Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260502016
Volume 5, Issue2 (May-August 2026), PP: 145-150.
www.indjcst.com
ISSN No: 2583-5300

Deepfake Voice Detection Techniques for Cybercrime Prevention and Secure Digital Communication

Gowthaman M[1], Gowri Shankari S[2], Ishwariya N[3], Janamithran K[4], Sowndarya V[5]
[1,2,3,4] B.E Computer Science and Engineering (Cyber Security), United Institute of Technology, Coimbatore, Tamilnadu, India.
[5]Assistant Professor, Department of Computer Science and Engineering (Cyber Security), United Institute of Technology, Coimbatore, Tamilnadu, India.

To Cite this Article: *Gowthaman M[1], Gowri Shankari S[2], Ishwariya N[3], Janamithran K[4], Sowndarya V[5], "Deepfake Voice Detection Techniques for Cybercrime Prevention and Secure Digital Communication", Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 145-150.*

Abstract: *The rapid advancement of AI-based voice synthesis and cloning technologies has introduced significant security challenges in modern digital communication systems. Voice-based cyberattacks such as vishing (voice phishing) and impersonation scams have increased by over 70%, resulting in severe financial losses and emotional distress. As AI-generated synthetic voices become increasingly realistic and indistinguishable from human speech, traditional voice authentication and verification methods are no longer sufficient to ensure security. This paper proposes an intelligent deepfake voice detection system leveraging audio signal processing and machine learning techniques. The system extracts discriminative acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch variations, and spectrogram-based patterns from voice samples. These features are then used to train a classification model capable of distinguishing between genuine human voices and AI-generated deepfake audio. The proposed approach enhances the accuracy and robustness of cybercrime detection by effectively identifying manipulated voice samples. Experimental evaluation indicates that the system achieves up to 97.5% detection accuracy, significantly outperforming traditional methods, thereby strengthening secure authentication mechanisms in digital communication environments. The study contributes to the development of reliable AI-driven cybersecurity frameworks for mitigating voice-based cyber threats.*

Key Words: *Deepfake Voice Detection, Cybersecurity, Voice Biometrics, Machine Learning, Audio Forensics, MFCC, Neural Networks, ZeroGuardian-XDR.*

I. INTRODUCTION

The rapid evolution of artificial intelligence has significantly transformed the landscape of digital communication, enabling advanced applications such as speech recognition, voice assistants, and text-to-speech systems. Among these innovations, deepfake voice synthesis has emerged as one of the most concerning technologies, as it allows the generation of highly realistic human-like speech using deep learning models. While these advancements have improved accessibility and human-computer interaction, they have simultaneously introduced serious security vulnerabilities in cyberspace.

Deepfake voice technology leverages sophisticated architectures such as Generative Adversarial Networks (GANs), autoencoders, and transformer-based speech models to replicate a target speaker's voice with high accuracy. These synthetic voices are increasingly being exploited in cybercrimes such as voice phishing (vishing), identity spoofing, financial fraud, and social engineering attacks. In many reported cases, attackers have successfully impersonated executives, bank officials, and even family members, leading to substantial financial and emotional damage.

Traditional voice authentication systems rely on biometric characteristics such as pitch, tone, and speech patterns. However, with the advancement of AI-generated speech, these conventional systems are no longer sufficient to differentiate between real and synthetic voices. The increasing similarity between human and AI-generated speech has made detection extremely challenging, thereby demanding more robust and intelligent detection mechanisms.

To address this issue, deepfake voice detection has become an active area of research in cybersecurity and audio forensics. Modern approaches utilize machine learning and deep learning techniques to analyze acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), spectrogram patterns, pitch variations, and temporal speech inconsistencies. These features help in identifying subtle artifacts introduced during voice synthesis that are often imperceptible to human listeners.

The objective of this study is to develop an intelligent and reliable deepfake voice detection system—termed ZeroGuardian-XDR—capable of distinguishing between genuine and synthetic speech with high accuracy. By integrating audio signal processing techniques with machine learning-based classification models, the proposed system aims to enhance cybersecurity defenses against voice-based attacks and ensure secure digital communication environments.

Research Contributions

The primary contribution of this study lies in the design and development of an intelligent deepfake voice detection

framework. The key contributions of this work are enumerated as follows:

- Extraction and utilization of robust acoustic features including Mel-Frequency Cepstral Coefficients (MFCC), pitch variations, and spectrogram-based representations that capture both temporal and spectral characteristics of speech signals.
- Development of a machine learning-based classification model trained on labeled datasets containing both human and synthetic speech to effectively learn distinguishing acoustic patterns.
- Integration of a feature fusion approach combining multiple acoustic features to enhance detection performance, reduce false positives, and improve system reliability in real-world scenarios.
- Focus on cybercrime prevention applications, particularly in detecting voice-based phishing, impersonation attacks, and fraudulent authentication attempts in digital communication channels.
- Design of a scalable and adaptable framework extendable to real-time detection systems for banking security, telecommunications, and forensic audio analysis.

II. RELATED WORK

The problem of deepfake voice detection has gained significant attention in recent years due to rapid advancements in speech synthesis and voice cloning technologies. Early research primarily focused on traditional acoustic feature extraction techniques to distinguish between genuine and spoofed speech signals. Mel-Frequency Cepstral Coefficients (MFCC) have been widely used as a foundational feature for speech analysis and spoof detection due to their ability to represent human auditory perception characteristics effectively [1]. Similarly, Constant-Q Cepstral Coefficients (CQCC) were introduced to improve robustness against various types of speech manipulation and have shown strong performance in speaker verification and spoof detection tasks [2].

With the advancement of machine learning techniques, traditional classifiers such as Support Vector Machines (SVM) and Gaussian Mixture Models (GMM) were initially used for classification of real and synthetic speech signals [3]. However, these methods showed limitations in handling complex patterns in large-scale datasets. To overcome these challenges, deep learning-based approaches were introduced. Convolutional Neural Networks (CNN) have been widely adopted for extracting spatial features from spectrogram representations of speech signals, achieving improved accuracy in distinguishing real and fake voices [4]. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks further enhanced detection performance by capturing temporal dependencies in speech sequences [5].

Hybrid architectures combining CNN and LSTM models have demonstrated superior performance by leveraging both spatial and temporal feature learning capabilities, and have been successfully applied in deepfake detection tasks with improved robustness and reduced error rates [6]. Recent advancements include transformer-based architectures and attention mechanisms for speech analysis, capable of capturing long-range dependencies in audio signals and demonstrating better generalization across unseen datasets compared to traditional deep learning models [7].

Benchmark datasets such as ASVspoof 2019 and ASVspoof 2021 have played a crucial role in evaluating deepfake voice detection systems. Studies using these datasets have reported that hybrid deep learning models significantly outperform traditional approaches in terms of Equal Error Rate (EER) and detection accuracy [8]. Feature fusion techniques combining MFCC, pitch variations, and spectrogram-based features have also been proposed to enhance detection robustness [9]. These approaches improve model stability by integrating multiple complementary acoustic representations.

Despite these advancements, existing systems still face challenges such as poor generalization across different datasets, vulnerability to adversarial attacks, and reduced performance in noisy environments. Recent surveys emphasize the need for adaptive and real-time detection frameworks capable of handling continuously evolving deepfake generation methods [10]. The present study addresses these gaps by proposing an integrated, real-time-capable framework that achieves state-of-the-art accuracy using the LCNN model.

III. PROBLEM STATEMENT AND MOTIVATION

The rapid growth of AI-based voice cloning and deepfake audio generation has introduced several critical challenges in cyber security and digital communication systems. The following subsections outline the key problems identified in this domain.

A. Increasing AI Voice Cloning Attacks

AI-driven voice cloning attacks have increased by more than 70% between 2023 and 2025. This sharp rise indicates that cybercriminals are increasingly using synthetic voice technologies to impersonate individuals and carry out fraudulent activities such as financial scams and identity theft. The sophistication of modern AI voice generation tools has lowered the technical barrier for malicious actors, enabling even non-experts to execute convincing impersonation attacks.

B. Low Human Detection Capability

Human beings are only able to distinguish between real and deepfake voices with approximately 73% accuracy. This limited detection ability makes individuals highly vulnerable to voice-based manipulation, as AI-generated speech closely resembles natural human speech patterns in pitch, cadence, and emotional intonation. This vulnerability is further amplified in high-pressure or time-sensitive scenarios where careful scrutiny is not feasible.

C. High Vulnerability in Business Communication

Approximately 41% of businesses report vulnerability to voice impersonation attacks. This highlights a serious security gap in organizational communication systems, where attackers can impersonate executives or employees to authorize fraudulent transactions or extract sensitive data. The consequences range from financial losses and reputational damage to regulatory

compliance violations and legal liabilities.

D. Increasing Financial and Social Impact

Voice phishing (vishing) attacks have increased by approximately 62%, with nearly 1 in 4 adults encountering such scams. Among these victims, approximately 7% have suffered direct financial losses. Global losses attributable to voice-based cybercrime are projected to exceed 40 billion USD by 2027. The emotional and psychological distress caused by impersonation scams further underscores the urgent need for robust automated detection systems.

E. Motivation for the Proposed ZeroGuardian-XDR Framework

The motivation for this work arises from the urgent need to secure modern digital communication systems against increasingly sophisticated AI-generated voice attacks. With the growing use of voice-based authentication in banking, customer service, and online platforms, the risk of voice impersonation has become a critical cybersecurity concern. Traditional security systems are no longer sufficient to detect advanced deepfake audio generated using modern machine learning techniques. Therefore, there is a strong need to develop an intelligent, automated, and reliable detection system that can accurately identify synthetic voices and prevent cybercrime. This study is motivated by the goal of enhancing digital trust, reducing financial fraud, and strengthening cybersecurity frameworks in an era dominated by AI-driven threats.

IV. PROPOSED SYSTEM ARCHITECTURE

The proposed ZeroGuardian-XDR system is designed to detect deepfake voice signals using a structured pipeline that integrates audio preprocessing, feature extraction, and machine learning-based classification. The architecture ensures accurate identification of synthetic speech and enhances cybersecurity in voice-based communication systems. The overall system consists of the following modules:

A. Dataset Collection

The first stage involves collecting a diverse dataset containing both real and synthetic voice samples. Real human speech data is obtained from publicly available datasets such as ASVspoof [2] and LibriSpeech [6]. Deepfake voice samples are collected from AI-generated speech datasets and voice synthesis models including WaveNet [8] and MelGAN [10]. The dataset is carefully curated to include a balanced distribution of genuine and synthetic audio samples, ensuring that the model is effectively trained without class imbalance bias.

B. Voice Preprocessing

The collected raw audio signals are preprocessed to improve quality, consistency, and suitability for feature extraction. Preprocessing steps include noise reduction using spectral subtraction to eliminate background disturbances, silence removal to discard non-speech segments using energy-based voice activity detection, and amplitude normalization to standardize audio levels. These operations improve the robustness and accuracy of the detection model by ensuring only relevant and clean audio data is processed.

C. Feature Extraction

In this stage, important acoustic features are extracted from the preprocessed audio signals. Mel-Frequency Cepstral Coefficients (MFCC) are extracted with 40 coefficients per frame to represent human auditory perception characteristics. Pitch and fundamental frequency (F0) variations are analyzed using autocorrelation methods to capture naturalness of speech. Time-frequency spectrograms are generated using Short-Time Fourier Transform (STFT) to visualize speech patterns and detect inconsistencies introduced by synthetic voice generation. These extracted features provide a strong discriminative foundation for distinguishing real human speech from AI-generated audio.

D. Model Training

The extracted features are used to train machine learning and deep learning models for binary classification. Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, Random Forest with 100 estimators, and Convolutional Neural Networks (CNN) are employed as baseline models. Advanced architectures including ResNet-34 and Light Convolutional Neural Networks (LCNN) are employed as the primary detection models. The training process employs cross-validation with an 80-10-10 train-validation-test split and uses the ASVspoof 2019 dataset as the primary benchmark.

E. Voice Classification and Output

Once the model is trained, it is used to classify incoming voice samples. The system analyzes the extracted features and categorizes input audio into either real voice or deepfake voice based on learned patterns and probability-based outputs. The classification result is presented to the end user through a notification interface, enabling users to take appropriate action. This module plays a crucial role in preventing voice-based cyberattacks in real-time communication systems.

F. Overall System Workflow

The overall workflow begins with the acquisition of an input audio sample. The audio signal is preprocessed using noise reduction, silence removal, and normalization. Next, acoustic features (MFCC, pitch, and spectrogram) are extracted and provided as input to the trained model, which classifies the audio as real or deepfake. The system generates a detection result that can be integrated into secure communication and authentication systems.

V. METHODOLOGY

The proposed methodology for deepfake voice detection is designed based on a real-time communication scenario, where synthetic audio generated by an attacker is transmitted through a communication platform and analyzed at the receiver side. The system integrates deepfake audio generation modeling, signal processing, and deep learning-based detection techniques to accurately identify manipulated voice signals.

A. System Initialization and Deployment

The detection system is initialized with pre-trained model weights derived from training on the ASVspoof 2019 dataset. Upon deployment, the system continuously monitors incoming audio streams from communication interfaces such as voice calls, video conferencing sessions, and voice authentication endpoints. The real-time processing pipeline is optimized to maintain low latency, ensuring minimal disruption to user experience during active communication.

B. Deepfake Audio Generation Modeling

In the source system context, deepfake audio is generated using advanced AI-based speech synthesis techniques including Text-to-Speech (TTS) systems, voice conversion models, and voice cloning techniques that replicate the voice characteristics of a target individual. Voice replay attacks may also be performed by injecting pre-recorded or synthesized audio into communication platforms. Understanding these generation methods is critical for designing robust detection countermeasures.

C. Audio Transmission and Reception

The generated audio is transmitted through real-time communication applications such as video conferencing systems. The attacker delivers the manipulated voice using a microphone interface, which is then transferred across the communication channel. At the target system, the incoming audio signal is received through the speaker interface of the communication platform and captured for analysis.

D. Audio Preprocessing Pipeline

The received signal undergoes multi-stage preprocessing. Noise reduction is performed using spectral subtraction algorithms to remove background disturbances common in real-world environments. Voice Activity Detection (VAD) is applied to identify and remove silence segments. Finally, amplitude normalization ensures consistent signal levels across samples from different recording conditions. These preprocessing steps are crucial for ensuring reliable feature extraction.

E. Feature Extraction and Deep Learning-Based Detection

After preprocessing, MFCC coefficients, pitch contours, and mel-spectrogram representations are extracted from the audio signal. These features are fed into the deep learning-based detection models. In this work, ResNet-based [16] and Light Convolutional Neural Networks (LCNN) models [15] are employed due to their effectiveness in speech classification tasks. LCNN uses max-feature-map activation functions that selectively filter the most discriminative features, while ResNet's residual connections enable training of deeper networks without gradient degradation.

F. Classification, Decision Making, and Notification

Based on the analysis performed by the detection models, the system classifies the input audio into either real or deepfake voice categories. The decision is made using learned feature representations and probability-based softmax outputs from the models, with a classification threshold of 0.5. The classification result is then presented to the end user through a notification interface indicating whether the received voice is authentic or manipulated, enabling users to take appropriate security measures.

VI. IMPLEMENTATION OF ZEROGUARDIAN-XDR

A. Development Environment

The system is developed using Python 3.10 due to its extensive support for data processing and machine learning. Libraries used include Librosa (v0.10) and NumPy for audio signal processing; Scikit-learn for SVM and Random Forest implementations; TensorFlow 2.12 and PyTorch 2.0 for building and training deep learning models. All experiments are conducted on a workstation with an Intel Core i7-12700K processor, 32 GB RAM, and an NVIDIA RTX 3080 GPU.

B. Dataset Preparation

Audio datasets are collected from ASVspoof 2019 (logical access partition) and LibriSpeech. The ASVspoof 2019 dataset contains 2,580 genuine and 22,800 spoofed utterances generated by 19 different TTS and voice conversion systems, providing a comprehensive benchmark for training and evaluation. The dataset is organized into training, development, and evaluation partitions. Class balancing is performed using oversampling of genuine samples.

C. Feature Extraction Implementation

Relevant acoustic features are extracted with a frame length of 25 ms and a hop length of 10 ms. MFCC features (40 coefficients) are extracted using a 512-point FFT with a Hamming window. Pitch estimation is performed using the YIN algorithm. Mel-spectrograms are computed with 128 mel filter banks spanning 0–8000 Hz. Feature normalization using mean-variance normalization is applied to each feature type before model input.

D. Model Architecture and Training

The LCNN model comprises six convolutional layers with max-feature-map activations, two fully connected layers, and a

sigmoid output neuron. The ResNet model adapts the ResNet-34 architecture with modifications for 1D audio feature processing. Both models are trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 64 for 50 epochs. Dropout (rate = 0.3) and L2 regularization are applied to prevent overfitting. Early stopping with patience = 10 is employed based on validation loss.

E. Real-Time Detection Module

The trained system is deployed in a simulated communication environment where incoming audio is processed in real time using a sliding window approach with a 2-second analysis window and 1-second hop. The system achieves an average inference latency of approximately 85 milliseconds per audio segment, making it suitable for integration into real-time voice communication applications. The output classification result is returned to the user interface within this latency window.

VII. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

The performance of the proposed deepfake voice detection system is evaluated using standard metrics: accuracy, precision, recall, and F1-score. The system is tested on the ASVspoof 2019 evaluation partition containing both genuine and AI-generated voice samples. The experimental setup includes training multiple machine learning and deep learning models using extracted features such as MFCC, pitch variations, and spectrogram representations.

A. Performance Metrics

The following standard evaluation metrics are employed: Accuracy (overall proportion of correct predictions); Precision (proportion of detected deepfakes that are truly deepfake); Recall (proportion of actual deepfakes correctly detected); and F1-Score (harmonic mean of precision and recall). Additionally, Equal Error Rate (EER) is computed as a supplementary metric consistent with ASVspoof challenge evaluation protocols.

B. Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	88.5	87.2	86.8	87.0
Random Forest	91.3	90.5	89.7	90.1
CNN	94.6	93.8	94.2	94.0
ResNet	96.8	96.2	95.9	96.0
LCNN	97.5	97.0	96.8	96.9

Table 1: Performance Comparison of Deepfake Voice Detection Models

Source: Experimental evaluation on ASVspoof 2019 dataset.

The results in Table 1 indicate that deep learning models substantially outperform traditional machine learning approaches. The SVM baseline achieves 88.5% accuracy, while Random Forest improves to 91.3%. Among deep learning models, CNN achieves 94.6% accuracy, ResNet achieves 96.8%, and the LCNN model attains the highest accuracy of 97.5% with an F1-score of 96.9%. These results demonstrate the effectiveness of LCNN's selective feature learning through max-feature-map activation for deepfake voice detection.

C. Detection Accuracy Analysis

The system demonstrates high detection accuracy across different audio samples, with deep learning models achieving accuracy above 94%. The substantial improvement over traditional methods is primarily attributable to the ability of CNN-based architectures to automatically learn complex discriminative patterns in speech spectrograms without manual feature engineering. The LCNN architecture's compact design also ensures computational efficiency, achieving the highest accuracy with fewer parameters compared to ResNet.

D. Discussion of Results

The experimental results demonstrate that the proposed ZeroGuardian-XDR system is highly effective in detecting deepfake voice signals. Feature extraction techniques such as MFCC and spectrogram analysis significantly contribute to model performance. Deep learning models, particularly LCNN and ResNet, provide superior accuracy due to their ability to learn complex speech patterns. The LCNN model's use of max-feature-map activations proves particularly effective in suppressing irrelevant acoustic features while amplifying discriminative cues.

However, the system faces potential challenges in highly noisy environments where preprocessing may not fully eliminate background interference. Additionally, performance may degrade when encountering novel deepfake generation techniques not represented in the training data, a known limitation in adversarial machine learning contexts. The relatively small dataset size also limits generalization, motivating future work on data augmentation and transfer learning strategies.

VIII. CONCLUSION AND FUTURE WORK

In this paper, an intelligent deepfake voice detection system, ZeroGuardian-XDR, has been proposed to address the growing threat of AI-generated voice-based cyberattacks. The system integrates audio signal processing techniques with machine learning and deep learning models to effectively distinguish between genuine and synthetic speech. Features such as Mel-Frequency

Cepstral Coefficients (MFCC), pitch variations, and spectrogram representations are utilized to capture the unique characteristics of voice signals. The experimental results demonstrate that deep learning models, particularly ResNet and Light Convolutional Neural Networks (LCNN), achieve high detection accuracy (up to 97.5%) and substantially outperform traditional machine learning approaches.

The proposed system demonstrates strong performance in identifying deepfake audio and provides a reliable solution for enhancing security in digital communication systems. The study highlights the importance of advanced AI-driven detection mechanisms in combating voice-based cyber threats and improving trust in voice authentication systems used across banking, telecommunications, and enterprise environments.

Although the proposed system achieves promising results, several areas for further improvement have been identified. Future work directions include:

- Development of real-time detection systems optimized for live communication environments such as video conferencing platforms and voice-based authentication APIs, with inference latency below 50 milliseconds.
- Integration of advanced transformer-based models and self-attention mechanisms, such as wav2vec 2.0 and HuBERT, to further improve detection accuracy and generalization across unseen deepfake techniques.
- Extension of the system to handle multilingual voice data and diverse accents to improve applicability in global communication systems with diverse speaker populations.
- Improving robustness in noisy environments through advanced audio enhancement techniques and adversarial training to prevent evasion attacks.
- Integration of the proposed detection framework with existing cybersecurity infrastructure, including SIEM systems and authentication middleware, to provide a comprehensive solution for preventing voice-based cybercrime.

Acknowledgment

The authors would like to express their sincere gratitude to their guide, Sowndarya V, Assistant Professor, for her continuous support, valuable guidance, and encouragement throughout the completion of this work. Her insightful suggestions, technical expertise, and constant motivation have played a crucial role in shaping this research and improving its overall quality. The authors also extend their heartfelt thanks to the Department of Computer Science and Engineering (Cyber Security), United Institute of Technology, Coimbatore, for providing the necessary resources, infrastructure, and academic environment required to successfully carry out this study. The authors acknowledge the support of all faculty members and peers who contributed directly or indirectly to the completion of this research work.

REFERENCES

1. M. Todisco, H. Delgado, and N. Evans, "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients," in Proc. IEEE Odyssey, 2017, pp. 283–290.
2. ASVspoof Consortium, "ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," 2019.
3. X. Wang, J. Yamagishi, M. Todisco, H. Delgado, and N. Evans, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," in Proc. IEEE ASRU, 2021, pp. 1–8.
4. T. Kinnunen et al., "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in Proc. Interspeech, 2017, pp. 2–6.
5. D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," arXiv: 1510.08484, 2015.
6. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," in Proc. IEEE ICASSP, 2015, pp. 5206–5210.
7. Y. Jia et al., "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis," in Proc. NeurIPS, 2018, pp. 4485–4495.
8. A. Oord et al., "WaveNet: A Generative Model for Raw Audio," arXiv: 1609.03499, 2016.
9. J. Donahue et al., "Adversarial Audio Synthesis," in Proc. ICLR, 2019.
10. K. Kumar et al., "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in Proc. NeurIPS, 2019.
11. S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech Enhancement Generative Adversarial Network," in Proc. Interspeech, 2017, pp. 3642–3646.
12. A. Lavrentyeva et al., "STC Anti-Spoofing Systems for the ASVspoof 2019 Challenge," in Proc. Interspeech, 2019, pp. 1033–1037.
13. H. Tak et al., "End-to-End Anti-Spoofing with RawNet2," in Proc. IEEE ICASSP, 2021, pp. 6369–6373.
14. Y. Zhang, F. Jiang, and Z. Duan, "One-Class Learning Towards Synthetic Voice Spoofing Detection," IEEE Signal Processing Letters, vol. 28, 2021, pp. 937–941.
15. X. Liu, X. Wu, and H. Meng, "A Light CNN for Deepfake Speech Detection," in Proc. Interspeech, 2020, pp. 971–975.
16. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.
17. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, 1997, pp. 1735–1780.
18. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM KDD, 2016, pp. 785–794.
19. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, 2001, pp. 5–32.
20. C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, 1995, pp. 273–297.
21. A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017, pp. 5998–6008.
22. J. Villalba et al., "State-of-the-Art Speaker Recognition with Neural Network Embeddings," in Proc. IEEE ICASSP, 2020, pp. 7184–7188.
23. Z. Wu et al., "Spoofing and Countermeasures for Speaker Verification: A Survey," Speech Communication, vol. 66, 2015, pp. 130–153.
24. H. Delgado et al., "Further Investigations on Deepfake Speech Detection," in Proc. Interspeech, 2020, pp. 2987–2991.
25. J. Patino et al., "Deep Learning-Based Countermeasures for Anti-Spoofing," IEEE Trans. Inform. Forensics Security, vol. 17, 2022, pp. 280–295.