System for detecting voice deepfake attacks

A.R. Fakhretdinov, A.M. Vulfin, A.D. Kirillova, A.V. Minko

Ufa University of Science and Technology, Ufa

Abstract: This paper addresses the limited accuracy of existing automatic systems for detecting deepfake audio content in real time. A solution is proposed to increase the efficiency of detecting signs of deepfake use by improving neural network models and algorithms for analyzing audio recordings of human voices. An algorithm and corresponding software for a voice attack detection system have been developed. For training and testing, datasets were created containing real voice audio recordings and deepfake audio samples. Evaluation on a real-world test set demonstrated an accuracy rate of 83%, confirming the effectiveness and practical applicability of the proposed solution in combating audio deepfake threats.

Keywords: deepfake, speech audio signal, machine learning models, convolutional neural network, vishing.

Introduction

Voice-based social engineering (vishing) attacks increasingly rely on deepfake models powered by deep neural networks. Extracting and then using key characteristics of a person's voice allows attackers to create audio messages with arbitrary content that are aurally indistinguishable from the real voice and manner of speech of the copied subject. However, automatically detecting deepfake attacks based on real-time analysis of audio message fragments is an important measure to combat social engineering attacks.

The number of cyberattacks using deepfake technology has increased more than 10-fold worldwide by 2024 [1-3]. This trend indicates growing accessibility of deepfake technologies are becoming increasingly accessible to attackers and raise serious security concerns.

Therefore, enhancing the accuracy of real-time deepfake voice detection remains a critical task in audio forensics.

The goal of this research is to increase the efficiency of detecting signs of deepfake use by improving neural network models and algorithms for analyzing audio recordings of human voices.

Analysis of approaches to detecting voice deepfake attacks

The uniqueness of the voice is determined by the anatomical features of the vocal tract. Acoustic characteristics of the voice are more reliable indicators, since diction, intonation, pronunciation, rhythm and other behavioral traits can vary greatly depending on the situation and are conditioned by social factors.

In the field of voice deepfake attack detection, several key approaches are distinguished. The key approaches are presented in Table № 1.

Table № 1

Comparison of approaches to detecting voice deepfake attacks

Approach	Characteristic	Advantages	Flaws
Analysis of psychophysical parameters	Evaluation of intonation, speech rate, pauses and emotional expressions.	High individuality of parameters, difficulty of precise counterfeiting.	Dependence on emotional state, instability of characteristics even in the same person. The method is not sufficiently developed for mass application.
Identification of semantic features	Analysis of logical stress, syntactic structures, speech style, semantic patterns.	High semantic uniqueness; can be useful for long-term monitoring and personalized security.	High-quality speech transcription required; difficulty of automatic analysis; linguistically complex; language dependence.
Acoustic signal analysis	Analysis of audio features: mel-frequency cepstral coefficients (MFCC), spectrogram, signal smoothing, sampling rate, presence of synthesis artifacts, editing traces.	High accuracy; compatible with neural network methods; real-time applicability.	Dependent on recording and microphone quality; performance degrades in noisy environments; requires processing power.

For automatic speech recognition, algorithms based on acoustic characteristics are most often used [4]. Such systems have become widely used both for automating the decoding of audio and video recordings and for biometric identification tasks (including multimodal) in the financial sector to combat cybercrime. Despite the promise of the first two approaches, their large-scale application is challenging due to insufficient study of the corresponding cognitive and psychophysiological aspects. For this reason the third approach focused on detecting anomalies in the structure of the acoustic signal is currently considered the most effective and practical solution [5].

Depending on the mathematical and algorithmic methods used, approaches to voice signal processing can be divided into three main categories [6]: frequency, time, and time-frequency.

Studies [7, 8] have shown that convolutional neural networks (CNNs) can be effectively used for detecting voice deepfake attacks (Table № 2). Various characteristics are used to analyze audio files: spectrograms, mel-spectrograms, chromograms, mel-cepstral coefficients. The effectiveness of neural network models based on Long Short-Term Memory (LSTM) is also noted – a recurrent neural network whose neurons have feedback [9]. LSTM neural networks are designed to work with temporal dependencies in an audio signal. However, the results of the detection efficiency of voice deepfake forgeries of LSTM-based neural network models are not presented.

The presented architecture of a CNN, featuring four convolution layers with 3x3 pixel filters, progressive down sampling with a stride 2, Softmax activation, and the Adam optimizer, achieves the highest accuracy in detecting voice deepfakes. CNNs can achieve high accuracy of up to 99% in detecting audio deepfakes.

In the work [10], a multilayer perceptron is used to build a deepfake detection model. The neural network was trained on a small and non-diversified dataset, which contained 8 voices, the total length of audio recordings of real voices was one hour, and the total length of audio recordings of fake voices was 8 hours. To train the neural network, only the MFCC averaged over the entire audio recording time were used, which leads to a loss of temporal information. The accuracy of the model after training was 86.9% on the training set without using the cross-validation scheme. The neural network recognizes fake audio recordings of familiar voices with 86.9% accuracy, but does not recognize fakes of other voices at all.

Table № 2

Comparison table of key technology

Research	Model	Input Features	Accuracy, % (on training set)
[7], [8]	CNN	Images of spectrograms, mel-spectrograms, chromograms	99
[10]	Multilayer Perceptron (MLP)	Mel-cepstral coefficients MFCC	86.9
[11]	Neural network model based on LSTM	20 MFCC, Zero Crossing Rate, oot Mean Square Energy, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, Chroma Short-Time Fourier Transform	98.4
	Support Vector Machine		91.2
	k-nearest neighbors method		80.6
	Logistic Regression Method		75.3

In the work [11] a neural network model based on LSTM was used. It was trained on the same dataset as the model [10]. However, 26 features were extracted from the audio files. The accuracy of the model after training was 98.4%, but this assessment was made on the same voices used to train the neural network model. Models based on decision tree committees (CatBoost and RandomForestClassifier) were used, achieving an accuracy of 98.7%, respectively. The support vector machine showed 91.2% accuracy on voice audio recordings from the training dataset. The k-nearest neighbors method showed 80.6% accuracy on voice audio recordings from the training dataset (the evaluation was conducted without using cross-validation). The logistic regression method showed 75.3% accuracy on voice audio recordings from the training dataset.

Commercial voice deepfake detection software solutions exist (e.g., Resemble.ai [12], Sensity.ai [13], Mts.ai [14], and Arya.ai [15]), but they are not available for study or testing.

Development of a system for detecting voice deepfake attacks

A structural and functional organization of a system for detecting voice deepfake attacks based on a composition of neural networks is proposed. The proposed algorithms and model for analyzing speech audio recordings are implemented in the form of software with a microservice architecture, followed by evaluation of their efficiency on real data. The proposed solution consists of a group of containerized, each of which implements the functionality of preprocessing, feature extraction and running machine learning models to analyze the selected set of features.

The structural diagram of the voice deepfake attack detection system comprises several components.

Data Loader. Module for loading and pre-processing initial data (the data are datasets containing real and deepfake voice audio recordings for training the neural network model and for testing the already trained neural network model).

Feature Extraction Module. MFCC, Zero Crossing Rate, Root Mean Square Energy, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, Chroma Short-Time Fourier Transform features are extracted, or images of mel-spectrograms are constructed and these are then recorded in the DB_1.

Model Management Module. Module for monitoring the training and management of neural network models, which can be built on the basis of such neural networks as MLP, LSTM (recurrent neural network, for identifying temporal dependencies in voice audio recordings) for analyzing mel-cepstral coefficients, as well as on the basis of CNN for recognizing images of mel-cepstral coefficients and mel-spectrograms. Trained neural network models for analyzing voice audio recordings are recorded in the DB_2.

Binary Classification Module. Module for voice audio recordings that will determine whether the analyzed voice audio recording belongs to the class of deepfake voices or to the class of real voices.

The structural diagram of the neural network system for detecting voice deepfake attacks is shown in fig. 1.

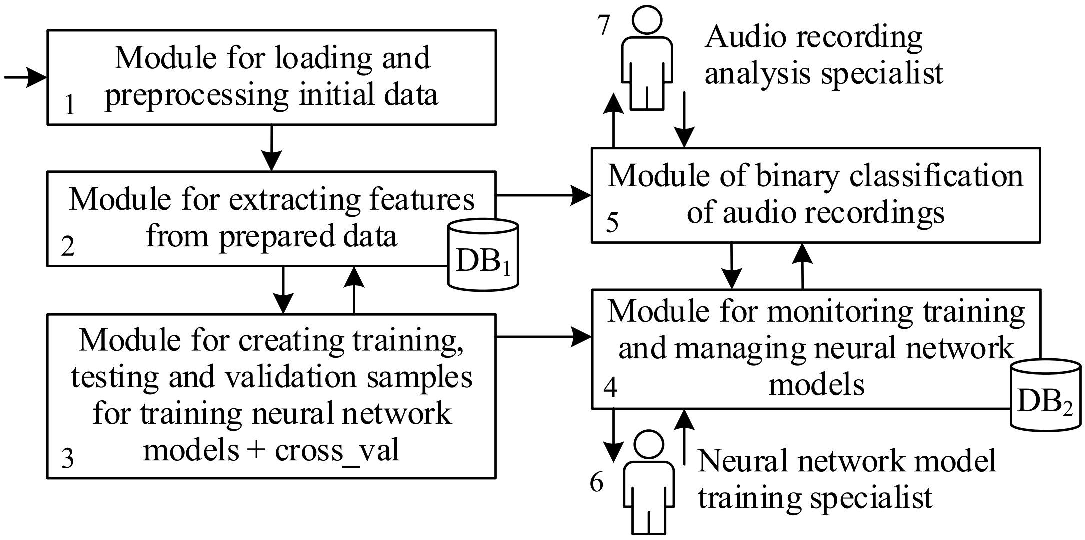

Module for loading and preprocessing initial data
1
Module for extracting features from prepared data
2
DB1
Module for creating training, testing and validation samples for training neural network models + cross_val
3
7
Audio recording analysis specialist
Module of binary classification of audio recordings
5
Module for monitoring training and managing neural network models
4
DB2
6
Neural network model training specialist

Fig. 1. – Structural diagram of a neural network system for detecting voice deepfake attacks

The functional diagram of the CNN-based voice deepfake attack detection system is shown in fig. 2.

The components in fig. 2 are described as follows:

1 – training dataset of voice audio recordings;

2 – test dataset of voice audio recordings;

3 – module for loading and preprocessing training data;

4 – CNN model;

5 – module for loading and preprocessing test data;

6 – decision module (binary classifier);

7 – mel-spectrogram images generated from training dataset audio segments;

8 – mel-spectrogram images generated from test dataset audio segments;

9 – trained CNN model;

10 – prediction output for analyzed audio recording.

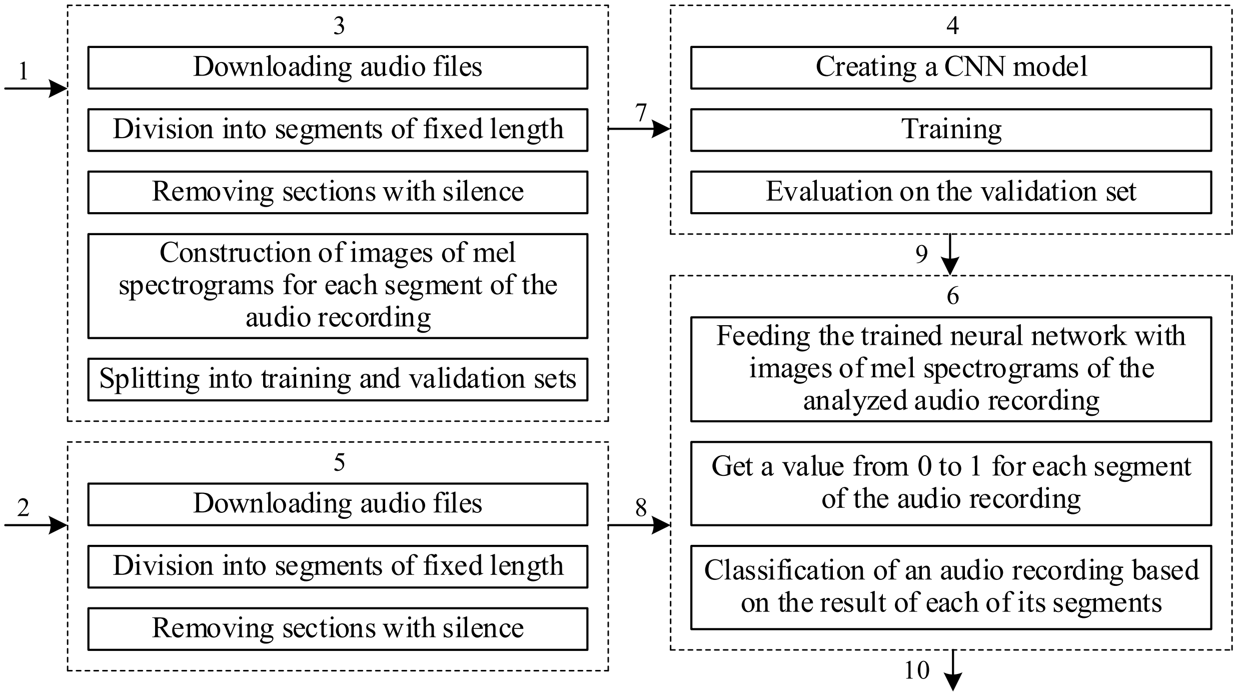

3
Downloading audio files
Division into segments of fixed length
Removing sections with silence
Construction of images of mel spectrograms for each segment of the audio recording
Splitting into training and validation sets
1
4
Creating a CNN model
Training
Evaluation on the validation set
7
9
6
Feeding the trained neural network with images of mel spectrograms of the analyzed audio recording
Get a value from 0 to 1 for each segment of the audio recording
Classification of an audio recording based on the result of each of its segments
5
Downloading audio files
Division into segments of fixed length
Removing sections with silence
2
8
10

Fig. 2. – Functional diagram of the CNN-based voice deepfake attack detection system

In addition to the images of the mel-spectrograms, images of the mel-cepstral coefficients for each segment of the audio recording are constructed.

Computational experiment on natural data

For the computational experiment, a dataset was prepared based on speech recordings of real users. The dataset used to train the models (the average length of one audio recording is 15 minutes) contains:

– the voices of 62 real speakers (equal male and female, 3 children’s voices [16]). The total duration of the audio recordings is 15 hours;

– 29 voices of fake speakers (108 audio recordings) with a total duration of about 25 hours.

The demo dataset (average length of one audio recording is 1 minute) contains

– 43 real voices (equally classified into bass, baritone, tenor, alto, mezzo-soprano, soprano, children's) with a total duration of about 43 minutes;

– 25 fake voices (320 audio recordings) with a total duration of about 5 hours.

To increase the diversity of the test data, an open-access English-language dataset consisting of 8 authentic and synthetic voice samples was used [17], so it was decided to create more complete Russian-language datasets. To assess the stability of the model, two sets of voice audio recordings were created: training and demonstration. The training dataset is used to train the neural network model, while the demo dataset containing completely different voices from those in the training dataset is used to test the trained model. This is necessary to test the generalization ability of the trained neural network model.

To create voice deepfakes, the(Retrieval-based Voice Conversion Version 2 (RVC v2) model was selected – the most well-known and accessible AI-based technology for voice copying, an open-source project [18-20]. The RVC v2 model transforms one person's voice into the sound of another's (speech to speech), while maintaining the intonation, tempo and other features of the original speech. RVC v2 uses a pre-trained Hidden-Unit BERT (HuBERT) model for feature extraction. It transforms the input audio signal into a sequence of features reflecting the phonetic content of speech. HuBERT is trained to predict masked latent representations, making it robust to variations in speaker voice. An acoustic model based on the Variational Inference Text-to-Speech architecture generates an audio signal of the target speaker, taking content and pitch features as input. Variational Inference Text-to-Speech incorporates variational inference and generative modeling capabilities for high-quality speech synthesis [21]. A distinctive feature of the RVC v2 model is the use of a retrieval mechanism, which, during speech reproduction, finds the most similar speech segments in the target speaker’s database. This allows for more accurate reproduction of the unique characteristics of the target’s voice and improves the quality of synthesized speech.

Subsequently after training the neural network model and testing it on the demo dataset, the next step is to evaluate its performance on voice deepfakes created by other models available online.

Below is a summary table № 3 of all created datasets (except for the English-language dataset from the Kaggle service) indicating the deepfake generation method, the number of voices, the number of audio recordings and their duration.

Table № 3

Overview of voice audio datasets used in the study

Dataset		Generation method	FAKE / REAL	Number of voices	Number of voice audio recordings	Average length of voice audio recording	Total length of audio recordings
Training, model RVC v2		Voice conversion speech to speech	REAL	62	62	15 min	15 h 29 min
			FAKE	29	108		25 h
Demo, model RVC v2			REAL	43	43	1 min	43 min
			FAKE	25	320		5 h 20 min
Dataset from Kaggle, model RVC v2 (English)			REAL	8	8	8 min	1 h 2 min
			FAKE	8	56		7 h 16 min
Internet services	ResembleAI		FAKE	5	46	1 min	45 min
	OpenAI fm	Text to speech	FAKE	11	62	1 min	1 h 7 min
	AnyVoice lab		FAKE	11	11	45 sec	9 min
	Typecast.ai		FAKE	22	22	25 sec	9 min
	Speechify		FAKE	28	44	25 sec	18 min

AnyVoice lab and Typecast.ai [22] datasets are easier to perceive by humans because they have unnatural pauses, stresses, intonations, and sound distortions.

Of particular interest is the detection of voice deepfakes generated using the RVC v2 model, since they sound are easier to perceive by humans compared to deepfakes made with other models. This model is widely accepted and used in many studies in the subject area. To demonstrate how humans can recognize

RVC v2 deepfake voice models, an auditory experiment was conducted with 11 participants. 32 deepfake voice audio recordings were selected from the demo dataset. The subjects were asked to listen to each audio recording once in a quiet environment and give one of three answers regarding each audio recording: “REAL”, “FAKE”, “DON’T KNOW”.

An assessment was made to determine the extent to which experts' evaluations were consistent using the following calculation:

– Fleiss’ coefficient [23] $\kappa = 0.086$ – experts generally agreed only slightly better than chance;

– Cronbach’s coefficient [24] $\alpha = 0.624$ – the internal agreement of the expert scales is below desired.

The obtained estimates indicate that:

– either the classes themselves are difficult to perceive by humans;

– or there are too many experts and each has their own “style” of marking;

– the experts have a common tendency, but many discrepancies.

It is necessary to take into account both the quality of experts and the hidden true label, therefore the following evaluation models are used:

– The Dawid-Skene Expectation-Maximization model [25] estimates prior class probabilities (π), each expert’s confusion matrices (θ), and posterior distributions of the true labels for each example:

1. It is assumed that there is a true label z_i.

2. For each expert j the error matrix $P(label = k \mid z = z')$ is estimated.

3. The Expectation-Maximization algorithm evaluates the parameters – both the “correctness” of the experts and the true labels.

– The Latent Class Analysis (LCA) model [26] implements Expectation-Maximization for a mixture model of multinomial distributions.

Table № 4 presents how each model classified the 32 examples into the three classes (“REAL”, “FAKE”, “DON'T KNOW”).

Table № 4

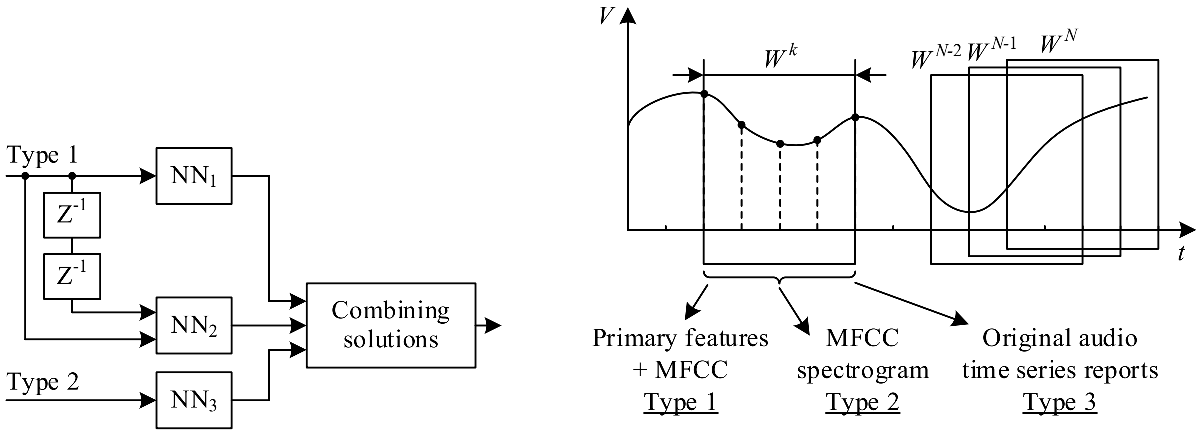
Assessment of the aggregation of expert decisions

Label	Dawid-Skene	LCA
DON’T KNOW	2	–
REAL	15	–
FAKE	15	32

The Dawid-Skene model assigned labels across all three categories, with 2 samples classified as “DON’T KNOW” due to high uncertainty, and the remaining examples nearly evenly split between “REAL” and “FAKE”.

LCA showed that there was very little difference between groups in the data. It was easier for the model to lump everything into one category. This again points to low inter-rater agreement and relative homogeneity of the labelling.

The experimental pipeline for constructing machine learning models for detecting deepfake is presented in fig. 3, where NN_1 – neural network models based on MLP, CNN 1D; NN_2 – neural network models based on LSTM; NN_3 – neural network models based on CNN 2D.



a) b)

Fig. 3. – The experimental pipeline: (a) classifier experimental pipeline; (b) feature formation experimental pipeline

The components in fig. 3, b are described as follows:

W^k – sliding window for audio signal analysis. Using a sliding window, the current fragment of the audio signal is selected for subsequent analysis;

S – the width of the sliding analysis window;

N – the number of audio signal fragments selected by the sliding window;

ΔW – sliding window step.

Several machine learning models based on fully connected feedforward neural networks and CNNs were implemented, the results of which are presented in table № 5. Fully connected neural network models use MFCCs as input features for processed audio signal fragments [27]. CNNs (4 layers of 2D convolution with BatchNormalization and Max pooling, fully connected output layer) were used to analyze the mel-spectrogram, a type of spectrogram in which the frequency scale is transformed into a nonlinear mel-scale. LSTM included 2 layers, followed by a fully connected output layer (table № 5).

Table № 5

Performance comparison of MLP, LSTM, and CNN models for 1-second audio fragments

Number of layers and number of neurons in a layer	Number of training epochs	Accuracy on training dataset, %	Percentage of correctly detected fake audio recordings, %	Percentage of correctly detected real audio recordings, %	Average detection time for one audio recording, s
MLP (1 layer, 10 neurons)	2	86.79	**60.94**	**71.43**	–
	3	89.69	35.94	66.67	–
LSTM (2 layers, 2 and 1 neurons)	1	82.34	43.75	67.44	
	2	83.21	**46.88**	**69.77**	
CNN (232 by 232 Image size in pixels)	2	98.04	96.88	76.74	1.11
	3	97.25	76.25	97.67	0.93
	4	96.9	83.12	97.67	–

Representing an audio signal as MFCCs more accurately reflects human perception of sound, since the human ear distinguishes low frequencies better than high frequencies, and human speech is focused on low frequencies in the range of 85-340 Hz [28]. Mel-spectrograms are widely used in speech recognition, music processing, detection of anomalies in audio signals and analysis of voice deepfakes.

Analysis of results

Tables № 6 and № 7 present the results of voice audio recognition, depending on the voice type and the internet service for creating the deepfake. In the previously collected dataset, voice types were identified to analyze the model's ability to accurately classify voices across different frequency ranges and acoustic features. The model demonstrates the least confidence in classifying bass-type voices. For a previously prepared demo dataset generated using well-known deepfake creation services and models, the model capabilities were also analyzed. The speechify service is the least confident in detecting deepfakes.

Table № 6

Accuracy of recognition of different types of voices (%)

Bass	Baritone	Tenor	Contralto	Mezzo-soprano	Soprano	Children
72.37	96.14	83.99	98.39	98.79	93.55	95.02

Table № 7

Accuracy of recognition of deepfakes obtained using various internet services (%)

anyvoice-lab	resemble-ai	speechify	typecastai	voicerai	openaifm
66.67	78.26	22.22	37.50	100.00	96.77

The selected fakes were recognized at 97.73% both by average and by quantity, duration from 20 to 30 seconds, average prediction time 0.34 seconds. In the dataset, the proportion of recognized real voices is 100% (averaging over the prediction coefficients for each segment).

Conclusions

A software prototype of a neural network system for detecting voice deepfakes based on CNN 2D was developed. For training and testing, datasets containing a total of 16 hours of real voice audio recordings and 30 hours of deepfake audio were created. Voice deepfakes were also created using various Internet services to test the neural network deepfake detection system.

The experiments conducted demonstrated that the developed system achieved 97% accuracy in recognizing deepfake audio recordings on the test sample while maintaining an acceptable level of false positive classifications. Furthermore, the analysis of mel-spectrograms was conducted at a speed four times greater than that of numerical acoustic feature analysis.

An experimental auditory recognition test of deepfakes was conducted, involving 12 participants. The experiment showed low response consistency and low reliability of human recognition, with the average proportion of deepfakes recognized being 45.8%.

Acknowledgment

The research was supported by the Ministry of Science and Higher Education of the Russian Federation (theme No. FSGF-2023-0004).

References

1. Fox J. Top 40 AI Cybersecurity Statistics // Cobalt URL: cobalt.io/blog/top-40-ai-cybersecurity-statistics#:~:text=AI%20Deepfakes,suite%20executive%20 (Deep%20Instinct) (accessed 29/05/2025).

2. Brett C. 2024 Deepfakes Guide and Statistics // Security.org URL: security.org/resources/deepfake-statistics (accessed 29/05/2025).

3. Smith S. Digital Identity Verification Checks to Pass the 70 Billion Mark in 2024, as Businesses Prioritise Fraud Prevention // Juniper Research URL: juniperresearch.com/press/digital-identity-verification-checks-to-pass/ (accessed 29/05/2025).

4. Moiseeva O.S., Skorobogatov I.A. Molodezh. Obshestvo. Sovremennaja nauka, tehnika i innovacii (Youth. Society. Modern science, technologies & innovations). Krasnoyarsk, 2022, №21, pp. 174-176.

5. Murtazin R.A. et al. Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki. 2021. №4. pp. 545-552.

6. Alimuradov A.K., Churakov P.P. Izmerenie. Monitoring. Upravlenie. Kontrol. 2015. №2. pp. 27-35.

7. Ponomarev K.G., Vereshchagina E.A. Inzhenernyj vestnik Dona. 2024. №6. URL: ivdon.ru/en/magazine/archive/n6y2024/9312.

8. Mcubaa M., Singha A., Ikuesanb R., Venter H. The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation // Procedia Computer Science. 2023. №219. pp. 211-219.

9. Iskhakov A.A., Makhmutova A.Z., Anikin I.V. Inzhenernyj vestnik Dona. 2024. №8. URL: ivdon.ru/ru/magazine/archive/n8y2024/9457.

10. Karaca A. Deep Fake Voice Recognition // Kaggle URL: kaggle.com/code/alperkaraca1/deep-fake-voice-recognition (accessed 29/05/2025).

11. Fitrahidayaat, Reza S. DEEP-VOICE: DeepFake Voice Recognition // Kaggle URL: kaggle.com/code/fitrahidayaat/deep-voice-deepfake-voice-recognition (accessed 29/05/2025).

12. Detecting Altered Voice with AI Deepfake Tools // Resemble.AI URL: resemble.ai/altered-voice-deepfake-detection/ (accessed 29/05/2025).

13. All-In-One Deepfake Detection // Sensity.AI URL: sensity.ai/ (accessed 29/05/2025).

14. Protect your business from AI attacks // MTS.AI URL: mts.ai/ru/deepfake_detection/ (accessed 29/05/2025).

15. Deepfake Detection API // Arya.ai URL: api.arya.ai/deepfake-detection (accessed 29/05/2025).

16. Oblomova O. Timbre and pitch of voice // 4brain URL: 4brain.ru/voice/tone.php (accessed 29/05/2025).

17. Bird J. DEEP-VOICE: DeepFake Voice Recognition // Kaggle URL: kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition (accessed 29/05/2025).

18. RVC-Project / Retrieval-based-Voice-Conversion-WebUI // GitHub URL: github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI (accessed 29/05/2025).

19. Gorshkov G.A., Zamshev V.M., Yashchun T.V. Vestnik vologodskogo gosudarstvennogo universiteta. 2024. №3. pp. 38-42.

20. Cochard D. RVC: An AI-Powered Voice Changer // Medium URL: medium.com/axinc-ai/rvc-an-ai-powered-voice-changer-39927cc83bee (accessed 29/05/2025).

21. Hao H.T. Understanding RVC - Retrieval-based Voice Conversion // GitHub URL: gudgud96.github.io/2024/09/26/annotated-rvc/ (accessed 29/05/2025).

22. AI Voice Generator with Emotion-Driven AI Voice Actors // Typecast URL: typecast.ai/ (accessed 29/05/2025).

23. Feng G.C. Mistakes and how to avoid mistakes in using intercoder reliability indices // Methodology. 2015. №11. URL: doi.org/10.1027/1614-2241/a000086

24. Schmitt N. Uses and abuses of coefficient alpha // Psychological assessment. 1996. Vol. 8, №4. pp. 350-353.

25. Dawid A.P., Skene A.M. Maximum likelihood estimation of observer error-rates using the EM algorithm // Journal of the Royal Statistical Society: Series C (Applied Statistics). 1996. Vol. 28, №1. pp. 20-28.

26. Weller B.E., Bowen N.K., Faubert S.J. Journal of black psychology. 2020. Vol. 46, №4. URL: doi.org/10.1177/0095798420930932.

27. Aksenov O.D. Jelektronnye sistemy i tehnologii: 55-ja jubilejnaja konferencija aspirantov, magistrantov i studentov. Minsk, 2019, pp. 45-46. URL: libeldoc.bsuir.by/handle/123456789/36578.

28. Markina Yu.Yu., Belov Yu.S. Mezhdunarodnyj studencheskij nauchnyj vestnik. 2018. №1. pp. 78.

Дата поступления: 30.07.2025

Дата публикации: 25.09.2025