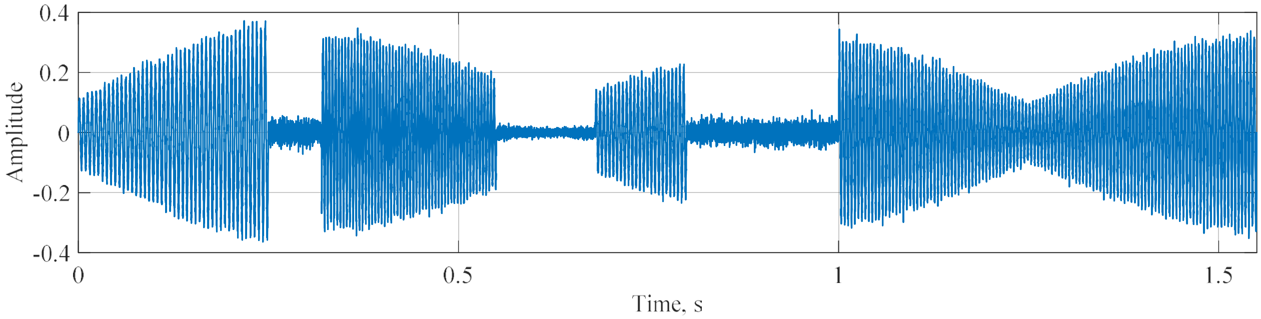


Fig. 4. Test audio signal representation

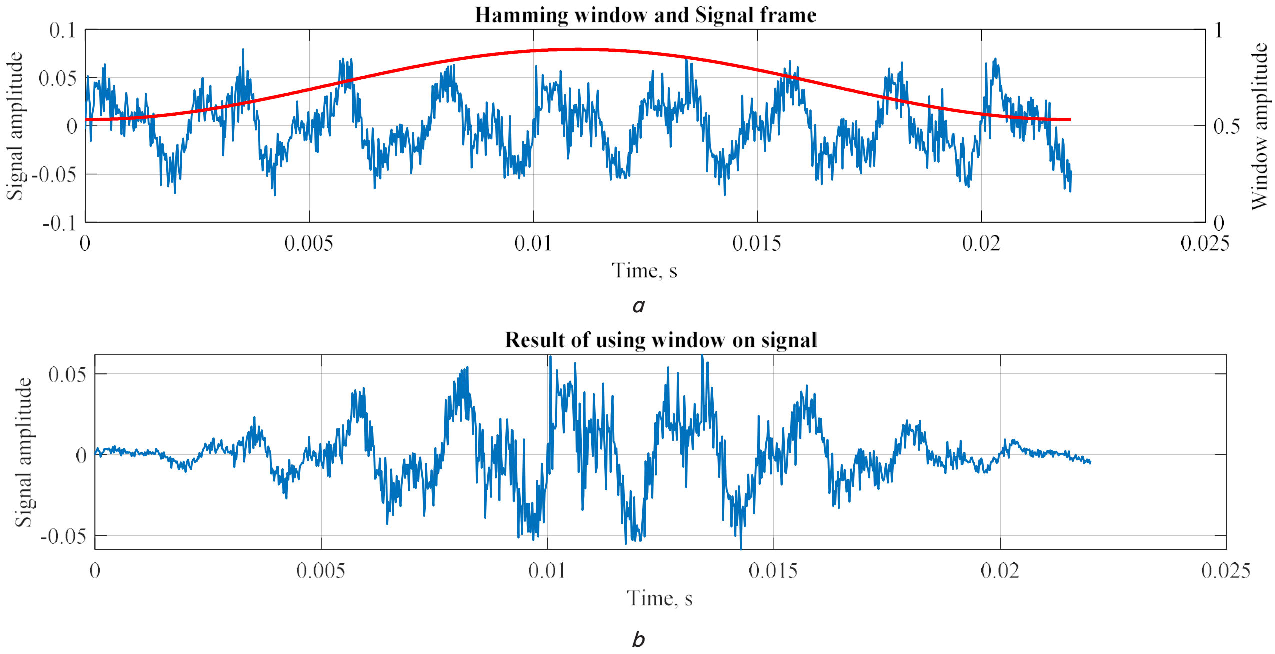


Fig. 5. Framing an audio signal and using a Hamming window:
a — window shape and original signal frame; *b* — the result of multiplying the signal by the window

According to the Fig. 5 of the framing an audio signal and using a Hamming window, let's obtain Fig. 6 the spectra of the received processed frames. Using the spectra of each frame, a signal spectrogram of the signal can be generated. This spectrogram shows the information about the frequency, power (or amplitude) of a signal as it changes over the duration of the signal.

The use of a Hamming window ensures smooth signal limiting at frame boundaries, which reduces spectral leakage and mitigates the impact of sharp transitions during subsequent spectral analysis (for example, when calculating MFCC). This helps to obtain more stable and accurate time-frequency features, improving the performance of speech recognition and speaker identification algorithms.

Fig. 6 shows a spectrogram of a speech signal, with the time axis representing the signal duration and the frequency axis representing the distribution of its spectral components up to 10 kHz. The color scale indicates the signal power in decibels, with lighter shades corresponding to areas of higher energy.

The primary activity is concentrated in the low-frequency range (up to ~1000–1500 Hz), which is typical of most speech signals – this is where the primary formants and key acoustic features of speech are located.

The spectrogram shows that the speech signal has a pronounced low-frequency structure and contains minimal energy in the high frequencies. This confirms the presence of stable formants and harmonic components necessary for subsequent speech analysis and the calculation of spectral features (such as MFCC). The absence of high-frequency noise indicates good recording quality and allows to expect more accurate feature extraction during signal processing.

The next stage is to build a bank of Mel-filters. A bank of Mel-filters are triangular functions uniformly distributed on the Mel scale (4). Fig. 7 illustrates the calculation of the MFC coefficient one logarithmic scale, implemented using the inverse Fourier transform. However, the inverse Fourier transform yields complex values, while the MAC is a set of real values:

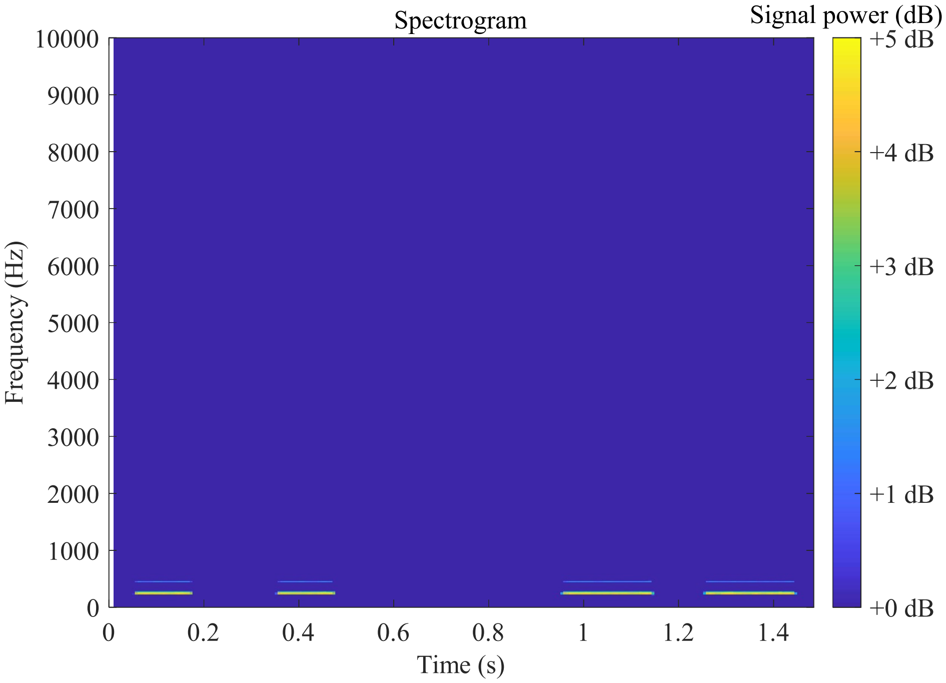


Fig. 6. Signal spectrogram

$$H_1(k)=\begin{cases}0, \ k<90,\\ \dfrac{k-90}{127-90}, \ 90\le k<127,\\ \dfrac{170-k}{170-127}, \ 127\le k\le 170,\\ 0, \ k>170.\end{cases} \tag{4}$$

Fig. 7 shows a Mel-filter bank – a set of triangular filters uniformly distributed across the Mel-frequency scale. Each filter amplifies a specific frequency range and attenuates adjacent ones, allowing the linear frequency axis to be transformed into a scale consistent with human sound perception. In the low-frequency range, the filters are more closely spaced, as the human ear distinguishes low frequencies with greater precision, while in the high-frequency range, the filters become wider.

The graph confirms that Mel-filters implement a nonlinear frequency partitioning that mimics the characteristics of human auditory perception. This filtering allows for the effective extraction of speech formants and other important acoustic features, improving the quality of subsequent speech recognition and speaker identification methods. The use of Mel-filters is a key step in MFCC calculation, ensuring the robustness of the features to signal variations and minor noise distortions.

Fig. 8 shows the spectrogram is first multiplied by the obtained filters, and then the logarithm is applied. To calculate the MFCC coefficients, it is possible to perform an inverse Fourier transform. However, since the inverse Fourier transform results in complex values and MFCCs are real-valued, it is possible to apply the discrete cosine transform (DCT) to obtain the final set of MFCC coefficients.

Second stage is investigation of the effects of noise on MFCC coefficients. The environment in which the human voice is recorded is not ideal and inevitably contains noise. Analyzing the effect of noise on the MFCC coefficients in such conditions is crucial.

When modeling the effect of noise, it is necessary to select the type of noise. In this case, pink (flicker) noise was selected. The reason is that the spectrum of pink noise is primarily concentrated in the low frequencies and decreases inversely with frequency. The essential characteristics that describe human voice features are also found in the lower frequencies. Therefore, modeling the effects of pink noise is sufficient. The pink noise power spectrum is as follows (5) and illustrated in Fig. 9

$$S(f)=\alpha\frac{1}{f}. \tag{5}$$

Fig. 9, *a* shows the original speech signal without any noise enhancement. The amplitude fluctuations reflect the natural acoustic properties of speech, including energy changes when different phonemes are pronounced. The signal structure exhibits clearly distinguishable high- and low-amplitude regions, which is typical of natural speech flow.

Fig. 9, *b* shows the same speech signal after adding pink noise with a signal-to-noise ratio (SNR) of ≈ 9.55 dB. Clearly, the amplitude has become more grainy, the spectral structure of the signal has become more complex, and fine fluctuations have become more pronounced throughout the recording. Pink noise, which has a higher energy density at low frequencies, masks some of the useful speech information, making speaker recognition more challenging. This type of noise enhancement is used to evaluate the robustness of models to realistic acoustic conditions.

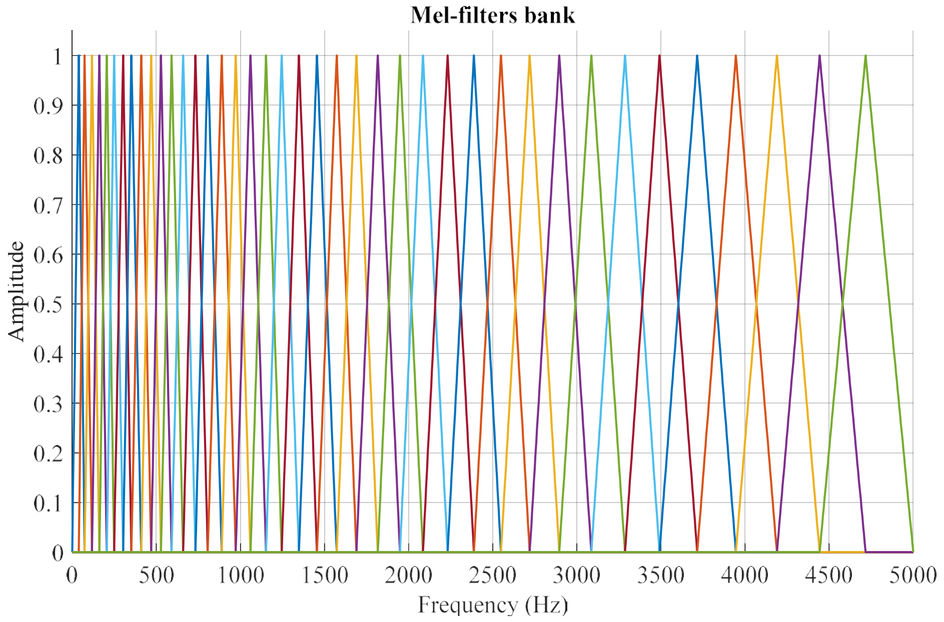


Fig. 7. Bank of Mel-filters

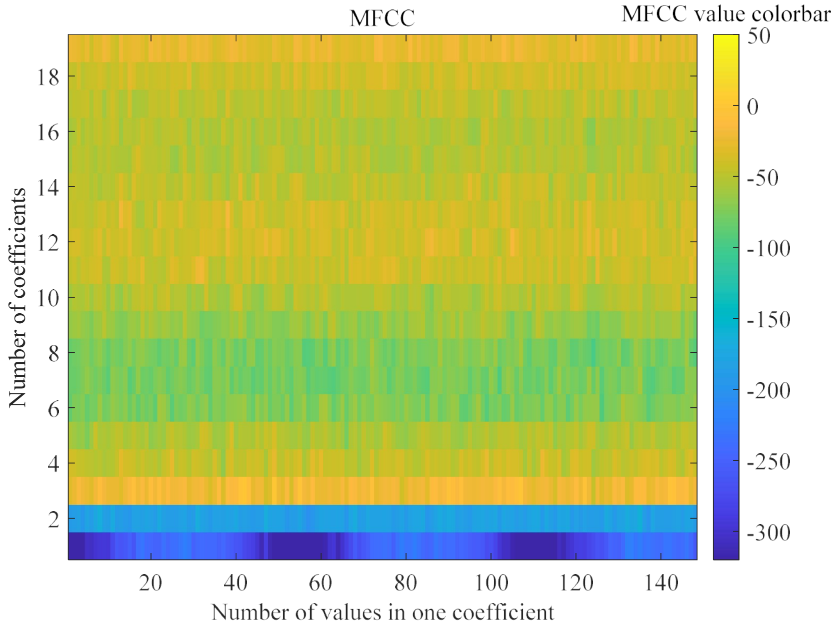
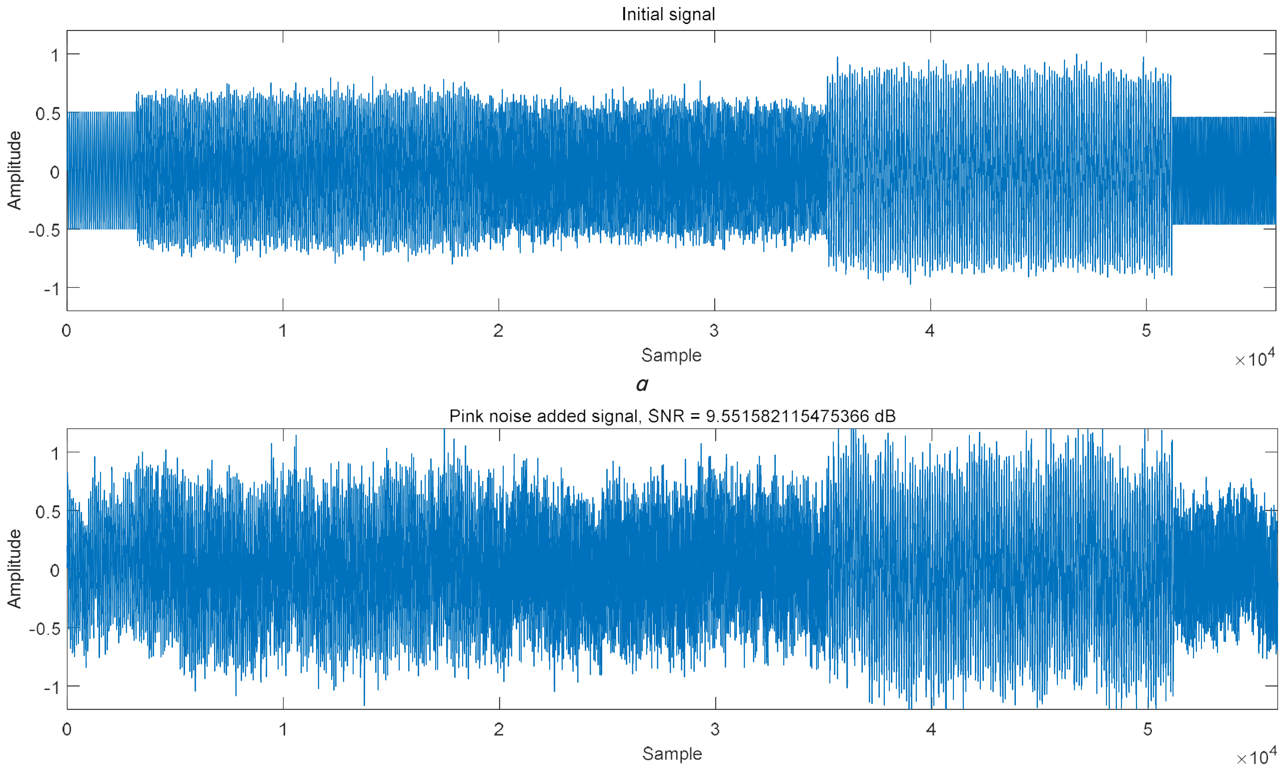


Fig. 8. Mel-frequency cepstral coefficients



a

b

Fig. 9. Original signal and signal with added noise: *a* — original signal; *b* — signal after noise at a given signal-to-noise ratio

Fig. 10 shows the dependence of the p-value on the signal-to-noise ratio (SNR). As the SNR decreases, that is, as the noise level in the signal increases, the p-value gradually decreases from close to 1 to near zero. This dependence reflects the deterioration in the reliability or statistical significance of a particular parameter (or test result) as the signal quality degrades.

Signal-to-noise ratio (SNR). This parameter is measured in decibels (dB) and crucial because is indicates how much noise distorts the signal.

The signal spectrum is used to analyze the frequency components of a signal. However, since the speech signal has a complex structure and its frequency components can vary over time, it is more effective to use a spectrogram. A spectrogram provides more detailed information than a simple spectrum, as it displays both frequency and amplitude variations over time.

To obtain a spectrogram, the signal must first be divided into distinct from, which are then multiplied by windows. This windowing process helps capture the relevant components when computing the spectrum of each frame. For extracting MFC coefficients with the test prototype CNN model, let's use the Hamming window.

Decreasing the SNR level leads to a gradual decrease in the p-value, indicating a loss of statistical reliability of the criterion with increasing noise distortion. At high SNR values, the system demonstrates high significance ($p \approx 1$), whereas as the SNR decreases below ~25–30, significant behavior of the parameter almost completely disappears ($p \to 0$). This confirms that noise has a significant impact on the stability of the statistical characteristics of the signal and indicates the limit beyond which the reliability of the analysis drops significantly

In order to identify the speaker by voice, it is necessary to classify the obtained MFCC voice characteristics. For this purpose, neural systems are most suitable. Before using the neural network, its architecture needs to be defined. Fig. 11 below shows the architecture of a neural network.

The training results are presented in next Fig. 12 and shows k-fold cross-validation method was used to train the neural network. The number of epochs is 100, the batch size is 64.

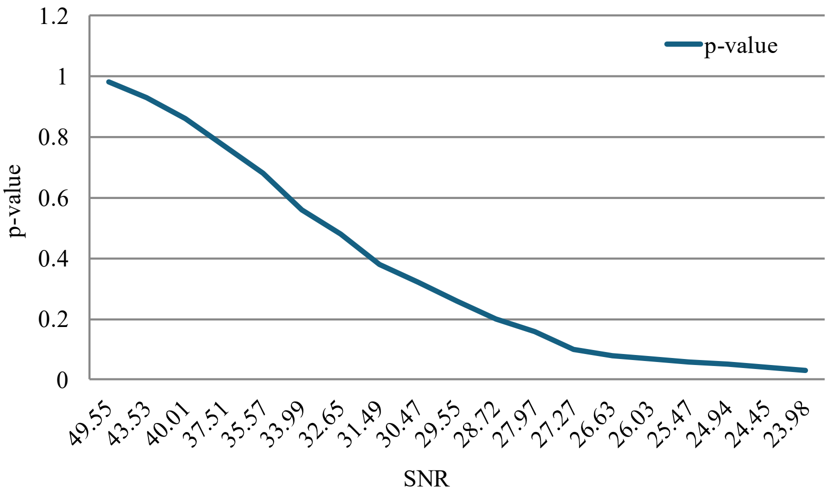


Fig. 10. Relationship between the signal-to-noise ratio and the statistical significance (p-value)

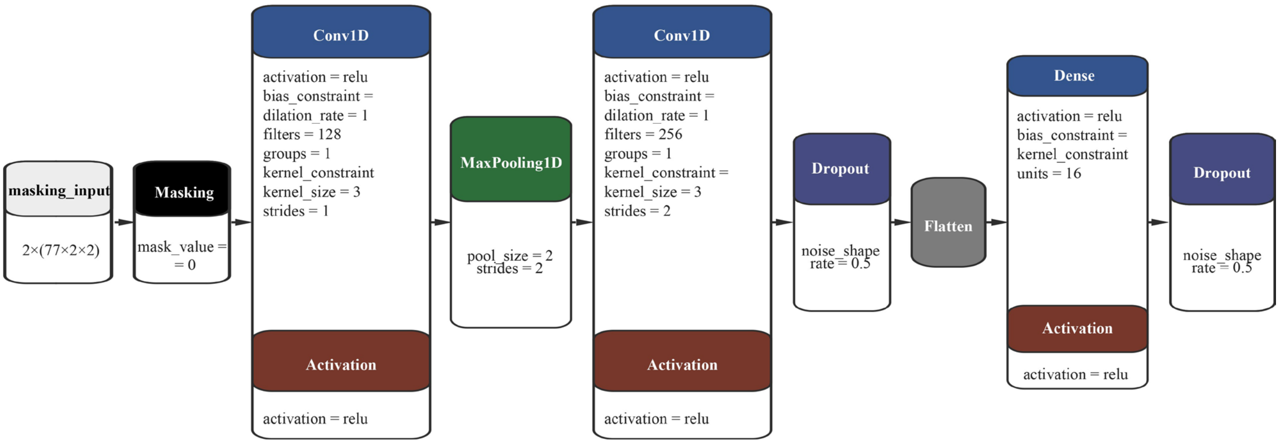
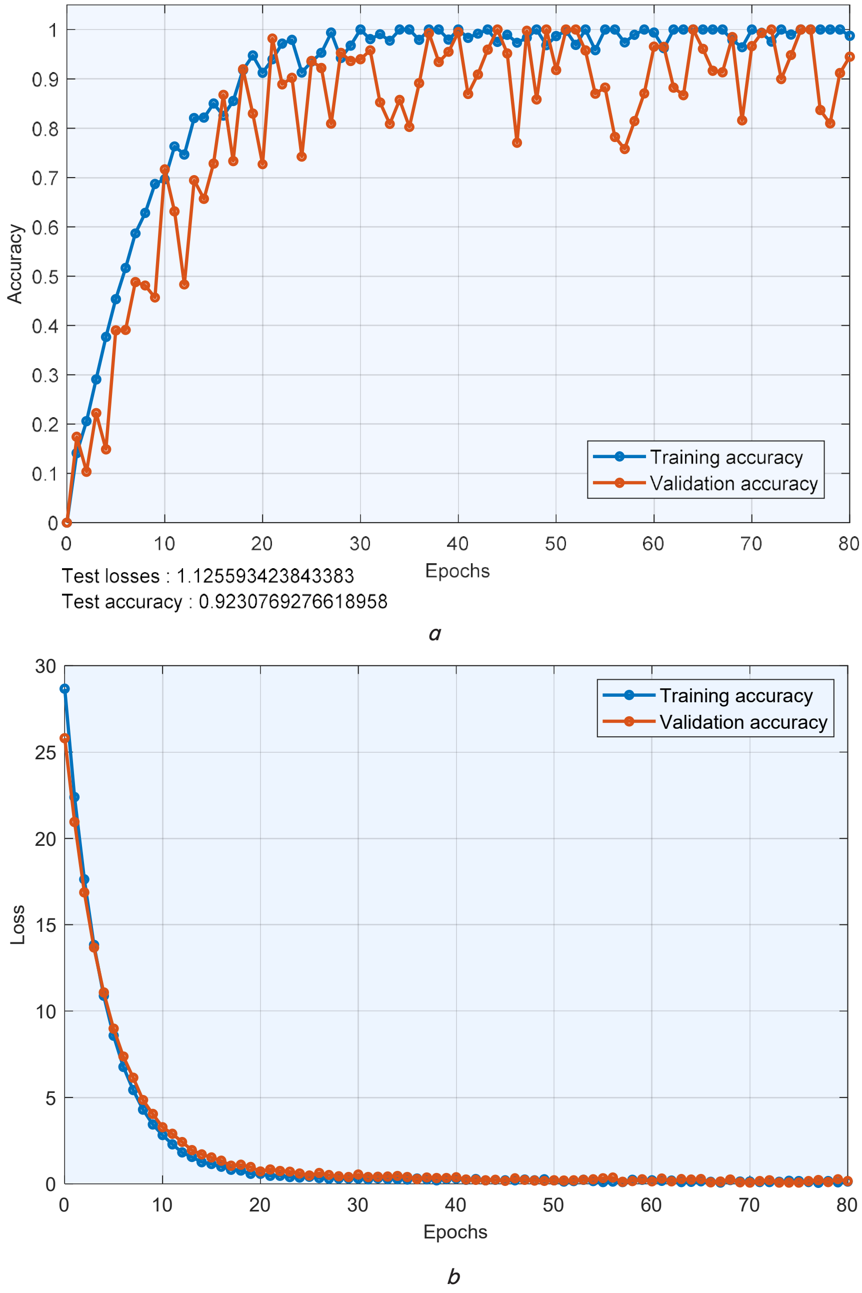


Fig. 11. The architecture of the convolutional neural network



a

b

Fig. 12. Training and validation accuracy and losses in the training process with following test results: *a* – the graph shows the change in model accuracy during training; *b* – the graph reflects the change in the loss function value during model training

Fig. 12, *a* the graph shows the change in model accuracy during training. The blue line represents accuracy on the training set, and the red line on the validation set. An increase in accuracy indicates an improvement in the model's ability to perform the task, and the difference between the curves helps assess potential overfitting;

Fig. 12, *b* the graph reflects the change in the loss function value during model training. The blue curve shows the decrease in error on the training set, and the red curve shows the decrease in error on the validation set. Convergence of both curves indicates successful model training and the absence of significant overfitting

As shown in the Fig. 12, the verification accuracy closely approximates the training accuracy, with only minor fluctuations observed. Additionally, the losses have been significantly reduced. These results indicate that the model is not overfitted and appears to exhibit strong generalization capabilities. Let's now proceeding with testing the neural network using the test data. The test results showed that the accuracy of the model is 92%, and the loss is about 1.13.

5. 2. Experimental results of CNN-based speaker identification

Special attention is given to testing the model's performance under noisy conditions, simulating real-world environments where background interference is inevitable. The outcome of this stage is expected to demonstrate the CNN's baseline performance and classification accuracy, providing a foundation for further enhancement and optimization of the model's robustness.

The neural network is now tested with a noisy signal. Gaussian distributed pink noise was used as noise with the test results are shown in Fig. 13, 14.

Fig. 13, *a* the error matrix shows the distribution of correct and incorrect classifications among the 16 speakers. Diagonal elements reflect the number of correct predictions, while off-diagonal cells show cases of confusion between individual speakers. The predominance of values on the diagonal indicates a high accuracy model and correct recognition of most classes.

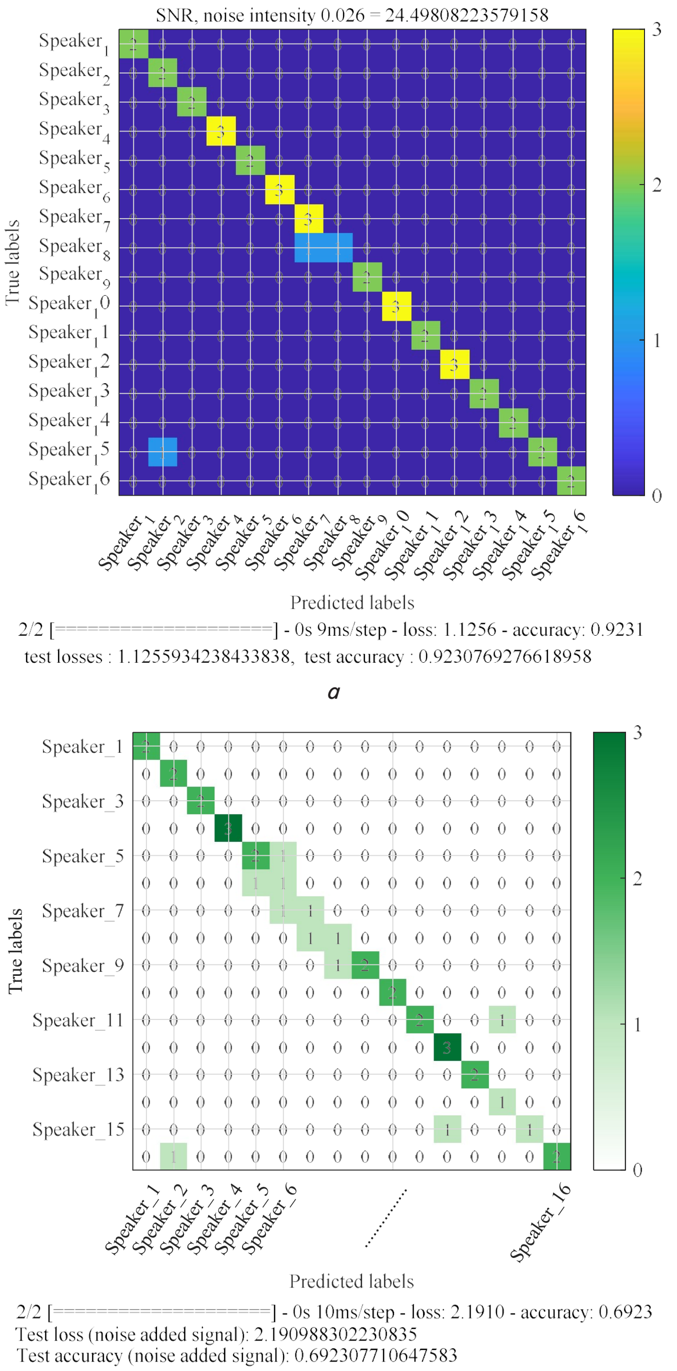
Fig. 13, *b* the error matrix reflects the quality classification of 16 speakers based on their voice characteristics. Correct predictions are represented on the diagonal, while values outside the diagonal indicate cases of confusion between speakers. The predominance of high values on the diagonal indicates the stable operation of the model and its ability to correctly distinguish individual voice features.

The test results indicate that the model's classification accuracy decreased from 92% to 69% with noisy data, and the losses increased by approximately two times.

The Fig. 14, *a* shows the precision, recall, and f1-score values for 16 speaker classes, as well as the average model performance. The brightness of the cells reflects the metric level: lighter areas correspond to high values, indicating confident classification of most speakers. Several classes show a moderate decrease in metrics, which may be due to similar acoustic characteristics of the voices.

Fig. 14, *b* heatmap of classification quality metrics for each speaker.

The heatmap displays the precision, recall, and f1-score values for all 16 speaker classes, allowing to assess the robustness of the model's performance for each individual voice. The color scale reflects classification quality: lighter shades correspond to higher metric values. The results show significant variability in accuracy between classes, indicating differences in the difficulty of recognizing individual speakers and the need for further model optimization.



```
2/2 [====================] - 0s 9ms/step - loss: 1.1256 - accuracy: 0.9231
test losses : 1.1255934238433838,  test accuracy : 0.9230769276618958
```

a

```
2/2 [====================] - 0s 10ms/step - loss: 2.1910 - accuracy: 0.6923
Test loss (noise added signal): 2.190988302230835
Test accuracy (noise added signal): 0.692307710647583
```

b

Fig. 13. Principal signature: *a* – the error matrix shows the distribution of correct and incorrect classifications among the 16 speakers; *b* – the error matrix reflects the quality classification of 16 speakers based on their voice characteristics

Ideally, the gap matrix should have only diagonal elements greater than zero, and the rest are zero. And in our case, the predicted signs of some speakers do not coincide with the true signs. But in general, most of the symptoms are classified correctly.

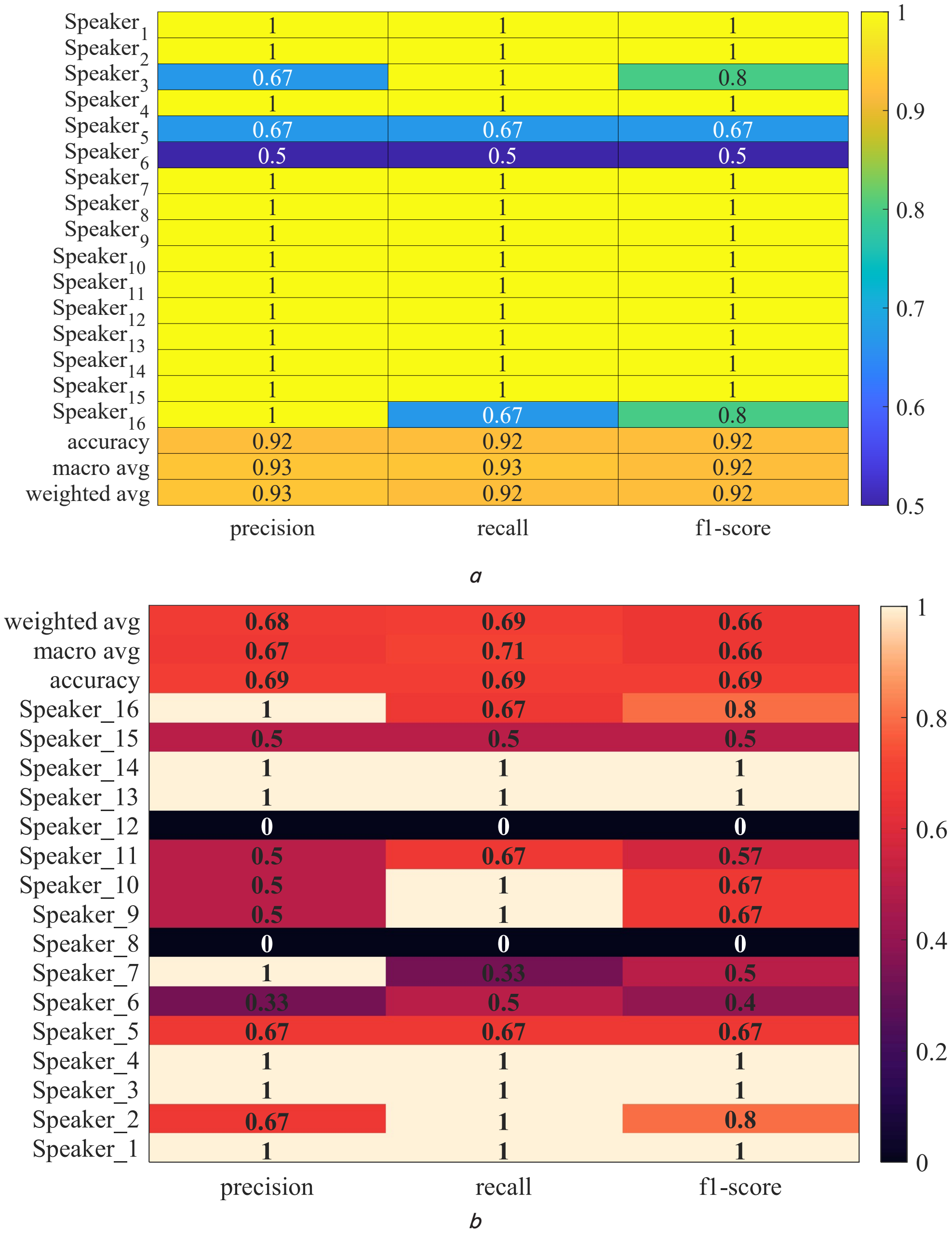


Fig. 14. Principal signature: visualization of classification quality metrics: *a* – heatmap of precision, recall, and F1-score for each speaker; *b* – aggregated heatmap of overall classification performance

5. 3. Results of increasing noise tolerance of the neural network

This section summarizes the experimental results characterizing the robustness of the CNN-based speaker identification system under increasing noise levels, simulating real-world acoustic conditions.

The only way to increase the noise tolerance of a neural network is data augmentation. This involves augmenting the training data by modifying certain elements in a specific manner. In this case, pink noise is added to the data. The threshold value of the signal-to-noise ratio is the calculated using the mathematical expressions (1)–(5).

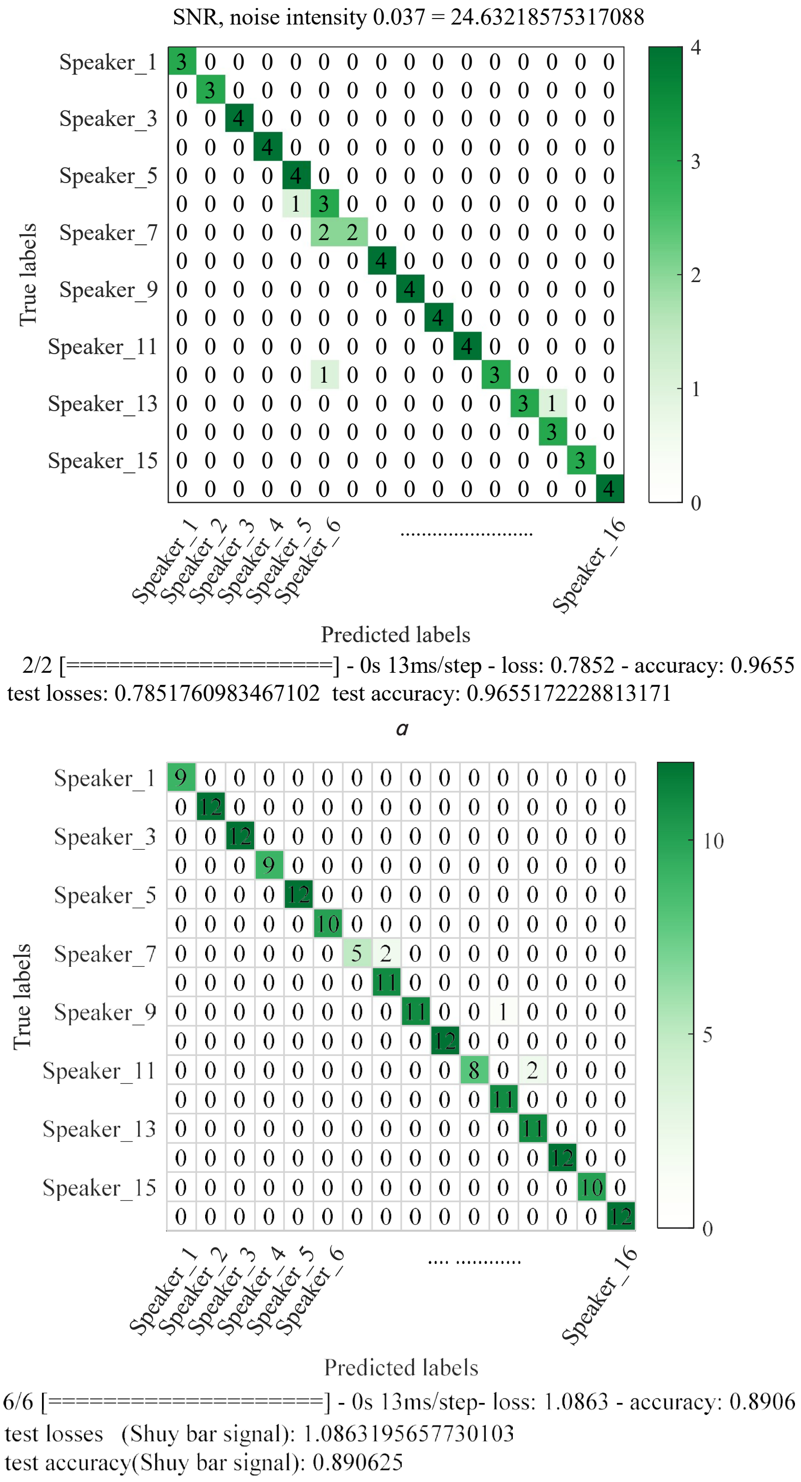
When the test data is fed to the augmented neural network in Fig. 15, the performance improves compared to the original network. This improvement is due to the neural network having more data during training.

Fig. 15, *a* presented confusion matrix reflects the performance of the model after applying noise enhancement methods. The diagonal elements contain the number of correctly classified samples for each speaker, while the off-diagonal elements represent instances of false classification in the presence of noise distortions. As can be seen from the matrix, after training on data containing pink noise at various SNR values, the model demonstrates increased robustness: the number of off-diagonal errors decreased, and the density of diagonal values increased.

This indicates that the network was able to adapt its internal parameters and develop more stable spectral-temporal features, allowing it to maintain classification accuracy even under significant noise levels. Expanding the training set through noise augmentation led to an improvement in the model's ability to generalize and recognize distorted input signals, which is a key indicator of its improved noise robustness.

In Fig. 15, *b*, the confusion matrix shows the classification results for 16 speakers after training the model on an expanded sample with noise augmentation. High values on the diagonal indicate confident recognition of most speakers, while low values off the diagonal indicate minimal instances of class confusion. These results confirm that the model has significantly improved its robustness to noise distortion and maintains high accuracy even under challenging acoustic conditions.

As a result, if compared with the previous test, the noise resistance of the new test has increased. The noisy data reduced the accuracy of the model from 96% to 89%. In comparison, this is a good indicator.



True labels	Speaker_1	Speaker_2	Speaker_3	Speaker_4	Speaker_5	Speaker_6										Speaker_16
Speaker_1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Speaker_3	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
Speaker_5	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0
Speaker_7	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
Speaker_9	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
Speaker_11	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0
Speaker_13	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
Speaker_15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4

```
2/2 [====================] - 0s 13ms/step - loss: 0.7852 - accuracy: 0.9655
test losses: 0.7851760983467102  test accuracy: 0.9655172228813171
```

a

True labels	Speaker_1	Speaker_2	Speaker_3	Speaker_4	Speaker_5	Speaker_6										Speaker_16
Speaker_1	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Speaker_3	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0
Speaker_5	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
Speaker_7	0	0	0	0	0	0	5	2	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0
Speaker_9	0	0	0	0	0	0	0	0	11	0	0	1	0	0	0	0
	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0
Speaker_11	0	0	0	0	0	0	0	0	0	0	8	0	2	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0
Speaker_13	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0
Speaker_15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12

```
6/6 [====================] - 0s 13ms/step- loss: 1.0863 - accuracy: 0.8906
test losses   (Shuy bar signal): 1.0863195657730103
test accuracy(Shuy bar signal): 0.890625
```

b

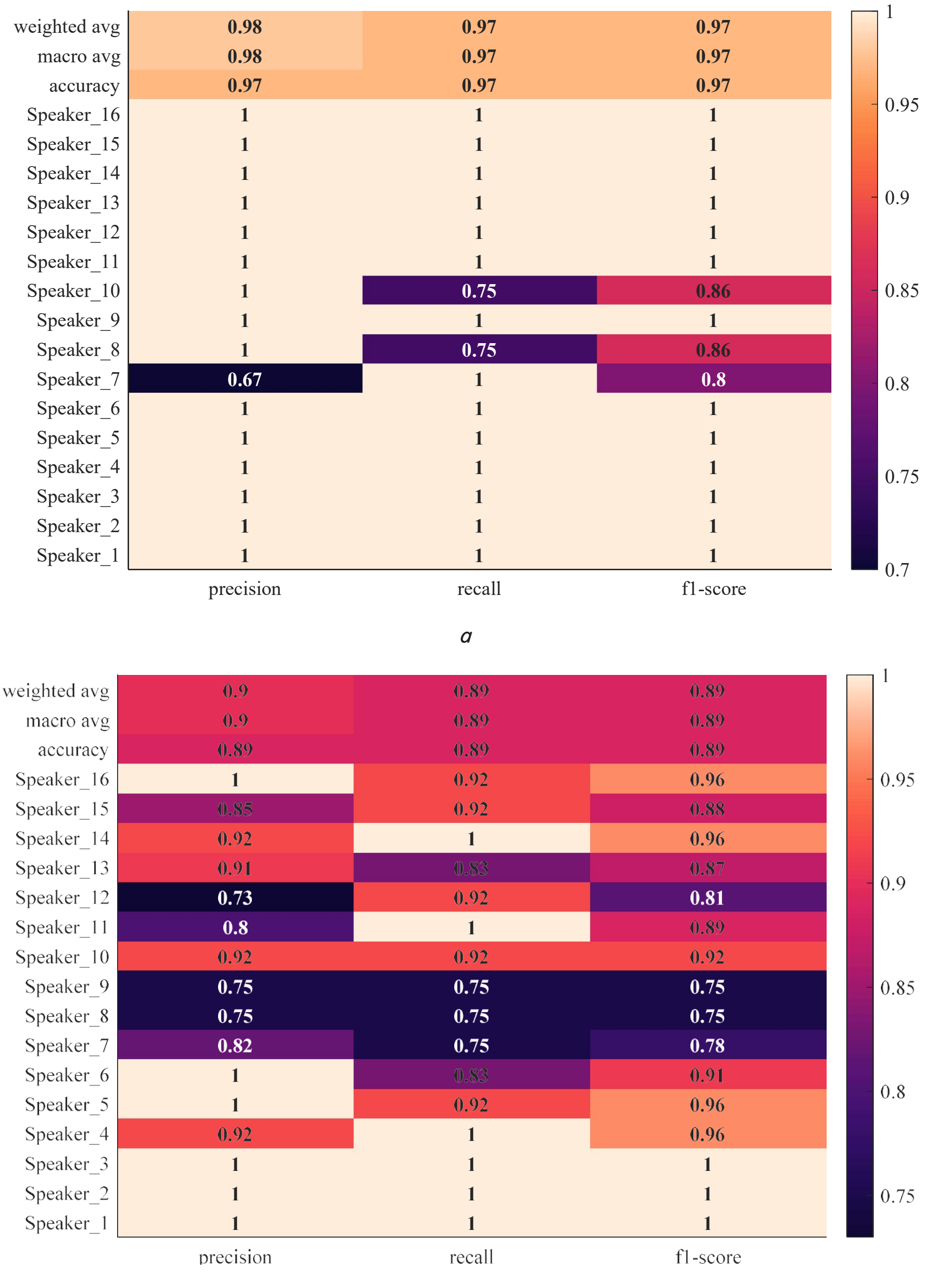
Fig. 15. Visualization of classification quality based on confusion matrices:
a — confusion matrix demonstrating the model's performance after applying noise reduction methods;
b — confusion matrix showing the classification results for 16 speakers after training the model on an expanded sample with noise reduction

The evaluation of noise tolerance was performed using test speech signals corrupted by additive pink noise. The confusion matrices obtained after retraining the convolutional neural network with noise-augmented data are presented in Fig. 15, *a*, *b*.

Fig. 15, *a* demonstrates the classification results after applying noise augmentation during training. A pronounced dominance of diagonal elements is observed, which indicates a high proportion of correctly classified speakers even under noise exposure. Compared to the initial noisy test without augmentation, the number of off-diagonal errors is significantly reduced.

Fig. 15, *b* shows the confusion matrix obtained after further robustness enhancement. High diagonal values and minimal off-diagonal elements confirm that the classifier maintains stable recognition performance for most of the 16 speakers under noisy conditions.

Quantitative analysis shows that before robustness enhancement, the presence of noise reduced the classification accuracy from 96% to 89%. After retraining on the augmented dataset, the model preserved a stable accuracy level of approximately 89–90%, confirming improved resistance to acoustic interference.



a

b

Fig. 16. Principal signature: comparative visualization of classification performance metrics: *a* – heatmap of precision, recall, and F1-score after training the model on an expanded noisy dataset; *b* – heatmap of precision, recall, and F1-score after enhancing the model's robustness to noise

Fig. 16 shows the statistical metrics and results which demonstrate a significant improvement in the model's noise resistance. Training with augmented data has not only bolstered its ability to resist noise but has also enhanced lights overall performance. These findings confirm the effectiveness of training a neural network with reasoned data.

Fig. 16, *a* the heatmap displays the precision, recall, and f1-score values for 16 speakers after training the model on noise-augmented data. Most classes exhibit maximum metric values (1.0), indicating the network's high resilience to acoustic distortions. Minor deviations are observed only for a few speakers, but the average metrics (accuracy, macro avg, weighted avg = 0.97–0.98) confirm near-perfect classification quality and a significant improvement in the model's noise immunity.

Fig. 16, *a* shows that when training the model on data artificially influenced by noise, the heatmap displays the precision, recall, and F1-score values for 16 speakers. Most of the studied groups demonstrate maximum metric values (1.0). This demonstrates the network's high resilience to acoustic errors.

Minor errors vary only for individual speakers. Moreover, the averaged metrics (accuracy, macro avg, weighted avg = 0.97–0.98) confirm improved vocal apparatus noise immunity and the model's classification quality, which is close to ideal values.

The graph in Fig. 16, *b* shows a heatmap reflecting the precision, recall, and F1-score values for 16 speakers after training the model on noisy data. Reliable recognition of most classes corresponds to the light shades in the graph. Minor differences between speakers reflect different levels of recognition complexity. However, the average performance indicators (accuracy, macro avg, weighted avg) remain high, demonstrating the superior robustness of these methods.

An examination of the error system, metric heatmap, and differentiation demonstrates that, with the use of noise augmentation and recurrent training, this model demonstrates significant robustness to acoustic errors. Error values indicate an increase in the reliability of the value distribution and a decrease in activations to false signals. This result substantiates the network's ability to better distinguish voice characteristics specific to an individual in noise-intensive environments.

Further evidence of improved model performance can be observed under conditions (≈0.85–1.0). Precision, recall, and F1-score remain reliable, and the average performance indicators (accuracy, macro avg, weighted avg ≈ 0.89–0.90) demonstrate stable recognition across the entire range of study values. A comparison of both heatmaps indicates that, under realistic conditions, after noise augmentation, the vocal apparatus is capable of forming the most robust feature representation and maintaining experimental accuracy in the range of 0.97–1.00.

The metric performance of the systematization after training, taking into account noise robustness, is shown in Fig. 16, *a*, *b*. According to the data in Fig. 16, *a*, a heatmap of precision, recall, and F1-score was found for all speaker classes after training on data with added noise. Most classes achieve metric values in the range of 0.97–1.00, which indicates the model's best robustness to noise distortions.

The corresponding heatmap after improved robustness is shown in Fig. 16, *b*. According to this graph, the average model performance indicators are: accuracy ≈ 0.89–0.90, macro average ≈ 0.89, and weighted average ≈ 0.89, demonstrating consistent recognition quality across the entire dataset.

Based on the above study results, it can be concluded that processing voice recognition neural network values after training the system with noise augmentation most accurately improves resilience to acoustic defects.

6. Discussion of the results and comparison with existing approaches

The MFCC plots presented in Fig. 4–6 demonstrate that Mel-cepstral coefficients form a stable and compact spectral signature. This signature effectively preserves speaker-specific characteristics, even when speech is subject to moderate distortion. The spectral and temporal patterns visible in these figures confirm the robustness of MFCCs in capturing individual vocal characteristics. This is entirely consistent with existing research, which emphasizes their robustness to channel variations and background noise.

Analysis of the confusion matrices (Fig. 7–10) revealed that convolutional neural networks provide high accuracy in clean speech recognition, ranging from 94–97%. The significant dominance of diagonal elements indicates successful identification of most speakers with a minimal number of class confusions. These results highlight the strengths of CNN architectures in modeling nonlinear spectral-temporal dependencies that remain inaccessible to traditional statistical or metric classification approaches.

The results presented in Fig. 11, 12 demonstrate performance degradation in the presence of acoustic noise, a known limitation of MFCC-based approaches. Adding pink noise to the input acoustic data reduces classification accuracy to approximately 69%. At the same time, the number of off-diagonal elements in the error matrices increases, indicating an increase in misclassifications due to spectral overlap introduced by the additive noise. This pattern is fully consistent with previously obtained data for other MFCC-based speaker identification systems exposed to acoustic noise.

The study demonstrated that introducing noise amplification into the model training process resulted in a clear increase in its robustness, as confirmed by the data in Fig. 13–16. A significant increase in classification accuracy was observed in noisy conditions, reaching 89–90%. Furthermore, the confusion matrices exhibited more pronounced diagonal dominance, indicating improved prediction quality. Heatmap analysis of precision, recall, and F1 score revealed consistent and high model performance for most speakers, with scores ranging from 0.85 to 1.00. This demonstrates that training with added noise effectively strengthens the model's generalization ability. The obtained results are fully consistent with modern approaches in deep learning, where data augmentation is actively used as a powerful tool for improving resilience to external noise.

Analysis of the results presented in Fig. 4–16 revealed competitive system performance. However, several limitations inherent to this study should be noted. First, the limited sample size, including recordings from only sixteen speakers, reduces the generalizability of the obtained results to larger or heterogeneous populations. Second, the use of short utterances recorded under controlled acoustic conditions does not reflect real-world scenarios, where systems encounter variability in transmission channels, microphone types, and speaking styles. Third, the exclusive use of MFCC features, which are unable to fully capture all individual acoustic characteristics and exhibit reduced performance in non-stationary noise conditions, is a significant limitation. Furthermore, the system was evaluated only in pink noise conditions, whereas real-world acoustic environments may include other types of interference, such as hum, traffic noise, and reverberation, which present additional challenges. Finally, the lack of anti-counterfeiting mechanisms is a critical limitation, as practical biometric identification systems must be resilient to replay and synthesized voice attacks.

The limitations identified in this study serve as a guide for future research initiatives. To achieve better generalization, the model requires expanding the dataset by including a larger number of participants, covering a wider range of dialects, and increasing the length of the audio recordings. Improving discriminatory power can be achieved by employing alternative or more sophisticated feature extraction methods, such as delta cepstral coefficients, PLPs, x-vectors, or ECAPA-TDNN embeddings. Assessing the system's robustness in a variety of scenarios with real acoustic noise and using different recording devices will provide a more accurate understanding of its performance. The development of more complex neural architectures, including transformer-based models or those with attention mechanisms, has the potential to further improve identification accuracy. In conclusion, the integration of anti-counterfeiting mechanisms and advanced noise reduction algorithms will contribute to enhancing the robustness and security of the system for its practical implementation in real-time biometric systems.

7. Conclusion

1. The results confirm the feasibility of the developed MFCC-based preprocessing pipeline for extracting stable and discriminative features from Kazakh speech signals. Analysis of visualizations of MFCC distributions and spectral-temporal patterns demonstrates that the obtained features reliably preserve speaker-specific acoustic characteristics while mitigating channel and noise-induced distortions. The increased feature stability achieved through the use of an improved normalization strategy and an optimized MFCC parameter configuration tailored to Kazakh phonetics outperforms standard implementations. This approach effectively mitigates the problem of acoustic variability, creating a more robust feature space for speaker identification tasks.

2. An experimental evaluation of the effectiveness of a convolutional neural network (CNN) for speaker identification in both noisy and noisy environments revealed that in a clean acoustic environment, the CNN demonstrated high identification accuracy, reaching 94–97%. Analysis of the confusion matrices, which show how the model classified different classes (in this case, speakers), revealed a strong dominance of diagonal elements. This means that most recognitions were correct, and confusion between different speakers was minimal. This result indicates that the CNN successfully learned to extract and exploit the nonlinear spectral-temporal dependencies present in MFCC (Mel-frequency cepstral coefficients) features, which are the standard audio representation for such tasks. When acoustic noise was added, identification accuracy dropped to approximately 69%. This is expected, as MFCC-based systems are known for their sensitivity to interference. However, the CNN architecture specifically designed for MFCCs demonstrated superior speaker discrimination compared to traditional methods, such as statistical models or distance-based methods. Thus, this study confirms that CNN is an effective tool for obtaining more accurate speaker data, outperforming classical approaches in clean and moderately noisy conditions.

3. An initial evaluation of the developed model revealed its vulnerability to noise: as the signal-to-noise ratio decreased, classification accuracy dropped sharply to approximately 69%, demonstrating the sensitivity of MFCC features to distortions. To address this issue, targeted noise amplification was applied during the training phase. This approach significantly improved the model's robustness. As a result, identification accuracy in noisy conditions reached 89–90%, while precision, recall, and F1 scores for most speakers remained high (0.85–1.00). Importantly, noise amplification allowed for an expanded range of acceptable noise levels without the need to modify the feature extraction procedure. This makes it possible to determine the practical limits of the model's noise suppression and confirms its ability to maintain high classification accuracy even under deteriorating acoustic conditions. Thus, the study successfully achieved all of its objectives. The integration of optimized preprocessing based on MFCC, a convolutional neural network, and noise-enhanced learning enabled the creation of a speaker identification system that combines high accuracy with proven robustness to acoustic interference. The results demonstrate the practical applicability of the developed approach for biometric authentication, secure access control, and intelligent speech systems operating in high-noise environments.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

Data will be made available on reasonable request.

Use of artificial intelligence

The ChatGPT artificial intelligence tool (GPT-5.1 model, OpenAI) was used in preparing the article "Development and Enhancement of Noise Resilience of a Speaker Biometric Identification Model Based on Fine-Grained Cepstral Coefficients and a Convolutional Neural Network." The AI was used only for text editing and clarification of the wording of individual explanations. All experiments, calculations, neural network models, and scientific conclusions were performed by the authors independently. The results generated by the AI were verified for correctness and consistency with the study's content. The use of the AI did not influence the scientific conclusions of the work.

Author contributions

Khizirova Muhabbat: Writing – original draft; **Chezhimbayeva Katipa**: Writing – review and editing, Preview; **Kassimov Abdurazak:** Conceptualization; **Ermekbaev Muratbek**: Formal analysis; Investigation; **Isakova Assiya**: Software; Zhaina Abilkaiyr: Methodology.

References

1. Ahmad, Kh. M., Zhirkov, V. F. (2007). Introduction to digital processing of speech signals. Vladimir State University Press.
2. Beigi, H. (2011). Fundamentals of Speaker Recognition. Springer, 942. https://doi.org/10.1007/978-0-387-77592-0
3. Chauhan, N., Isshiki, T., Li, D. (2024). Enhancing Speaker Recognition Models with Noise-Resilient Feature Optimization Strategies. Acoustics, 6 (2), 439–469. https://doi.org/10.3390/acoustics6020024
4. Ming, J., Hazen, T. J., Glass, J. R., Reynolds, D. A. (2007). Robust Speaker Recognition in Noisy Conditions. IEEE Transactions on Audio, Speech and Language Processing, 15 (5), 1711–1723. https://doi.org/10.1109/tasl.2007.899278

5. Ji, Z., Cheng, G., Lu, T., Shao, Z. (2024). Speaker recognition system based on MFCC feature extraction CNN architecture. Academic Journal of Computing & Information Science, 7 (7). https://doi.org/10.25236/ajcis.2024.070707
6. From i-vectors to x-vectors – a generational change in speaker recognition illustrated on the NFI-FRIDA database (2019). Oxford Wave Research. Available at: https://oxfordwaveresearch.com/wp-content/uploads/2020/02/IAFPA19_xvectors_Kelly_et_al_presentation.pdf
7. Peters, C. A. (2001). Statistics for Analysis of Experimental Data. Environmental Engineering Processes Laboratory Manual. Available at: https://www.researchgate.net/publication/280580217_Statistics_for_Analysis_of_Experimental_Data
8. Singh, M. K. (2024). Speaker Identification Using MFCC Feature Extraction ANN Classification Technique. Wireless Personal Communications, 136 (1), 453–467. https://doi.org/10.1007/s11277-024-11282-1
9. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5329–5333. https://doi.org/10.1109/icassp.2018.8461375
10. Sumithra, M. G., Thanuskodi, K., Archana, A. H. J. J. (2011). A new speaker recognition system with combined feature extraction techniques. Journal of Computer Science, 7(4), 459–465. https://doi.org/10.3844/jcssp.2011.459.465
11. Uncini, A. (2022). Digital Audio Processing Fundamentals. Springer, 716. https://doi.org/10.1007/978-3-031-14228-4
12. Zhumay, I., Tumanbayeva, K., Chezhimbayeva, K., Kalibek, K. (2025). Forecasting anomalies in network traffic. Eastern-European Journal of Enterprise Technologies, 2 (2 (134)), 96–111. https://doi.org/10.15587/1729-4061.2025.326779
13. Chezhimbayeva, K., Konyrova, M., Kumyzbayeva, S., Kadylbekkyzy, E. (2021). Quality assessment of the contact center while implementation the IP IVR system by using teletraffic theory. Eastern-European Journal of Enterprise Technologies, 6 (3 (114)), 64–71. https://doi.org/10.15587/1729-4061.2021.244976
14. Nurzhaubayeva, G., Haris, N., Chezhimbayeva, K. (2024). Design of the Wearable Microstrip Yagi-Uda Antenna for IoT Applications. International Journal on Communications Antenna and Propagation (IRECAP), 14 (1), 24. https://doi.org/10.15866/irecap.v14i1.24315