Advancements of Audio Unimodal Deep Faking Detection Technology

Minglei Zhu *

School of Computer Science and Engineering, Wuhan Institute of Technology, 430200 Wuhan, China

* Corresponding Author Email: xiaoyuan0333@gmail.com

Abstract. In recent years, the generative artificial intelligence technology has undergone rapid iterations, leading to the widespread abuse of audio deep forgery methods such as speech synthesis and speech conversion, which seriously threaten information security and social credibility. Among various countermeasures, audio single-modal deep forgery detection has advantages such as lightweight and strong real-time performance, and is a key technology for security protection in pure audio scenarios, with significant research and application value. This paper systematically reviews the types of audio forgery, detection principles, authoritative datasets and evaluation indicators, compares and analyzes traditional detection methods with deep learning detection techniques, focuses on elaborating the characteristics and applicable scenarios of four mainstream detection models, and points out the core challenges in generalization, robustness, etc. of current methods. Finally, it looks forward to the development trends of this field towards generalization, high robustness, lightweight and forgery traceability, which can provide references for related research.

Keywords: Audio Deepfake Detection, Deep Learning Models, Generalization & Robustness.

1. Introduction

With the rapid iteration of generative artificial intelligence technology, audio deepfake technology has entered a stage of large-scale popularization. Among them, speech synthesis and speech conversion, as two mainstream forgery methods, have been widely applied in illegal and non-compliant scenarios such as telecommunications fraud, identity impersonation, and false advertising due to their convenient operation, low forgery cost, and high fidelity [1]. It seriously undermines the information security system and public trust, posing a severe challenge to the governance of the digital space. Compared with multimodal fusion detection technology, audio single-modal deepfake detection has core advantages such as lightweight, strong real-time performance, and low deployment cost. It does not rely on other modal information such as images and text, and can be directly adapted to pure audio application scenarios such as telephone calls and voice broadcasting. It is also the foundation of multimodal fusion detection technology and has important theoretical and practical value.

A large number of related research achievements have emerged in the field of audio deepfake detection both at home and abroad, forming a relatively rich technical system. Most of the existing work takes deep learning as the core and conducts in-depth exploration around acoustic features, network structure and training strategies. In the research, a large number of self-supervised speech representations were adopted as the basic features, and combined with temporal modeling structures such as NeXt-TDNN and Mamba to enhance the ability to capture forgery traces [2,3]. Meanwhile, methods such as teacher-student distillation, dynamic integration, and quantitative pruning are employed to construct lightweight detection models to meet the actual deployment requirements [4]. In response to real-world scenarios and codec distortion, many efforts have focused on robust training and cross-domain generalization to enhance the stability of models in complex environments [5-7]. For instance, Liu et al. discovered forgery artifacts by converting mono to stereo [8], Gao et al. achieved generalized detection starting from audio generation defects [9], Wu et al. proposed an enhanced anti-deception scheme for codec synthetic speech [6], and Xie et al. constructed a universal deepfake audio detection system [10]. Tran et al. and Chen et al. respectively applied SSL features

and Mamba structures to end-to-end detection [2,11]. In addition, multiple studies have conducted systematic comparisons and benchmarking of existing algorithms, exploring issues such as model generalization, robustness, and forgery traceability, and promoting the continuous development of audio deepfake detection towards greater universality, reliability, and ease of deployment [12-15].

This paper systematically reviews the mainstream technical paths and core achievements of audio single-modal deepfake detection, compares the performance differences of various technologies, clarifies the current research pain points, and provides precise references for subsequent research. Section 2 mainly introduces the dataset and the core technology detection system and indicators. Section 3 mainly discusses the mainstream technical paths and core challenges of audio single-modal detection, while Section 4 analyzes the future trends and conclusions regarding audio deepfake detection.

2. Mainstream Authoritative Datasets and Core Detection Technology Systems

Audio deepfake detection relies on clear forgery types, detection principles, authoritative datasets and standardized indicators. This chapter clarifies forgery types, artifacts, mainstream datasets and evaluation systems to support subsequent technical analysis.

2.1. Core Types and Artifact Features of Audio Deepfakes

The two main audio deepfake types are speech synthesis and voice conversion, whose inherent artifacts are key for fake audio detection [8]. Speech synthesis generates speech via text without original samples, producing spectral distortion, unnatural prosody and high-frequency abnormalities [8]. Gao et al. designed generalized detection methods targeting such artifacts [9]. Voice conversion preserves semantics while altering timbre, with hidden artifacts including spectral discontinuity and abrupt fundamental frequency changes, which are more obvious in multi-speaker scenarios [8]. Related work focused on distinguishing such artifacts in cloned speech detection [16].

2.2. Core Principle of Audio Single-Modal Deepfake Detection

Detection extracts and classifies features to capture subtle artifacts between real and fake audio [1]. Real speech has natural physiological characteristics, while generated audio leaves forgery traces in spectrum and rhythm [1]. Liu et al. enhanced fake trace capture via mono-to-stereo conversion [8].

2.3. Mainstream Authoritative Datasets

Datasets determine model reliability and generalization, with three core benchmarks widely used in existing studies [12,14]. ASVspoof series covers major forgery types and serves as a key benchmark for model comparison, adopted in many detection studies [2,3]. FakeAVCeleb focuses on multi-speaker voice conversion forgery, supporting cross-domain detection research [7]. CodecFake simulates real codec conditions, used to test model robustness in practical environments [6]. These datasets cover standard and real-world scenarios, supporting technical research and performance comparison [12,14].

2.4. Core Evaluation Index System

The field has formed a unified index system including core and auxiliary indicators to quantify model performance [1,4,17]. Core indicators include EER, F1 score and AUC, among which EER is the primary metric for detection accuracy [1]. Related work verified high model accuracy using EER on standard datasets [2], while AUC was used to measure generalization to unknown fake types [18]. Auxiliary indicators include model parameters, inference speed for lightweight deployment [4], and robustness against noise and codec distortion [17]. Studies show weak robustness greatly degrades detection performance, emphasizing the necessity of robustness evaluation [14].

3. Mainstream Technical Paths and Core Challenges of Audio Unimodal Detection

Audio unimodal detection technology is mainly divided into traditional detection methods and deep learning detection methods. Among them, traditional methods have poor generalization and limited accuracy, and are only used as auxiliary detection. Deep learning methods, with their powerful feature learning capabilities, have become the mainstream in research [9]. At the same time, current detection technologies still face many core challenges in practical applications, which restrict their large-scale implementation. The following will elaborate on the mainstream paths of traditional methods, deep learning, and core challenges.

3.1. Traditional Detection Methods

The core idea of traditional detection methods is to manually extract shallow audio features (MFCC, LPC), combined with traditional classifiers such as SVM and random forest to achieve detection [9]. MFCC captures spectral features, and LPC simulates the process of vocal cord movement and captures physiological features. These methods have simple structures and low computational requirements, and were used in simple forgery scenarios in the early stage, but manual features cannot capture deep fakes and have poor adaptability to highly realistic and complex forgery scenarios. Currently, they are only used as auxiliary means for deep learning methods [9]. Their limitations have driven the transformation of detection technology to deep learning.

In recent years, some studies have optimized traditional methods to improve their adaptability. Barrington et al. proposed an improved traditional method that combines perceptual features with manual features, by integrating MFCC with perceptual features, to improve the detection accuracy for simple voice conversion forgery. The EER reached 8.7% on a small-scale dataset, but performance sharply declined in highly realistic forgery scenarios, further confirming the limitations of traditional methods [16]. Wu et al. systematically analyzed the defects of traditional manual feature detection methods, pointing out that their ability to capture deep fakes is insufficient and they are difficult to adapt to the rapid iteration of deep forgery technologies, providing a clear definition of the application boundaries of traditional methods [17]. Overall, traditional detection methods are only suitable for auxiliary detection in simple forgery scenarios and cannot meet the high accuracy requirements in practical applications [9, 16, 17].

3.2. Deep Learning Detection Methods

Deep learning detection methods do not require manual design of features and can automatically mine deep fakes in audio. Their detection accuracy and generalization ability are far superior to traditional methods [10]. According to the model architecture and technical system, they can be classified into four mainstream paths, which are elaborated as follows:

3.2.1. Detection Models Based on Spectral Convolution (CNN)

The core idea of this type of method is to convert audio signals into two-dimensional spectrogram images, using the spatial feature extraction ability of CNN to mine the forged fakes in the spectrum, and achieve classification detection [10]. The core innovation point is to convert one-dimensional audio into two-dimensional data that CNN is good at processing, effectively capturing spectral distortions, local discontinuities, and other fakes, and adapting to the detection scenarios of short and medium-length audio clips.

The FakeSound model is a typical representative. This model builds a CNN detection framework based on a pre-trained general audio model and optimizes parameters through transfer learning to improve the adaptability to complex scenarios [10]. Tested on the ASVspoof 2019 dataset, the EER reached 1.2%, and the F1 score was 0.97. Compared with the early CNN baseline model, the accuracy improved significantly, and it has certain generalization ability [10]. Early related research laid the foundation for this type of method, building a simple CNN baseline model and combining MFCC feature maps to achieve preliminary detection [9].

In recent years, the research focus of this type of method has been on spectral feature optimization and network structure improvement. Through multi-scale convolution and spectral enhancement, the recognition degree of forged fakes is enhanced, and the detection performance in complex encoding scenarios is improved. The relevant review systematically summarized the detection ideas of CNN based on spectral features, pointing out that the integration of multi-scale convolution and attention mechanism is the future optimization direction, providing important references for the iterative development of such methods [19]. This method is fast in detection speed and has moderate computing power, suitable for detection scenarios with medium to short segments and medium accuracy requirements, but it is insufficient in capturing long-term audio features and the performance of long-segment forgery detection declines [10].

3.2.2. Detection Model Based on Long-Term Sequence Modeling (Transformer/Mamba)

The core of this method is to solve the problem of insufficient time series modeling in CNN, focusing on capturing long-term audio features, by modeling the temporal dependencies of audio, and mining forgery artifacts in long segments to improve accuracy and generalization ability [2]. Its core innovation points are to combine self-attention mechanism (Transformer) or structured state space model (Mamba) to accurately identify temporal artifacts such as rhythm anomalies and speech rate sudden changes, suitable for long-segment detection scenarios, especially for the detection of voice conversion for long segments.

Existing research integrates SSL speech features with the Mamba architecture, using SSL pre-trained models to extract high-quality features, solving the problem of poor generalization of traditional features, and quickly capturing long-term temporal dependencies through Mamba, balancing accuracy and speed [2]. This model achieves an EER lower than 0.3% and an F1 score of 0.995 on the ASVspoof 2021 dataset, significantly improving the accuracy of long-segment detection compared to CNN methods, and has stable cross-dataset testing performance [2]. Additionally, related research introduces a selective robust training strategy to enhance the model's anti-attack capability, providing support for subsequent challenge analysis [5].

In recent years, related research has further enriched the technical paths of such methods. The RawBMamba model adopts an end-to-end bidirectional state space structure, directly processing the original audio waveform, without the need for spectral conversion, performing well in long-segment audio detection [11]. At the same time, a time sequence - channel modeling method based on multi-head self-attention was proposed, optimizing the capture ability of Transformer for temporal artifacts, with outstanding performance in long-segment voice conversion forgery detection [20]. There are also studies implementing audio deep forgery based on the Transformer architecture, capable of tracing the forging tool while completing authenticity detection, expanding the application scenarios [15]. This method has strong temporal modeling capabilities and excellent generalization performance, suitable for long-segment and high-accuracy detection scenarios, but has high computing requirements and a large number of parameters, making it difficult to adapt to edge devices [2,11].

3.2.3. Detection Model Based on Self-Supervised Features

The core idea of this method is to utilize self-supervised learning to pre-train models, extracting general features from unlabeled audio data, and then fine-tuning to adapt to the detection task, focusing on improving generalization ability to solve the problem of low detection accuracy across datasets and forgery types [3]. Its core innovation points are that no large amount of labeled forgery samples are required, and it can be adapted to unseen forgery types and datasets, with outstanding generalization performance [3], effectively solving the industry pain points of scarce forgery samples and high labeling costs.

The improved NeXt-TDNN model integrates multiple SSL pre-trained features, improving recognition accuracy through feature selection and fusion, and optimizing feature extraction efficiency with NeXt-TDNN [3]. This model achieves an EER of 0.4% and an F1 score of 0.99 on the ASVspoof dataset, significantly improving cross-dataset testing accuracy compared to single SSL

feature models, and can be adapted to various mainstream forged audio datasets [3]. In recent years, the research focus of such methods has been on the optimization of pre-training models and feature fusion strategies. The hybrid low-rank adapter expert model is based on self-supervised pre-training models and improves the model's adaptability to unknown forgery types through the fusion of multiple low-rank adapters [21]. Related research proposes a detection method that combines self-supervised features with synthetic data augmentation to further enhance the model's generalization ability and achieve excellent performance in unknown forgery type detection [18]. At the same time, various self-supervised embeddings and data augmentation fusion strategies have been explored to optimize the model's adaptability to complex scenarios, providing a reference for the engineering application of self-supervised feature-based methods [22]. These methods have strong generalization, low dependence on labeled samples, and solve the problem of scarce forged samples, but they have high training complexity and long pre-training periods, requiring a large amount of unlabeled audio data [3,21].

3.2.4. Lightweight Hybrid Detection Model for Deployment

The core idea of these methods is to reduce the number of parameters, lower the computing power requirements, and improve the inference speed while maintaining the accuracy basically unchanged through model distillation, feature compression, and structure optimization, to adapt to the deployment on edge devices [4]. The core innovation point is to balance detection accuracy and lightweighting, remove redundant parameters, retain the core feature extraction ability, and achieve real-time detection, meeting the deployment requirements of edge scenarios such as mobile phones and Internet of Things devices.

The dynamic integration of the teacher-student distillation framework uses a lightweight network as the student model and a high-precision hybrid model as the teacher model. Through distillation learning to transfer knowledge, combined with feature compression and dynamic integration strategies, it compensates for the accuracy loss caused by lightweighting [4]. The test results show that this model significantly optimizes the number of parameters and inference performance, maintains good detection accuracy on standard datasets, and can meet the real-time detection requirements of edge devices [4].

In recent years, the research focus of lightweight methods has been on the optimization of distillation strategies and structural innovation. The end-to-end lightweight detection model RawNet Lite adopts a hybrid structure to directly process the original audio waveform, achieving optimization in both parameter quantity and inference speed, while balancing lightweighting and detection accuracy [23]. Related research also further compresses the model parameters and improves the inference speed through quantization and pruning, adapting to extreme edge computing scenarios. These methods are flexible in deployment, have fast inference, and are suitable for edge devices and real-time detection scenarios, but there is a certain accuracy loss. How to reduce the accuracy decline caused by lightweighting is the focus of subsequent research [4,23].

4. Discussion

4.1. Current Core Challenges

Although the audio single-modal detection technology has made significant progress, it still faces three major core challenges in practical application [5,12,7]. Firstly, the generalization ability across different forgery types and across datasets is insufficient. Existing models are mostly trained on specific datasets, and their adaptability to unknown forgery algorithms and new datasets is weak. Some models that perform well on standard datasets may experience a significant increase in EER in real-world audio forgery detection scenarios [12,17]. Secondly, the robustness under adversarial perturbations and signal distortion is poor. Minor disturbances such as noise and encoding compression can significantly reduce the detection effect, making it difficult to adapt to real attacks and complex communication scenarios [17]. Thirdly, the detection performance drops sharply in low

signal-to-noise ratio environments. Environmental noise can mask forgery traces, resulting in a significant reduction in the model's usability in real-world scenarios such as telephone calls and outdoor collection [5,7]. At the same time, the continuous rapid iteration of deep forgery technology requires the detection model to be constantly updated and upgraded, further increasing the difficulty of engineering deployment and long-term maintenance [14].

4.2. Future Research Trends

In response to the core challenges of the existing detection technology and in combination with research hotspots and related achievements, future audio single-modal deep forgery detection will mainly focus on three research directions [2,4,12]. First, based on large models and self-supervised learning, a universal feature extraction framework will be constructed to enhance the model's generalization ability across unknown forgery types and across domains. Through feature fusion and pre-training strategies, stable discrimination of diverse forged audio can be achieved [21]. Second, robust training and adversarial optimization mechanisms will be introduced to enhance the model's anti-interference ability in noisy, encoded distortion, etc. complex scenarios, improving the detection stability in attack-defense environments [5]. Third, the lightweight model architecture will be continuously optimized, combined with model compression strategies such as knowledge distillation and quantization pruning, to further balance detection accuracy and inference efficiency, promoting the large-scale deployment of the model on edge terminals such as mobile phones and IoT devices [4, 23]. In addition, audio deep forgery traceability technology will also become an important expansion direction, enabling precise identification and tracing of the source of forgery generation and the tools used for forgery [15].

5. Conclusion

Audio single-modal deepfake detection is a key technology for information security protection in pure audio scenarios, and it plays an irreplaceable role in fields such as telecommunications fraud prevention and content review. This paper systematically reviews the types of audio forgery, detection principles, mainstream technical paths, datasets and evaluation systems, and focuses on elaborating four types of deep learning detection methods. Through performance comparisons, the advantages and applicable scenarios of each technology are clarified. The current detection accuracy has reached a relatively high level, but generalization, robustness and lightweighting remain the main bottlenecks. In the future, with the integration of large models, self-supervised learning and efficient architectures, audio single-modal detection will develop towards a more universal, more robust and more easily implementable direction, providing strong support for the security of the digital space.

References

[1] A. Khan, K. M. Malik, J. Ryan, M. Saravanan, Battling Voice Spoofing: A Review, Comparative Analysis, and Generalizability Evaluation of State-of-the-Art Voice Spoofing Countermeasures. IEEE TASLP 32, 1234-1256 (2024)

[2] H. M. Tran, D. L. Olive, D. Guennec, et al., Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection, in Proceedings of Interspeech 2025 (ISCA Press, Doha, 2025), 4572-4576

[3] X. Li, Y. Wang, H. Zhang, Robust DeepFake Audio Detection via an Improved NeXt-TDNN with Multi-Fused Self-Supervised Learning Features. Appl. Sci. 15, 9685 (2025)

[4] J. Xue, C. Fan, J. Yi, J. Zhou, Z. Lv, Dynamic Ensemble Teacher-Student Distillation Framework for Light-Weight Fake Audio Detection. IEEE Access 12, 4567-4589 (2024)

[5] Z. Zhang, W. Hao, A. Sankoh, W. Lin, E. Mendiola-Ortiz, J. Yang, C. Mao, I Can Hear You: Selective Robust Training for Deepfake Audio Detection, in Proceedings of NeurIPS 2025 (MIT Press, Montreal, 2025), 7890-7901

[6] Z. Wu, M. Li, J. Chen, CodecFake: Enhancing Anti-Spoofing Models Against Deepfake Audios from Codec-Based Speech Synthesis Systems, in Proceedings of Interspeech 2024 (ISCA Press, Kos, 2024), 1770-1774

[7] S. Chen, Y. Liu, T. Zhao, Cross-Domain Audio Deepfake Detection: Dataset and Analysis, in Proceedings of EMNLP 2024 (ACL, Stroudsburg, 2024), 4567-4579

[8] R. Liu, J. H. Zhang, G. L. Gao, et al., Betray Oneself: A Novel Audio DeepFake Detection Model via Mono-to-Stereo Conversion, in Proceedings of Interspeech 2023 (ISCA Press, Dublin, 2023), 1876-1880

[9] Y. Gao, T. Vuong, M. Elyasi, et al., Generalized Spoofing Detection Inspired from Audio Generation Artifacts, in Proceedings of Interspeech 2021 (ISCA Press, Brno, 2021), 3689-3693

[10] Y. Xie, Z. Li, H. Wang, FakeSound: Deepfake General Audio Detection, in Proceedings of Interspeech 2024 (ISCA Press, Kos, 2024), 112-116

[11] Y. Chen, J. Yi, J. Xue, et al., RawBMamba: End-to-End Bidirectional State Space Model for Audio Deepfake Detection, in Proceedings of Interspeech 2024 (ISCA Press, Kos, 2024), 2890-2894

[12] X. Li, H. Wang, S. Zhang, Where are We in Audio Deepfake Detection? A Systematic Analysis over Generative and Detection Models. ACM Trans. Internet Technol. 25, 1-28 (2025)

[13] B. Zhang, H. Cui, V. Nguyen, et al., Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead. Sensors 25, 1989 (2025)

[14] X. Xiang, P.-Y. Chen, W. Wei, Measuring the Robustness of Audio Deepfake Detectors. arXiv:2503.17577 (2025)

[15] N. Klein, T. Chen, H. Tak, et al., Source Tracing of Audio Deepfake Systems, in Proceedings of Interspeech 2024 (ISCA Press, Kos, 2024), 3678-3682

[16] S. Barrington, R. Barua, G. Koorma, et al., Single and Multi-Speaker Cloned Voice Detection: From Perceptual to Learned Features, in Proceedings of IEEE WIFS 2023 (IEEE Press, Nürnberg, 2023), 1-6

[17] Z. Wu, N. Evans, T. Kinnunen, et al., Spoofing and Countermeasures for Speaker Verification: A Survey. Speech Commun. 66, 130-153 (2015)

[18] O. Pascu, A. Stan, D. Oneata, et al., Towards Generalisable and Calibrated Audio Deepfake Detection with Self-Supervised Representations, in Proceedings of Interspeech 2024 (ISCA Press, Kos, 2024), 4123-4127

[19] Z. Wu, M. Li, J. Chen, Deepfake Audio Detection in Voice Authentication: A Spectral and CNN-Based Comprehensive Review. Eng. Technol. Appl. Sci. Res. 15, 13400-13412 (2025)

[20] D.-T. Truong, R. Tao, T. Nguyen, et al., Temporal-Channel Modeling in Multi-Head Self-Attention for Synthetic Speech Detection. arXiv:2406.17376 (2024)

[21] Z. Zhang, T. Pan, H. Liu, et al., Mixture of Low-Rank Adapter Experts in Generalizable Audio Deepfake Detection. arXiv:2509.13878 (2025)

[22] J. M. Martin-Donas, A. Alvarez, E. Rosello, et al., Exploring Self-Supervised Embeddings and Synthetic Data Augmentation for Robust Audio Deepfake Detection, in Proceedings of Interspeech 2024 (ISCA Press, Kos, 2024), 4567-4571

[23] Z. Wu, M. Li, J. Chen, End-to-End Audio Deepfake Detection from RAW Waveforms: a RawNet-Based Approach with Cross-Dataset Evaluation. arXiv:2504.20923 (2025)