ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

UDC 004.932: 004.5

DOI: https://doi.org/10.30837/2522-9818.2025.3.005

O. BARKOVSKA

TWO-FACTOR AUTHENTICATION BASED ON KEYWORD SPOTTING AND SPEAKER VERIFICATION

The **subject** matter of the article is the development and evaluation of a two-factor speaker authentication method based on voiceprint identification and keyword spotting (KWS), designed for secure voice-based access in human-machine interfaces, especially for users with limited mobility. The **goal** of the work is to create a method for managing speaker authentication using convolutional neural networks (CNNs), comparing the efficiency of two widely used spectral feature extraction techniques – Mel-Frequency Cepstral Coefficients (MFCC) and Short-Time Fourier Transform (STFT) spectrograms. The following tasks were solved in the article: a model of a two-factor authentication method is proposed, which includes speaker identification and voice password recognition; the quality of MFCC and STFT spectrograms features is compared; the influence of the number of epochs, CNN architecture and training parameters on the system accuracy is evaluated; the effect of the sampling rate on the performance of the models was investigated. The following **methods** are used: deep learning methods with CNN architecture, fine-tuning, MFCC, and STFT feature extraction, mathematical and statistical analysis of training efficiency, and system performance metrics. The following **results** were obtained: the method achieved 97.95% accuracy in speaker identification using MFCCs after 60 training epochs, and 99.82% accuracy in voice password verification using the same CNN structure after 20 epochs. The average accuracy of the entire authentication process was 98.75%. Moreover, using MFCC features reduced training time by a factor of 23 and memory consumption by a factor of 7 compared to STFT spectrograms. **Conclusions**: the effectiveness of a two-factor voice authentication method that combines speaker identification by acoustic voice characteristics and voice password verification was implemented and studied. Further research directions include studying the impact of alternative spectral features (in particular, CQCC, GFCC, prosodic parameters) on improving accuracy and resistance to spoofing. Special attention will be paid to optimizing the model for energy-efficient use on portable devices.

Keywords: voice authentication, identification, voice password, MFCC, spectrogram, CNN, MHI, biometrics.

Introduction

Interaction between humans and computers has become a common form of communication in the modern world. Human-machine interaction (HMI) can be facilitated through hardware devices such as keyboards, touchscreens, or mice, as well as via a more natural communication modality for humans – voice [1–2]. The proliferation of voice-based interfaces has led to significant scientific progress in areas such as speech modeling, linguistic pattern analysis, and acoustic signal processing. Voice technologies are emerging as a key component of the Fourth Industrial Revolution and are expected to have a growing impact on how people interact with machines.

Technological advancements in natural language processing (NLP), text-to-speech (TTS) systems, real-time speech pattern recognition, noise filtering, and multi-speaker separation (e.g., for conference calls), as well as system personalization based on speaker-specific attributes such as accent, speech rate, or physiological speech impairments, represent major scientific challenges with considerable practical value.

Voice and speech recognition methods, along with voice-driven human-machine interfaces, have particular practical significance in the following domains:

– providing accessibility for individuals with disabilities (e.g., voice control for those unable to use a keyboard) [3–4];

– supporting globalization and cultural preservation through universal translators and tools for low-resource languages [5].

A voice sample contains rich information – from the speaker’s gender and age to, in some cases, emotional state. Speaker recognition and voice authentication tasks typically aim to identify or verify a speaker based on one or more biometric parameters. In contrast to passwords or PIN codes, which can be guessed, stolen, or observed, voice biometric data is inherently unique, making it much harder to spoof. As such, voice-based identification and authentication offer an effective and secure method for protecting sensitive data and personal information.

Speaker identification can also serve as a component of multifactor authentication (Figure 1), alongside other modalities such as fingerprint recognition, facial identification, or PIN codes [6–8]. This layered approach enhances security, requiring an attacker to overcome multiple verification barriers to gain unauthorized access.

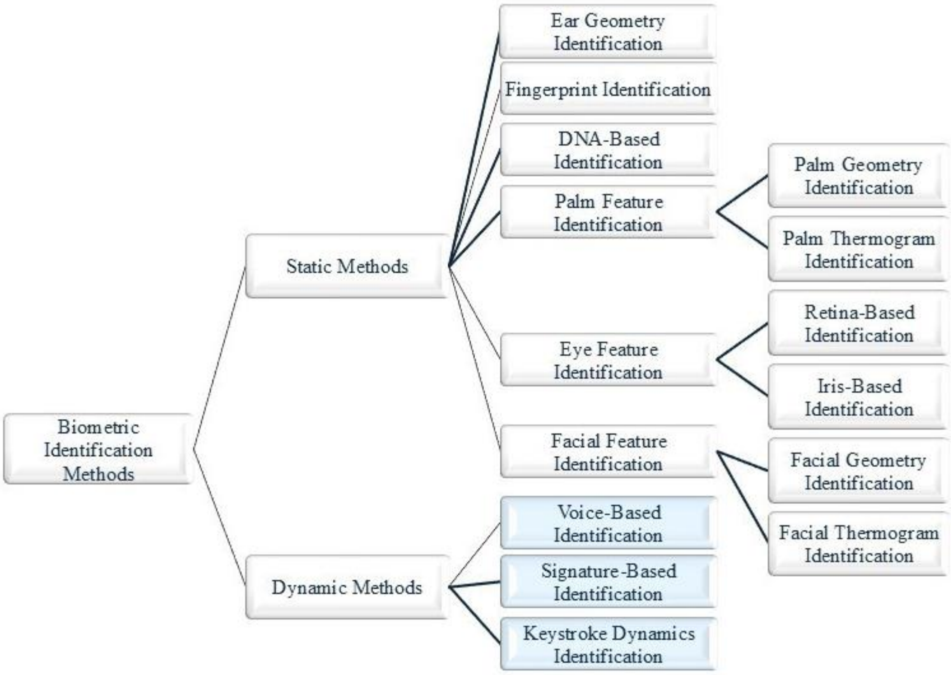

Biometric Identification Methods
Static Methods
Dynamic Methods
Ear Geometry Identification
Fingerprint Identification
DNA-Based Identification
Palm Feature Identification
Eye Feature Identification
Facial Feature Identification
Voice-Based Identification
Signature-Based Identification
Keystroke Dynamics Identification
Palm Geometry Identification
Palm Thermogram Identification
Retina-Based Identification
Iris-Based Identification
Facial Geometry Identification
Facial Thermogram Identification

Fig. 1. The role of voice identification among other biometric authentication methods

The analysis of a human voice sample enables the extraction of a wide range of psychophysiological, social, and biometric information, extending far beyond the basic task of speaker identification (Figure 2).

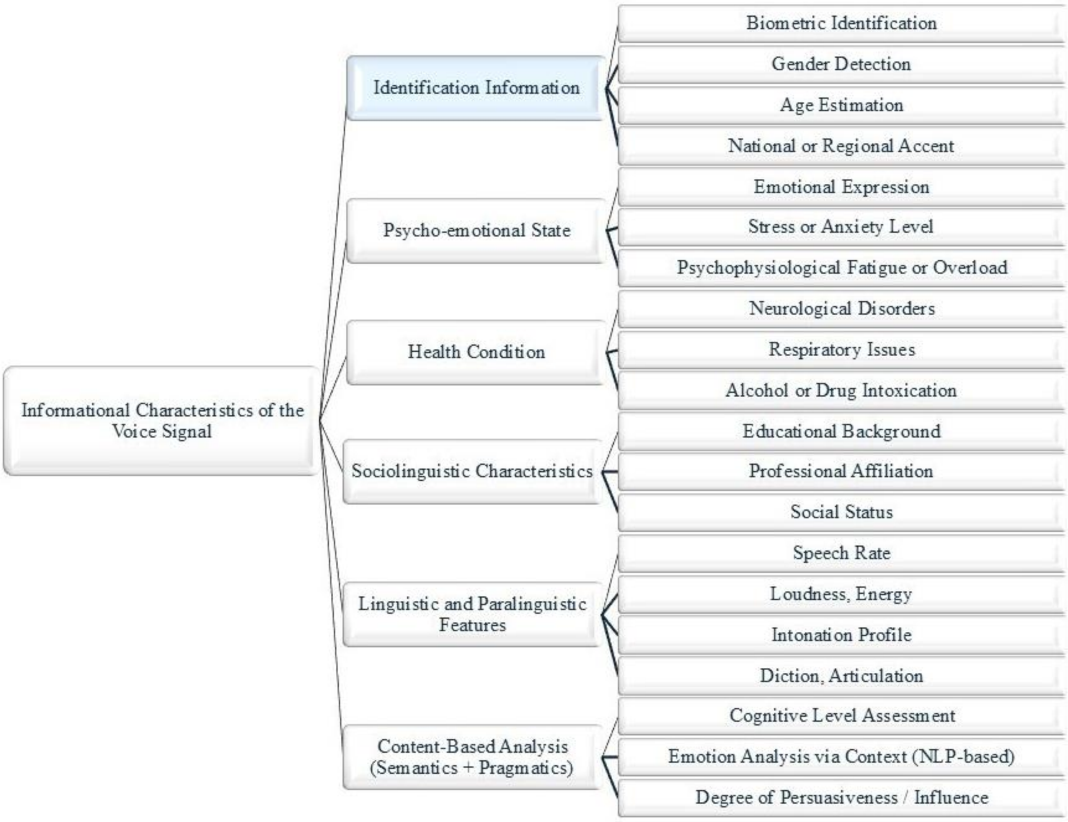

Informational Characteristics of the Voice Signal
Identification Information
Psycho-emotional State
Health Condition
Sociolinguistic Characteristics
Linguistic and Paralinguistic Features
Content-Based Analysis (Semantics + Pragmatics)
Biometric Identification
Gender Detection
Age Estimation
National or Regional Accent
Emotional Expression
Stress or Anxiety Level
Psychophysiological Fatigue or Overload
Neurological Disorders
Respiratory Issues
Alcohol or Drug Intoxication
Educational Background
Professional Affiliation
Social Status
Speech Rate
Loudness, Energy
Intonation Profile
Diction, Articulation
Cognitive Level Assessment
Emotion Analysis via Context (NLP-based)
Degree of Persuasiveness / Influence

Fig. 2. Categorization of tasks based on voice sample analysis

Analyzing a person's voice sample enables the extraction of detailed information not only about the individual but also about their psychophysiological state at the moment of speaking. Primarily, the voice is used for biometric identification, as each person possesses unique acoustic characteristics, including timbre, pitch, and articulatory movement patterns. These features allow for not only speaker identification, but also the estimation of gender, approximate age, and even linguistic background based on accent or dialect.

Advanced voice analysis can reveal the speaker's psycho-emotional state: parameters such as intonation, speech rate, loudness, pauses, and rhythm provide insight into emotions (e.g., joy, anxiety, anger), stress level, or even fatigue. Additionally, the voice contains significant health markers – neurological, respiratory, and even viral disorders may manifest through vocal tremor, hoarseness, irregular breathing, or dysarthria. These indicators are increasingly studied in the context of conditions such as Parkinson's disease and COVID-19.

The sociolinguistic dimension of speech further enables assessment of the speaker's educational level, professional affiliation, or social status through analysis of vocabulary, stylistic choices, and linguistic behavior. Paralinguistic features such as speech tempo, loudness, articulation clarity, and emotional expression offer additional cues regarding the speaker's intentions and confidence level.

Finally, content-level speech analysis can provide insight into the cognitive complexity of utterances, strength of argumentation, emotional polarity, and semantic richness – particularly when supported by natural language processing (NLP) tools. All these data can be captured and interpreted using modern acoustic signal processing techniques (e.g., formant analysis, spectrograms, MFCC), as well as deep learning methods, including neural network architectures such as CNNs, RNNs, LSTMs, and transformers, enabling the voice to function as a multidimensional channel of personal information.

Analysis of last achievements and publications

It should be noted that speaker identification and verification are different processes that can complement each other in voice authentication systems. According to the ISO/IEC 19794-13:2018 standard, identification is the process of determining a person from a set of registered users based on the acoustic characteristics of speech. Instead, verification (or authentication) checks whether the user's voice matches the reference sample, confirming the right to access. At the same time, authentication systems can use a fixed phrase (voice password), combining keyword recognition (KWS) and voice verification. In this way, voice identification determines who is speaking, and verification determines whether this person has access rights.

A review of the literature confirms that spectrograms (based on STFT) and MFCCs are the de facto standards in the architectures of modern deep learning models for speaker identification and verification tasks (e.g., x-vector, ECAPA-TDNN, CNN-KWS).

In the study by [9], the practical efficiency of MFCCs for speech feature extraction is demonstrated. The authors used 13 cepstral coefficients obtained from speech signals segmented with 50 ms frames and 50% overlap at a 16 kHz sampling rate. Despite limitations of the Madaline Type I neural network used for classification, the recognition accuracy within the database reached 61%, and rejection accuracy for unknown utterances reached 84%. This highlights the robustness of MFCCs in encoding speaker-specific spectral patterns independently of lexical content, which is critical for real-time speaker identification systems.

In the work by [10], the ECAPA-TDNN architecture is introduced as an enhanced version of the x-vector model. The authors implement channel attention and aggregation mechanisms that allow the model to better capture speaker-relevant features. The paper also notes that the use of MFCC as input features is a common practice in such architectures.

The study by [11] explores automatic speaker identification based on features extracted from spectrograms using a convolutional neural network (CNN). The authors demonstrate that spectrograms are effective input features for speaker identification tasks, as they preserve local spectral properties of the speech signal and result in high recognition accuracy.

A systematic review by [12] analyzes key feature extraction methods for speaker identification. The authors conclude that MFCCs and spectrograms are among the most effective and widely used approaches due to their capability to represent acoustically meaningful features relevant for distinguishing between speakers.

Other features – such as LPCC (Linear Predictive Cepstral Coefficients), PLP (Perceptual Linear Prediction), GFCC (Gammatone Frequency Cepstral Coefficients), CQCC (Constant-Q Cepstral Coefficients), as well as prosodic features (e.g., speech rate, intonation, pauses) and wavelet-based representations (DWT, CWT) –

have significant scientific value, particularly in multimodal systems, noisy conditions, or cross-domain applications [13–14].

In future work, the analysis will be extended to include CQCC and prosodic features, with particular attention to cross-lingual and intermodal verification scenarios.

This paper focuses on the most widely adopted feature types, as evidenced by their presence in state-of-the-art toolkits and frameworks such as Kaldi, SpeechBrain, and the PyTorch Speaker Verification Toolkit.

The extracted features serve as the input for the next stage of the standard audio sequence analysis pipeline – the analyzer or classifier [15–17].

The purpose of this study is to evaluate the authentication accuracy of the proposed two-factor authentication (2FA) method based on keyword spotting (KWS) and speaker verification by analyzing the impact of STFT-based spectrograms and MFCC features on the accuracy of speaker identification and verification. The system is implemented using a convolutional neural network (CNN) architecture with the application of fine-tuning techniques.

To achieve this goal, the following objectives were addressed:

– a model of a two-factor authentication method is proposed, which includes speaker identification and voice password recognition;

– the quality of MFCC and STFT spectrograms features is compared;

– the influence of the number of epochs, CNN architecture and training parameters on the system accuracy is evaluated;

– the effect of the sampling rate on the performance of the models was investigated.

Materials and methods

There are two main groups of methods on which voice identification and authentication systems are based (Figure 3) [18]:

– the reference method is based on comparing some voice features (these can be either physiological or articulatory features) with some reference. A group of individual words is used as a reference;

– the phoneme-oriented method is based on the extraction of individual phonemes from the speech stream.

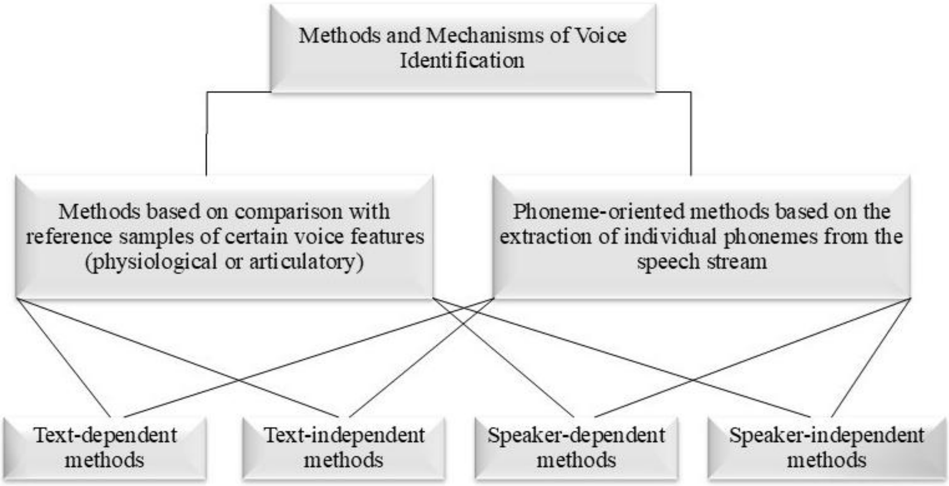

Methods and Mechanisms of Voice Identification
Methods based on comparison with reference samples of certain voice features (physiological or articulatory)
Phoneme-oriented methods based on the extraction of individual phonemes from the speech stream
Text-dependent methods
Text-independent methods
Speaker-dependent methods
Speaker-independent methods

Fig.3. Methods and mechanisms of speaker identification

A speaker-dependent model requires preliminary training on a specific speaker. It generates a speaker embedding, which is subsequently used to compare new voice samples. In contrast, a speaker-independent model is trained on a large, diverse corpus of speakers and can estimate the degree of similarity between samples with high accuracy, even when the speaker is not included in the training data.

Text-dependent identification is effective when the spoken content is known in advance (e.g., a predefined passphrase), as this reduces variability in the linguistic content. Text-independent models, on the other hand, are more flexible and capable of handling a wider range of input, but they require larger datasets and more complex architectures that can generalize across varying speech content.

These categories are closely related to challenges posed by acoustic variability (e.g., background noise, microphone quality), speech variability (e.g., emotional tone, intonation), and contextual variability (e.g., speaker stress, fatigue).

A generalized standard pipeline for speaker identification – across both speaker-dependent/independent and text-dependent/independent approaches – includes the following stages: preprocessing, feature extraction, feature selection, and decision-making or classification (Figure 4) [19].

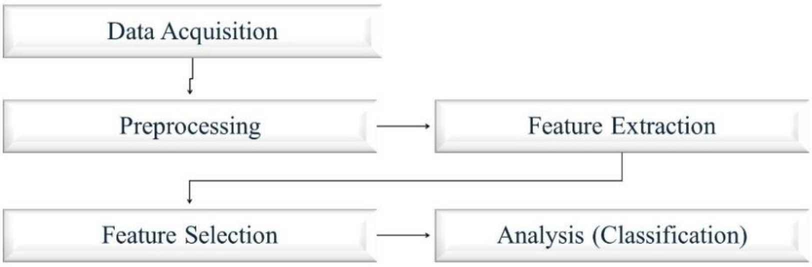


Fig.4. A standard pipeline for audio analysis (keyword spotting and voice verification)

Pre-processing is required to remove background noise and convert the input signal into a form suitable for feature extraction [20, 21]. The following feature extraction methods are commonly considered: MFCC (Mel-Frequency Cepstral Coefficients) – regarded as the standard for speaker identification tasks, though it requires signal normalization and pre-processing, LPC/LPCC (Linear Predictive Coding / Linear Predictive Cepstral Coefficients) – often used in combination with MFCC to enhance classification accuracy, PLP/RASTA-PLP (Perceptual Linear Prediction) – well-suited for noisy environments due to perceptual spectrum filtering, GFCC (Gammatone Frequency Cepstral Coefficients) – an alternative to MFCC, designed for more challenging acoustic conditions, DNN/CNN-based features – enable automatic, data-driven feature learning but demand significant computational resources, Prosodic features – serve as complementary inputs to spectral features, capturing suprasegmental information (intonation, rhythm, stress), STFT-based features (Short-Time Fourier Transform) – serve as a foundational representation for time-frequency analysis and underlie most spectral methods (e.g., MFCC, GFCC). Typically used as a pre-processing step rather than as standalone features, CQCC (Constant-Q Cepstral Coefficients) – apply a logarithmic frequency scale that aligns with musical tones and low-frequency speech components; frequently used in modern biometric security systems (Table 1).

The strengths and limitations of these techniques allow for selecting an appropriate trade-off tailored to specific application requirements and computational constraints [20].

The generalization of the table indicators leads to the following conclusions regarding the recommended applications and research directions for feature extraction methods in audio analysis:

– speech recognition: MFCC, LPCC, GFCC;

– music processing: CQCC, STFT;

– biometric authentication: CQCC for speaker recognition;

– medical diagnostics: DWT, CWT for ECG/EEG signal analysis;

– audio compression: DWT, e.g., in MPEG-4 encoding.

This work focuses on the analysis of deep neural architectures such as x-vector, ECAPA-TDNN, ResNet, CRNN, among others, as these models represent the state of the art (SOTA) in speaker verification and identification in both current research and practical systems.

Traditional classifiers such as CNN, RNN, and MLP are considered as building blocks within larger composite models and, therefore, are not separately featured in the comparative analysis (Table 2).

Classical machine learning classifiers such as SVM, K-NN, and HMM are excluded from the comparison due to their limited scalability and effectiveness in large-scale speaker recognition tasks.

The observed balance between accuracy and computational efficiency of STFT-based spectrograms and MFCCs ensures high performance with moderate resource consumption, which is essential for real-time applications and deployment on resource-constrained devices.

Moreover, the compatibility of these feature types with convolutional neural network (CNN) architectures motivates continued research on CNN-based models incorporating fine-tuning techniques on domain-specific or task-specific datasets.

Table 1. *Comparison of local feature extraction methods for speaker identification*

Type / Principle	**Typical Parameters / Input Data**	**Primary Application Area**	**Advantages**	**Limitations**
MFCC (Mel-Frequency Cepstral Coefficients)				
Spectral analysis using Mel-scale filterbanks, followed by cepstral transformation; short-time spectral analysis	Frame length: 20–40 ms; number of coefficients: typically 13–39	Application areas: speaker identification, speech recognition, voice password systems (e.g., ASV) (наприклад, ASV).	simplicity, computational efficiency, established standard in ASR systems, good performance on clean speech, incorporates perceptual characteristics of human hearing	sensitive to noise, limited temporal dynamics
LPCC, Linear Prediction Cepstral Coefficients				
Modeling of the vocal tract as a linear system for signal prediction	10–16 coefficients per frame	speaker identification	simplicity, effective in clean environments	low robustness to noise, limited relevance in modern ASR
PLP, Perceptual Linear Prediction				
Incorporates psychoacoustic properties (e.g., Bark scale critical bands) and spectral filtering	13–20 coefficients per frame	speaker identification, speech recognition	improved noise robustness compared to MFCC; more closely aligned with human auditory perception	still limited in temporal resolution; higher computational complexity than MFCC
GFCC, Gamma-tone Frequency Cepstral Coefficients				
Uses gammatone filterbanks to simulate the human auditory system	13–20 coefficients per frame	speaker identification in noisy environments, voice password systems	high noise robustness; well-suited for challenging acoustic conditions due to high frequency resolution	higher computational complexity compared to MFCC
CQCC, Constant Q Cepstral Coefficients				
Constant-Q transform with a logarithmic frequency scale	29–40 coefficients; long time windows	voice password systems, anti-spoofing, synthetic speech detection, speaker identification under adversarial conditions	high resolution at low frequencies; effective against spoofing and synthetic attacks	high computational cost; less commonly used in traditional ASR systems
Prosodic Features (e.g., pitch, duration, intensity, rhythm)				
Suprasegmental temporal features	computed over longer utterances (500 ms – 2 s)	speaker profiling, disfluency detection in ASR	convey speaking style and emotional state; complement spectral features; useful for speaker discrimination	not suitable for short utterances or passphrases due to context dependency
Spectrogram (based on STFT – Short-Time Fourier Transform)				
Time-frequency representation based on STFT; amplitude spectrogram	frame length 20–40 ms; FFT windowing	speaker identification, voice password systems	visual representation of energy distribution; rich information content	sensitive to noise; requires CNNs for feature learning
Wavelet Transform (DWT – Discrete Wavelet Transform, CWT – Continuous Wavelet Transform)				
Multilevel time-frequency analysis	choice of mother wavelet; multi-scale decomposition	speaker identification, anti-spoofing	captures short-term phenomena; adaptive localization in time and frequency	sensitive to parameter selection; lack of standardization

Table 2. *Analytical comparison of ai-based voice feature extractors*

Model	Principle	Input Features	Sensitivity To	Accuracy	Remarks and Resource Usage
Speaker Identification					
x-vector (TDNN)	Embeddings from Time Delay Neural Network	MFCC/STFT spectrogram	Noise, speech distortion	~90–95% (VoxCeleb1)	Stable model, well integrated in Kaldi. Does not capture long-term context. Moderate resource usage.
d-vector (LSTM)	Averaging LSTM-based embeddings	MFCC	Noise, signal length	~85–92% (VoxCeleb)	Captures temporal structure. Slow, less scalable. Moderate resource usage.
Siamese CNN	Speech spectrum comparison (embedding-distance)	STFT spectrogram	Quality of positive/negative pairs	~85–90%	Few-shot capability. Depends on pair generation. Moderate resource usage.
Voice Password (KWS – Keyword Spotting)					
CNN-KWS	CNN for keyword classification	MFCC/STFT spectrogram	Noise, pronunciation, speech rate	95–97% (Google Speech Commands)	Easy implementation, good performance. Moderate resource usage.
DS-CNN	Mobile-optimized deep CNN	MFCC	Background noise, limited vocabulary	92–95%	Optimized for mobile devices. Lower accuracy. Low resource usage.
CRNN	CNN + recurrent layers for temporal modeling	STFT spectrogram	Utterance length, tempo variation	96–98%	Context-aware. More complex training. Moderate resource usage.

Research results

The proposed voice access system enables both speaker identification (i.e., determining who is speaking) and speaker verification (i.e., confirming whether the correct passphrase has been spoken). The authentication process is organized in two sequential stages: first, the system verifies the speaker's identity, and then it checks the correctness of the spoken password.

The study focuses on evaluating the impact of two commonly used acoustic representations – the STFT-based spectrogram and the Mel-Frequency Cepstral Coefficients (MFCC) – on the accuracy of speaker identification and verification. The extracted features are analyzed using a convolutional neural network (CNN) architecture with the application of fine-tuning techniques to improve model generalization and performance (Figure 5).

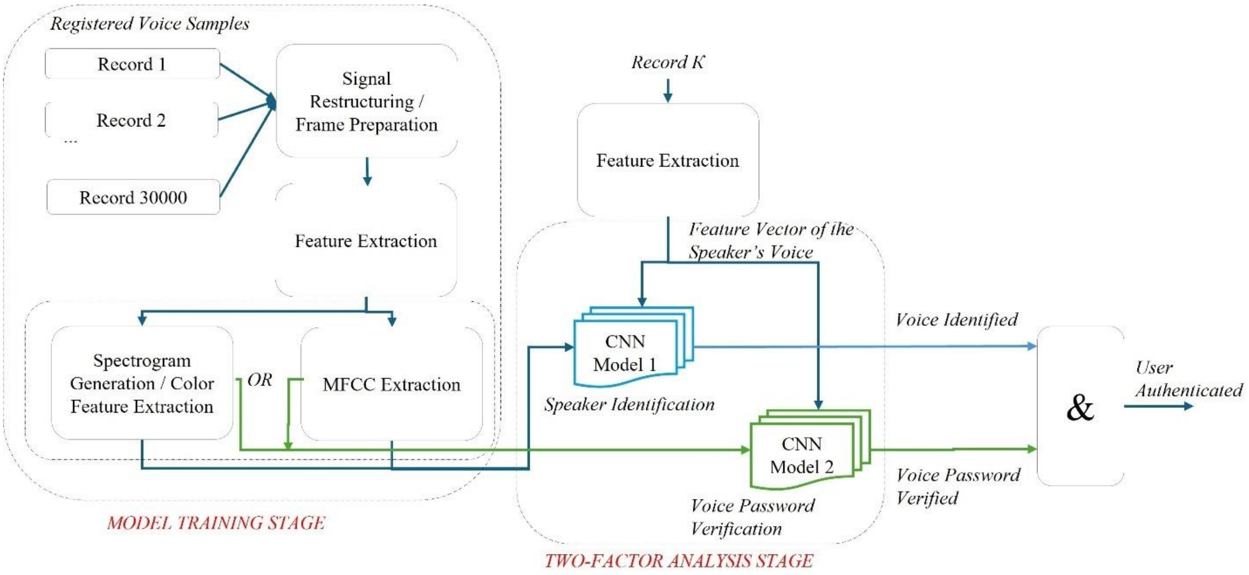


Fig. 5. Functional model of the proposed two-factor speaker authentication method

A series of experiments was conducted based on the proposed method to validate and analyze its performance. In particular, the effectiveness of two audio signal processing methods – spectrogram-based representation and Mel-Frequency Cepstral Coefficients (MFCC) – was evaluated and experimentally confirmed for the task of extracting salient acoustic features.

The study further examined the design of convolutional neural network (CNN)-based analyzers, the impact of dataset partitioning into training, validation, and test subsets, as well as the influence of training duration and hardware requirements on the system's stability and real-time applicability.

Neural network models with convolutional architectures were trained using a dataset organized as described in Table 3:

– Speaker_ID – speaker identifier; the dataset includes 60 speakers (labeled Speaker_1 through Speaker_60);

– Digit – the digit spoken by the speaker (from 0 to 9);

– File_ID – unique identifier for each audio file;

– Volume – recording volume level (low, medium, high);

– Pronunciation – articulation type (clear, muffled, fast, slow);

– Duration – length of the recording in milliseconds (ranging from 500 ms to 2499 ms).

Experiment 1 focuses on comparing the accuracy of speaker identification based on Mel-Frequency Cepstral Coefficients (MFCC) and STFT-based spectrograms. The processing sequence for implementing the system using MFCC features (Figure 6) is as follows:

– the initial dataset, consisting of WAV-format audio recordings of digits (0–9) pronounced in English 50 times by each of 60 speakers, is loaded into memory for further processing;

– each audio sample undergoes a resampling procedure to standardize the sampling rate across the dataset;

– MFCC features are extracted from each audio sequence, and the first 13 coefficients are stored in a data array;

– a feature set is formed by pairing the extracted MFCC vectors with their corresponding speaker identifiers. The resulting dataset is then split into three subsets: training, validation, and test;

– a convolutional neural network (CNN) is trained using the training and validation data subsets;

– the trained model is evaluated on the test set to determine the speaker identification accuracy.

Table 3. *Dataset characteristics for training neural network analyzers*

Speaker_ID	**Digit**	**File_ID**	**Volume**	**Pronunciation**	**Duration**
Speaker_1	0	File_1	low	clear	500ms
Speaker_1	1	File_2	medium	muffled	501ms
Speaker_1	2	File_3	high	fast	502ms
Speaker_1	3	File_4	low	slow	503ms
Speaker_1	4	File_5	medium	clear	504ms
...	...	...	...	...	...
Speaker_60	9	File_30000	high	slow	2499ms

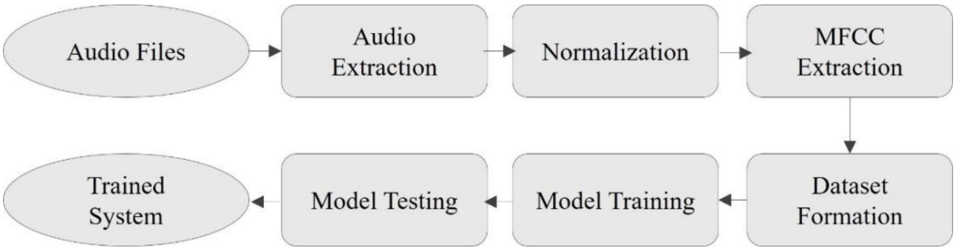


Fig. 6. Sequential steps for implementing the system using mel-frequency cepstral coefficients (MFCC)

The workflow for the system based on spectrograms (Fig. 7) differs in the initial preprocessing steps. Instead of extracting MFCC features, the following operations are performed:

– each audio signal is converted into a spectrogram using the Short-Time Fourier Transform (STFT), and the resulting image is resized to 305 × 184 pixels and stored for further use;

– the spectrogram image is then loaded and its color features are extracted across three channels (RGB) (Fig. 7);

– the extracted color channel data from the spectrograms are used to construct the training dataset.

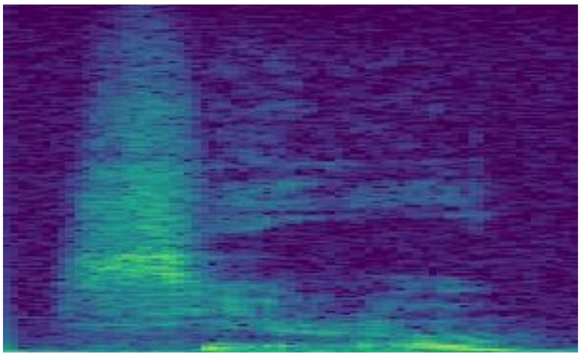


Fig. 7. Example of a spectrogram

To compare the accuracy of systems based on Mel-Frequency Cepstral Coefficients (MFCC) and spectrograms, two neural networks with identical configurations, architectures, and training parameters were developed and trained. The only difference between the two systems lies in the type of input features used by the neural network: MFCCs in the first case and spectrograms in the second.

Both models were trained using the same initial dataset, consisting of 30,000 WAV-format audio files. Each file contains the recording of one of ten digits, spoken by one of 60 different speakers, with each digit repeated 50 times per speaker under varying conditions of volume, pronunciation, and duration. The training was performed over the same number of epochs for both systems (Fig. 8). To evaluate the identification accuracy, both models were tested using a hold-out test dataset that was not included in the training process.

The results of evaluating training time and speaker identification accuracy depending on the type of feature representation (spectrogram or Mel-Frequency Cepstral Coefficients) using a convolutional neural network are presented in Table 4.

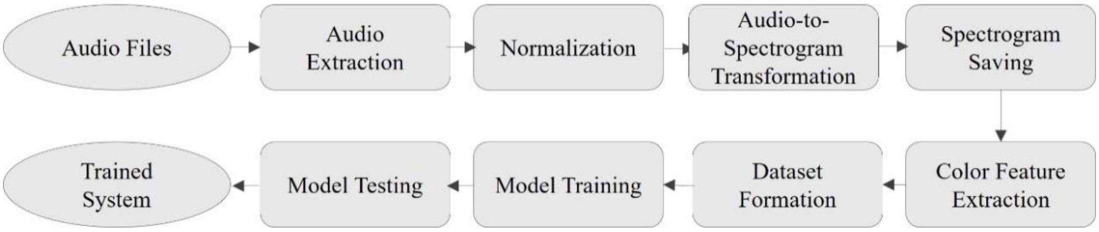


Fig. 8. Sequence of steps for implementing the system using spectrograms

Table 4. *Training results of systems using mel-frequency cepstral coefficients and spectrograms*

	Spectrogram			**Mel-Frequency Cepstral Coefficients (MFCC)**		
Number of Epochs	Training Time	Memory Usage	Accuracy	Training Time	Memory Usage	Accuracy
20	7 год. 48 хв.	15 Гб	73.13%	6 хв. 43 сек.	2 Гб	96.84%
60	23 год. 57 хв.	15 Гб	80.9%	14 хв. 58 сек.	2 Гб	97.64%

The model trained using Mel-Frequency Cepstral Coefficients (MFCCs) demonstrated 17% higher accuracy compared to the model trained on spectrograms. In addition, the training time for the MFCC-based model was 23 hours and 42 minutes shorter.

Accordingly, further experiments focused on analyzing the impact of different training/validation/test splits within the working dataset were conducted exclusively using MFCC features.

The proportion of validation data was gradually increased by reducing the training portion, in order to assess the relationship between dataset composition, training time, and model performance.

The results of these experiments (see Figure 9) show that training time decreases as the size of the training set decreases. This outcome is expected, as fewer training samples require less processing time per epoch.

However, the model accuracy also declines with a reduction in training data. This indicates that a larger training set improves the model's ability to make accurate predictions.

The highest accuracy, 97.4%, was achieved with a 70% training / 15% validation / 15% test split.

It is important to consider that training time and computational resources may be limited. Therefore, the choice of data split should be based on a trade-off between the desired accuracy and acceptable training time. The conducted experiment enables the observation of accuracy trends depending on the ratio between training, validation, and test subsets.

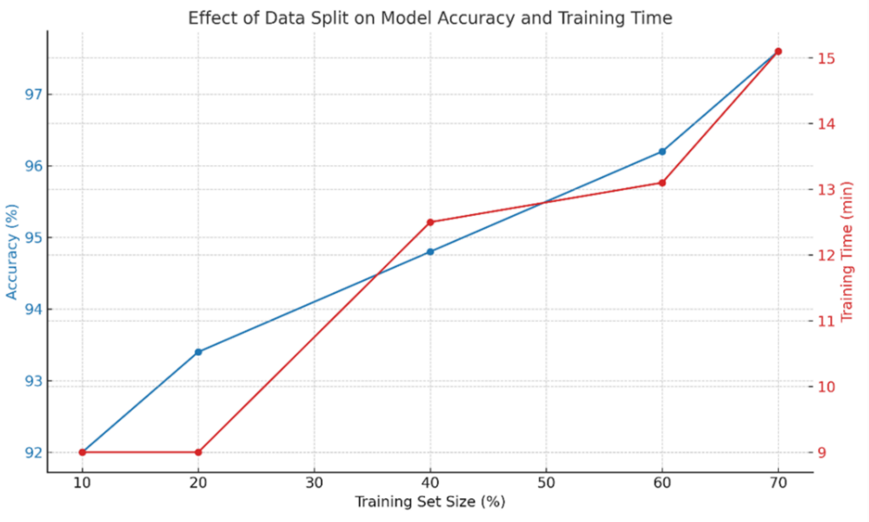


Fig. 9. Impact of data split proportions on model accuracy and training time

In the course of this work, an additional experiment was performed to determine the optimal sampling rate that ensures the highest accuracy of speaker identification by voice. The effect of various sampling rate values on model accuracy was investigated. The initial dataset consisted of audio files recorded at a sampling rate of 48 kHz. For comparative evaluation, the most common alternative sampling rates were also tested: 8 kHz, 16 kHz, 20.05 kHz, and 32 kHz. The results of this experiment are presented in Table 5.

Table 5. *Model accuracy as a function of audio sampling rate*

Sampling Rate (kHz)	Training Time	Accuracy (%)
8	14 хв. 57 сек.	97.62
16	15 хв. 13 сек.	98.02
20.05	16 хв. 59 сек.	97.84
32	17 хв. 05 сек.	98.06
48	17 хв. 52 сек.	98.28
88.2	19 хв. 35 сек.	97.77

Based on the results of the conducted experiments, it was determined that the most optimal sampling rate is 48 kHz, which provides the highest accuracy, as well as efficient training time and faster audio loading. Higher sampling rates capture more detailed information from the signal. In audio processing tasks, a higher sampling rate may help the model detect subtle nuances and fine-grained acoustic patterns, potentially improving accuracy. Although such high sampling rates are not traditionally used in speaker recognition tasks, each dataset and task may have unique characteristics that influence optimal system settings.

The experiments were also conducted to determine the fine-tuning configuration of the convolutional neural network (CNN) architecture – specifically, the number of convolutional layers, training epochs, and hyperparameter settings – to achieve the highest possible speaker identification accuracy while avoiding overfitting.

During the course of this work, a CNN was tested with varying numbers of convolutional layers, different training durations, and regularization techniques to mitigate overfitting.

The process began with a baseline architecture of three convolutional layers, each followed by MaxPooling and BatchNormalization layers to enhance learning stability and improve feature extraction accuracy. Each subsequent convolutional layer doubled the number of filters compared to the previous one, increasing the model's representational capacity and enabling it to capture more complex patterns.

Throughout the testing phase, the number of convolutional layers was gradually increased in order to determine the optimal architecture and number of training epochs. The initial model was trained on a dataset consisting of 60 speakers, using the optimal data split: 70% for training, 15% for validation, and 15% for testing.

The results of these experiments are summarized in Table 6.

Based on the results of preliminary testing, the optimal neural network architecture was determined to consist of four convolutional layers and 60 training epochs, providing the best trade-off between model accuracy and training time.

Table 6. *Evaluation of CNN architectures and training epoch counts*

Number of Layers	Number of Epochs	Training Time	Accuracy (%)
3	20	4 хв. 45 сек.	94.3
3	60	12 хв.	96.1
3	150	30 хв.	96.4
4	20	5 хв. 32 сек.	95.66
4	60	14 хв. 58 сек.	97.95
4	150	39 хв.	97.64
5	20	6 хв. 24 сек.	94.6
5	60	19 хв. 25 сек.	96.18
5	150	45 хв.	97.1
6	20	7 хв. 43 сек.	93.4
6	60	22 хв. 57 сек.	96.2
6	150	53 хв.	97.05

The achieved speaker identification accuracy was 97.95%, with a total training time of 14 minutes and 58 seconds.

A detailed description of the neural network architecture is provided in Table 7.

Table 7. *CNN architecture and configuration details*

Layer Type	Output Dimension	Number of Parameters
Conv2D	(None, 87, 13, 32)	320
MaxPooling2D	(None, 44, 7, 32)	0
BatchNormalization	(None, 44, 7, 32)	128
Conv2D	(None, 44, 7, 64)	18496
MaxPooling2D	(None, 22, 4, 64)	0
BatchNormalization	(None, 22, 4, 64)	256
Conv2D	(None, 22, 4, 128)	73856
MaxPooling2D	(None, 11, 2, 128)	0
BatchNormalization	(None, 11, 2, 128)	512
Conv2D	(None, 11, 2, 256)	295168
MaxPooling2D	(None, 6, 1, 256)	0
BatchNormalization	(None, 6, 1, 256)	1024
Flatten	(None, 1536)	0
Dense	(None, 128)	196736
Dropout	(None, 128)	0
Dense	(None, 61)	7869

As a result of fine-tuning the model on a dataset of 3,000 audio recordings from six new speakers, the recognition accuracy significantly improved due to the increased diversity of training data.

An accuracy level of 99.5% was achieved, while the training time remained nearly unchanged, enabling efficient integration of new users within a relatively short period of time.

The second component of the proposed system (Figure 5) is responsible for voice password verification. The authentication principle is as follows: the user is considered authenticated only if both the voice is correctly identified and the spoken password matches the one stored in the database; otherwise, access is denied.

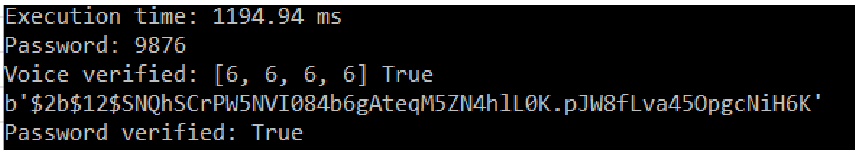
At this stage, the analyzer is implemented as a convolutional neural network (CNN) with four convolutional layers, trained over 20 epochs.

The main difference from the architecture used for speaker identification lies in the output layer – in the verification model, it contains ten output neurons, corresponding to the ten digits that need to be recognized as part of the personal voice password (a digit sequence). A shared dataset consisting of 30,300 audio files from 66 different speakers was used for training. In this case, each audio file corresponds to one of ten digits. The training results of the digit recognition model are summarized in Table 8.

Table 8. *Training results of the neural network for spoken digit recognition*

Number of Layers	Number of Epochs	Training Time	Accuracy (%)
4	20	6 хв. 22 сек.	99.82
4	60	18 хв. 4 сек.	99.4

The model trained with fewer epochs demonstrated higher accuracy, as the network trained over 60 epochs encountered overfitting on the initial dataset. The result of the complete multifactor authentication system for one of the users – who was successfully identified by voice and verified through correct password recognition – is illustrated in Figure 10.



```
Execution time: 1194.94 ms
Password: 9876
Voice verified: [6, 6, 6, 6] True
b'$2b$12$SNQhSCrPW5NVI084b6gAteqM5ZN4hlL0K.pJW8fLva45OpgcNiH6K'
Password verified: True
```

Fig. 10. Output of the multifactor authentication method

Based on the console output, a general conclusion can be drawn confirming the successful execution of all experimental stages:

– the total processing time was 1194.94 ms, indicating high execution efficiency;

– the password "9876" was successfully verified;

– voice verification was successfully completed, yielding output values of [6, 6, 6, 6], which confirm the authenticity of the speaker;

– the password was securely hashed using the bcrypt algorithm, ensuring a high level of cryptographic security;

– the password was successfully matched and validated, confirming its correctness and consistency with the expected value.

These results demonstrate that the proposed system effectively performs both password processing and voice verification, delivering a high degree of security and reliability. Among the key advantages of the proposed solution is the maximum voice password verification accuracy, which reached 99.82% using a convolutional neural network with four layers, trained for 20 epochs over a duration of 6 minutes and 22 seconds, with MFCC features used for acoustic feature extraction. The maximum speaker identification accuracy achieved was 97.95%, using the same CNN architecture with four convolutional layers, trained for 60 epochs over 14 minutes and 58 seconds. The average speaker authentication accuracy of the proposed two-factor biometric authentication system is 98.75%.

Conclusions

The relevance of the research is due to the growing number of people with visual impairments and the need to create As a result of this work, a two-factor voice authentication method was developed and evaluated. It combines speaker identification based on acoustic voice characteristics with keyword verification (voice password recognition). It is built using convolutional neural networks (CNNs) and analyzes two types of spectral features: STFT-based spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs).

A series of experiments was conducted to compare the accuracy, training time, and computational resource requirements associated with each feature type.

The highest speaker identification accuracy – 97.95% – was achieved using MFCC features in a four-layer CNN architecture trained for 60 epochs. The voice password verification model demonstrated even greater effectiveness, reaching 99.82% accuracy with the same architecture trained over 20 epochs. The overall average accuracy of the proposed two-factor method was 98.75%, which is a competitive result for modern biometric security applications.

The use of MFCCs was shown to reduce training time by a factor of 23 and memory consumption by a factor of 7, compared to models using STFT spectrograms. This makes MFCCs the preferred choice for real-time applications, especially on resource-constrained devices.

The method was validated on a synthetic dataset of over 30,000 audio recordings from 60 speakers, and was further fine-tuned with new user data, achieving an identification accuracy of up to 99.5%, demonstrating its adaptability.

Analysis of sampling rate impact revealed that 48 kHz offers the best trade-off between accuracy (98.28%) and computational efficiency.

The method is of particular practical value in the context of human-computer interaction for individuals with limited mobility or bedridden patients, as it enables contactless and keyboard-free access to digital devices via voice alone.

The findings confirm that the proposed method can be successfully integrated into assistive technologies or home automation solutions to enable reliable, accurate, and adaptive biometric voice authentication.

The scientific novelty of the study lies in the comprehensive analysis of the effectiveness of spectral features – MFCC and STFT spectrograms – for the simultaneous tasks of speaker identification and keyword-based voice verification (KWS) within a unified convolutional neural network architecture, with a focus on fine-tuning techniques.

The practical significance of the research lies in the fact that its results can be directly applied to the development of real-world contactless authentication systems, particularly:

– in access systems for individuals with physical disabilities, where traditional input methods are not feasible;

– in medical or household devices that require hands-free interfaces;

– in mobile or resource-constrained environments, where computational efficiency is critical – the use of MFCC reduces training time by a factor of 23 and memory usage by a factor of 7 compared to STFT;

– in biometric security systems where high accuracy is essential (97.95% for identification and 99.82% for verification), along with adaptability to new users (achieving up to 99.5% accuracy after fine-tuning).

Future research directions aim to expand the functionality and robustness of the proposed method. In particular, the study will focus on implementing few-shot learning techniques for dynamic adaptation to new users, and enhancing method noise robustness using spectral denoising and data augmentation strategies. Additionally, it is proposed to integrate alternative spectral and suprasegmental features (e.g., CQCC, GFCC, and prosodic parameters) to improve accuracy and resistance to spoofing attacks.

Further efforts will be directed toward optimizing the model for energy-efficient deployment on portable and embedded devices, and extending the method's capabilities toward multimodal biometric authentication, combining voice with gaze tracking or emotion analysis. The proposed method is also planned to be tested in real-world human-machine interaction scenarios, particularly for users with limited mobility or speech impairments.

References

1. Mourtzis, D., Angelopoulos, J., Panopoulos, N. (2023), "The Future of the Human–Machine Interface (HMI) in Society 5.0". *Future Internet,* № 15, 162 p. DOI: https://doi.org/10.3390/fi15050162
2. Grobelna, I., Mailland, D., Horwat, M. (2025), "Design of Automotive HMI: New Challenges in Enhancing User Experience, Safety, and Security". *Appl. Sci*. № 15, 5572 p. DOI: https://doi.org/10.3390/app15105572
3. Esquivel, P. et al. (2024), "Voice Assistant Utilization among the Disability Community for Independent Living: A Rapid Review of Recent Evidence", *Human Behavior and Emerging Technologies*, Vol. 2024, №. 1, 6494944 p. DOI: https://doi.org/10.1155/2024/6494944
4. Semary, H. E., Al-Karawi, K. A. (2024), "Abdelwahab M. M. Using voice technologies to support disabled people", *Journal of Disability Research,* 2024. Vol. 3. №. 1. DOI: https://doi.org/10.57197/jdr-2023-0063
5. Lawrence, I. D., Pavitra, A. R. R. (2024), "Voice-controlled drones for smart city applications", *Sustainable Innovation for Industry 6.0*. P. 162–177. DOI: DOI: 10.1109/ICUFN.2017.7993759
6. Ryu, R., Yeom, S., Kim, S. H., Herbert, D. (2021), "Continuous multimodal biometric authentication schemes: a systematic review", *IEEE Access*. Vol. 9. P. 34541–34557. DOI: 10.1109/ACCESS.2021.3061589
7. Barkovska, O., Liapin, Y., Muzyka, T., Ryndyk, I., Botnar, P. (2024), "Gaze direction monitoring model in computer system for academic performance assessment. Civil law aspect", *Information Technologies and Learning Tools*, Vol 99, №1, P. 63–75. DOI: 10.33407/itlt.v99i1.5503
8. Shaheed, K., Mao, A., Qureshi, I. et al. (2021), "A Systematic Review on Physiological-Based Biometric Recognition Systems: Current and Future Trends". *Arch Computat Methods Eng 28*, P. 4917–4960. DOI: https://doi.org/10.1007/s11831-021-09560-3
9. Sasongko, S. M. A., Tsaury, S., Ariessaputra, S., Ch, S. (2023), "Mel Frequency Cepstral Coefficients (MFCC) Method and Multiple Adaline Neural Network Model for Speaker Identification". *International Journal on Informatics Visualization*, № 7(4), P. 2306–2312. DOI: https://doi.org/10.30630/joiv.7.4.1376
10. Desplanques, B., Thienpondt, J., & Demuynck, K. (2020), "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification". *In Interspeech 2020*, P. 3830–3834. DOI: https://doi.org/10.21437/Interspeech.2020-2650
11. Jahangir, R., Alreshoodi, M., Alarfaj, F. K. (2025), "Spectrogram Features-Based Automatic Speaker Identification for Smart Services". *Applied Artificial Intelligence,* № 39(1). DOI: https://doi.org/10.1080/08839514.2025.2459476
12. Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., Wang, R. (2017), "Speaker Identification Features Extraction Methods: A Systematic Review". *Expert Systems with Applications,* № 90, P. 250–271. DOI: https://doi.org/10.1016/j.eswa.2017.08.015
13. Iliev, Y.; Ilieva, G. (2023), "A Framework for Smart Home System with Voice Control Using NLP Methods". *Electronics 2023,* № 12, 116 p. DOI: https://doi.org/10.3390/electronics1201011614
14. Kim, Y., Hyon, Y., Lee, S., Woo, S. D., Ha, T., Chung, C. (2022), "The coming era of a new auscultation system for analyzing respiratory sounds", *BMC Pulmonary Medicine,* Vol. 22, №. 1. 119 p. DOI: 10.1186/s12890-022-01896-1

15. Barkovska, O, Havrashenko, A. (2024), "Research of the impact of noise reduction methods on the quality of audio signal recovery", *Information and control systems at railway transport,* 2024, Vol. 29, №. 3. P. 57–65. DOI: https://doi.org/10.18664/ikszt.v29i3.313606

16. Zaman, K., Sah, M., Direkoglu, C., Unoki, M. (2023), "A Survey of Audio Classification Using Deep Learning", *IEEE Access*, Vol. 11, P. 106620–106649. DOI: 10.1109/ACCESS.2023.3318015

17. Xie, X., Cai, H., Li, C., Wu, Y., Ding, F. (2023), "A Voice Disease Detection Method Based on MFCCs and Shallow CNN", *Journal of Voice*, Oct. 2023, DOI: https://doi.org/10.1016/j.jvoice.2023.09.024

18. Tu, Y., Lin, W., Mak, M. W. (2022), "A survey on text-dependent and text-independent speaker verification", *IEEE Access*. Vol. 10. P. 99038–99049. DOI: DOI: 10.1109/ACCESS.2022.3206541

19. Luitel, Sophina, Mohd, Anwar. (2022), "Audio Sentiment Analysis Using Spectrogram and Bag-of-Visual- Words", *IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), IEEE,* P. 200–205. DOI: https://doi.org/10.1109/IRI54793.2022.00052

20. Singh, V. K., Sharma, K., Sur, S. N. (2023), "A survey on preprocessing and classification techniques for acoustic scene", *Expert Systems with Applications*, Vol. 229, 120520 p. DOI: https://doi.org/10.1016/j.eswa.2023.120520

21. Labied, M., Belangour, A., Banane, M., Erraissi, A. (2022), "An overview of Automatic Speech Recognition Preprocessing Techniques", *2022 International Conference on Decision Aid Sciences and Applications (DASA),* Chiangrai, Thailand, P. 804–809, DOI: 10.1109/DASA54658.2022.9765043

Надійшла (Received) 25.05.2025

Відомості про авторів / About the Authors

Barkovska Olesia – Ph.D (Engineering Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor of the Department of Electronic Computers, Kharkiv, Ukraine; e-mail: olesia.barkovska@nure.ua; ORCID ID: https://orcid.org/0000-0001-7496-4353

Барковська Олеся Юріївна – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри Електронних обчислювальних машин, Харків, Україна.

ДВОФАКТОРНА АВТЕНТИФІКАЦІЯ НА ОСНОВІ МЕТОДУ KWS ТА ГОЛОСОВОЇ ВЕРИФІКАЦІЇ

Предметом статті є розробка та оцінка двофакторного методу автентифікації мовця на основі ідентифікації голосового відбитка та верифікації ключових слів (KWS), призначеного для безпечного голосового доступу в інтерфейсах «людина-машина», особливо для користувачів з обмеженою мобільністю. **Метою** роботи є створення методу управління автентифікацією мовця з використанням конволюційних нейронних мереж (CNN), порівняння ефективності двох широко використовуваних методів вилучення спектральних ознак – спектрограм Мел-частотних кепстральних коефіцієнтів (MFCC) та короткочасного перетворення Фур'є (STFT). У статті вирішено такі **завдання**: запропоновано модель двофакторного методу автентифікації, що включає ідентифікацію мовця та розпізнавання голосового пароля; порівняно якість ознак спектрограм MFCC та STFT; оцінено вплив кількості епох, архітектури CNN та параметрів навчання на точність системи; досліджено вплив частоти дискретизації на продуктивність моделей. Використовуються такі **методи**: методи глибокого навчання з архітектурою CNN, точне налаштування, вилучення ознак MFCC та STFT, математичний та статистичний аналіз ефективності навчання та показники продуктивності системи. Отримано такі **результати**: метод досяг 97,95% точності в ідентифікації мовця за допомогою MFCC після 60 епох навчання та 99,82% точності в перевірці голосового пароля за допомогою тієї ж структури CNN після 20 епох. Середня точність всього процесу автентифікації становила 98,75%. Більше того, використання MFCC-ознак дозволило скоротити час навчання в 23 рази, а споживання пам'яті – в 7 разів порівняно зі спектрограмами STFT. **Висновки**: було реалізовано та досліджено ефективність двофакторного методу голосової автентифікації, що поєднує ідентифікацію мовця за акустичними характеристиками голосу та перевірку голосового пароля. Подальші напрямки досліджень включають вивчення впливу альтернативних спектральних характеристик (зокрема, CQCC, GFCC, просодичних параметрів) на підвищення точності та стійкості до підробки. Особлива увага буде приділена оптимізації моделі для енергоефективного використання на портативних пристроях.

Ключові слова: олосова автентифікація, ідентифікація, голосовий пароль, MFCC, спектрограма, CNN, МНІ, біометрія.

Бібліографічні описи / Bibliographic descriptions

Барковська О. Ю. Двофакторна автентифікація на основі методу KWS та голосової верифікації. *Сучасний стан наукових досліджень та технологій в промисловості*. 2025. № 3 (33). С. 5–18. DOI: https://doi.org/10.30837/2522-9818.2025.3.005

Barkovska, O. (2025), "Two-factor authentication based on keyword spotting and speaker verification", *Innovative Technologies and Scientific Solutions for Industries*, No. 3 (33), P. 5–18. DOI: https://doi.org/10.30837/2522-9818.2025.3.005