Defence Against the Deepfake Arts : Improving Audio Deepfake Detection With Context Awareness

Anonymous submission to Interspeech 2025

Abstract

Abhay : Exceeding 408 characters.

The increasing use of generative AI models to create realistic deepfakes poses significant challenges to information security, particularly in the realm of audio manipulation. This paper addresses the pressing need for improved detection of audio deepfakes by proposing a novel approach that incorporates context awareness. We begin by discussing the various types of audio deepfakes and their potential consequences, emphasising their ethical considerations and threats to privacy, security, and democratic processes. We review existing deepfake detection techniques and currently identified challenges in the field. We then present our method, which leverages textual context extracted from transcriptions as well as audio features extracted from deepfake detection models. We propose two fusion strategies, late fusion and mid fusion, to integrate these features and enhance detection accuracy. We conduct experiments using benchmark datasets and state-of-the-art deepfake detection models to evaluate the performance of our proposed approach. Our results demonstrate promising improvements in detecting audio deepfakes, highlighting the potential of context awareness in enhancing detection capabilities. Finally, we discuss the implications of our findings, including the privacy-security trade-off and limitations of our approach, and suggest directions for future research to address these challenges.

Index Terms: audio deepfake detection

1. Introduction

With the rapid advances of the digital landscape and new media technologies, users become increasingly exposed to different types of manipulations. Nowadays, everyone can create super realistic AI-generated content such as images, audio, and even videos. The challenge of distinguishing between those and the real ones becomes nearly impossible for the ordinary user. Audio deepfakes or voice clones are like digital ventriloquists, using various deep learning techniques to create voices that convincingly impersonate real people [1]. We can distinguish between a couple of different types of deepfakes. [2]. The first approach is imitation-based (voice conversion), wherein the original audio signal is modified to mimic another targeted voice. The manipulations can be of the speech style, tone, and articulation. This technique is mainly used in show business and, therefore, is applied by a third person or deep learning software.

The second approach is synthetic-based, which produces artificially generated speech by using text-to-speech techniques (TTS). This process involves three stages: text analysis, acoustic, and vocoder. The text analysis step gathers related audio files and the corresponding transcripts. Next, the acoustic module extracts relevant features from the collected audio files to train the new model. Techniques like Tacotron 2, Deep Voice 3 [3], and Flat Speech 2 [4] are being used for this module. Finally, the vocoder generates the final deepfake audio output by creating a speech waveform based on the information extracted from the steps before.

The third technique is centred on replaying existing recordings of the target speaker. Two techniques are notable here: cut-and-paste detection, where a victim's recorded segment is played through a phone handset for capture, and far-field detection, which involves crafting sentences for manipulation of a text-dependent system.

There are other types of deepfakes, such as emotion fake. This technique alters the emotions of a speaker's voice while preserving the original words and speaker identity. For example, a happy statement can be transformed into a sorrowful one, potentially twisting the meaning and intent of the message [5]. Another type is scene fake, which involves modifying the background sounds of a recording, leaving the speaker and content untouched. For instance, an office, bus or park conversation can be made to sound like it took place at an airport. Altering the acoustic scene can raise concerns about the recording's authenticity and even influence its interpretation [6]. The last type of deepfake is the partially fake, which involves replacing specific words within the original audio signal with either real or synthetic audio segments. The speaker remains the same throughout, but the content is altered [7].

All these different techniques can have far-reaching consequences, affecting domains ranging from political discourse to personal security. These deepfakes are not only used in creating political propaganda to manipulate public opinion [8] but have also become tools for conducting phone scams [9], posing significant threats to individual privacy and security. The ability to convincingly replicate or alter voices using artificial intelligence has underscored the urgent need for robust detection mechanisms. This paper discusses existing deepfake detection techniques and proposes a new pipeline leveraging these techniques.

This paper contributes to proactive measures in detecting audio deepfakes, particularly focusing on detecting audio deepfakes of public figures. Public figures are especially susceptible to the detrimental effects of audio deepfakes as their extensive online presence renders them prime targets for malicious actors seeking to disseminate misinformation.

2. Related Work

This section reviews the different methods of audio deepfake detection, focusing on several competitions that benchmark progress and pivotal studies that mark the development in this

field. We propose a detailed overview of various approaches spanning from signal processing to advanced deep learning techniques and consider the challenges of generalising detection across diverse deepfake datasets.

2.1. Feature Extraction

Feature extraction is an essential phase of the pipeline of every detection method. Features are categorised into short-term spectral, long-term spectral, prosodic, and deep features, each with unique characteristics and extraction methods.

Short-term spectral features, derived through digital signal processing techniques like the Short-Time Fourier Transform (STFT), focus on the immediate magnitude and phase spectrum being computed at each time step [10] but fall short in capturing temporal dynamics. Long-term spectral features, on the other hand, aim to grasp extended information across speech signals, using different transform methods such as STFT, Constant-Q Transform, Hilbert Transform, and Wavelet Transform to enhance detection accuracy by covering a broader temporal scope.

Prosodic features include elements like fundamental frequency, duration, and energy distribution, which leverage the rhythmic and intonation patterns of the speech, offering a complementary detection dimension by exploiting the nuances of natural language that are often poorly replicated in synthetic speech. In contrast to the short-term features, prosodic ones are extracted from longer speeches such as phone calls and audio recordings. [11].

Deep features, extracted through neural network models, represent a significant advancement in feature extraction by learning directly from data, thus potentially overcoming the biases inherent in hand-crafted features. These deep features can be divided into partially and fully learnable spectral features, supervised embedding features, which can be one of the four different types: spoof, emotion, speaker, and pronunciation embeddings and self-supervised embedding features, which do not require the costly annotated audio data [12]. Some pre-trained self-supervised speech models are publicly available, such as Wav2vec, LS-R and HuBERT.

2.2. Competitions

There have been a series of competitions over the years that have played a crucial role in advancing state-of-the-art deepfake detection methods. The most well-known challenges are ASVspoof and Audio Deepfake Detection (ADD), which have systematically tried to leverage the detection of spoofed and deepfake audio. These competitions try not only to benchmark the advancement in this field of study but also to emphasise the importance of robust detection methods.

2.2.1. ASVspoof

The ASVspoof 2021, 2022 and 2023 [13, 14, 15] challenge focuses on improving the security of Automatic Speaker Verification (ASV) systems against spoofing attacks, which aim to imitate a legitimate user's voice. The challenge comments on the different ways to generate attacks and the difficulties of finding a robust detection method that can tackle all of them at once. To do so, the competition is divided into three different sub-tasks. The first and the most important for our paper is the deepfake task (DF), which focuses on detecting manipulated or synthetic speech, often used to impersonate a target speaker and potentially harm their reputation. This task consists of genuine and manipulated speech utterances processed using different lossy codecs commonly used in media storage. These codecs introduce distortions during the encoding and decoding process, which vary depending on the codec and its configuration.

This part aims to assess the robustness of spoofing detection solutions when used to detect compressed manipulated speech data of varying characteristics posted online. The second subchallenge is Logical Access (LA), which differentiates between actual human speech and speech generated using artificial intelligence, like text-to-speech (TTS) and voice conversion (VC) technologies.

The final task is Physical Access (PA), focusing on detecting replay attacks, where a pre-recorded voice sample is used to spoof the system. It simulates real-world scenarios with compressed and potentially distorted audio. Overall, high error rates were observed, though several systems outperformed baseline models, indicating progress yet significant room for improvement.

2.2.2. Audio Deepfake Detection (ADD)

Following the problems of the previous ADD 2022 challenge [16], particularly the binary classification approach and limited evaluation rounds for the Fake Game Track, ADD 2023 [17] introduces a new set of sub-challenges and tasks to improve audio deepfake detection capabilities. The 2023 challenge consists of three main tracks: Audio Fake Game (FG), Manipulation Region Location (RL) and Deepfake Algorithm Recognition (AR).

The FG task aims to improve track features in two evaluation rounds for generating and detecting fake audio. The generation task (FG-G) aims to create fake audio that bypasses the detection model in the detection task (FG-D), which attempts to identify these generated utterances. The RL sub-challenge focuses on pinpointing the specific regions within partially fake audio that have been manipulated with either real or generated audio. Moreover, the AR task aims to identify the specific algorithms used to generate the deepfake audio.

The evaluation dataset even includes samples generated by unknown algorithms, adding further complexity. While the results show progress, challenges remain, particularly in accurately locating manipulated regions and recognising unknown deepfake algorithms. This indicates the need for further research and development in these areas [18].

2.3. Methodological Evolution

Studies have explored various dimensions of audio deepfake detection, ranging from signal processing to multimodal and semantic approaches. Key findings and methodologies from recent literature include:

2.3.1. Traditional Machine Learning Classification

Traditional methods including a wide range of machine learning classifiers such as logistic regression [19], probabilistic linear discriminant analysis, random forest [20], support vector machine (SVM) based models, Gaussian Mixture Models (GMM)[13] and k-nearest neighbour (KNN) which have been widely utilised for their robustness and efficacy in discriminating between genuine and spoofed speech. The GMM and SVM-based classifiers have been used in various challenges over the years and have produced promising results. However, the problem with the traditional classifiers is that they cannot be used as one robust detection method against all audio deepfakes, given that the nature of the attack is usually unknown.

2.3.2. Deep Learning Classification

This subsection discusses audio feature extraction methods to improve detection performance against sophisticated deepfake algorithms. Time-Delay Neural Networks (TDNN) [21] and Convolutional Neural Networks (CNN) [22] trained on mel-spectrograms have been pivotal in learning discriminative features from audio data. These methods introduce an approach to improve detection in multimodal (audio-visual) deepfakes by training a detector on a concatenation of features learnt from video and audio channels. The deep learning models outperform traditional classifiers by effectively modelling the important patterns in audio data, with specific architectures like Light CNNs and ResNets being highlighted for their success in competitions and their innovative adaptations to address the challenges of deepfake detection.

There have also been studies on the use of Explainable AI (XAI) techniques to interpret deepfake detection models, which offer insights into decision-making processes, potentially guiding the development of more robust detection mechanisms. [22] uses simple CNN backbones trained on mel-spectrograms and LSTM in the detection to maintain as much simplicity as possible. Deep Taylor Decomposition (DTD) and Integrated Gradients and Layer-wise Relevance Propagation (LRP) were used to explain the neural networks by highlighting the most prominent features of the audio signals. These methods found that dependency varies on the frequency bands and characteristics of the audio that cannot be recognised by a human.

2.3.3. Semantic Approach

Semantic approaches targeting the emotional properties of speech have shown promise, exploiting deepfake algorithms' inability to model complex emotional expressions accurately. [23] introduces one such method for identifying deepfake audio by focusing on the emotional nuances that current deepfake generators struggle to replicate accurately. The methodology leverages a Semantic Approach to detect emotion features through a Speech Emotion Recognition (SER) system, which is more adept at recognising complex emotional cues in speech that are not typically well-simulated by deepfake technologies.

The process begins by analysing the audio signal to estimate the expressed emotion and extract relevant features, which are then given to a Synthetic Speech Detector (SSD) system to classify the audio as fake or real. The SER employs a neural network, including 3D convolutional layers and an attention mechanism, to categorise emotions and enhance feature extraction, while the SSD utilises a Random Forest classifier for final determination of the final set of possible emotions such as happy, sad and angry. The approach demonstrated high accuracy, around 0.8, for all deepfake generation algorithms over various datasets (ASVspoof2019 and Cloud2019). The paper reports significant success in detecting deepfakes, particularly when the model is finetuned with a diverse set of emotions and trained on adversarial datasets to improve robustness against various deepfake generation algorithms.

More recently, research has started focusing on multimodal and modality-agnostic methods for deepfake detection [24, 21, 25], but these works are primarily aimed at audio-visual deepfake detection. These methods, along with previous studies in improving speech recognition through multi-task training [26], consistently show, however, that combining cues from different modalities can achieve more comprehensive and reliable deepfake detection/recognition by learning robust cross-modal features. We demonstrate the same for audio deepfake detection by combining acoustic cues from the raw audio and linguistic cues from their transcriptions.

3. Method

4. Experiments and Results

5. Conclusions / Discussion

6. Acknowledgements

Acknowledgement should only be included in the camera-ready version, not in the version submitted for review. The 5th page is reserved exclusively for acknowledgements and references. No other content must appear on the 5th page. Appendices, if any, must be within the first 4 pages. The acknowledgments and references may start on an earlier page, if there is space.

The authors would like to thank ISCA and the organising committees of past Interspeech conferences for their help and for kindly providing the previous version of this template.

7. References

[1] P. Kawa, M. Plata, and P. Syga, "Defense against adversarial attacks on audio deepfake detection."

[2] A. Dixit, N. Kaur, and S. Kingra, "Review of audio deepfake detection techniques: Issues and prospects," *Expert Systems*, vol. 40, 04 2023.

[3] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," 2018.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2022.

[5] Y. Zhao, J. Yi, J. Tao, C. Wang, X. Zhang, and Y. Dong, "Emofake: An initial dataset for emotion fake audio detection," 2023.

[6] J. Yi, C. Wang, J. Tao, Z. Tian, C. Fan, H. Ma, and R. Fu, "Scenefake: An initial dataset and benchmarks for scene fake audio detection," 2022.

[7] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-truth: A partially fake audio detection dataset," 2023.

[8] G. STĂNESCU, "Informational war: Analyzing false news in the israel conflict."

[9] Y. Mirsky, "Df-captcha: A deepfake captcha for preventing fake calls," 2022.

[10] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2119–2123.

[11] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[12] H. Yu, Z.-H. Tan, Y. Zhang, Z. Ma, and J. Guo, "Dnn filter bank cepstral coefficients for spoofing detection," *Ieee Access*, vol. 5, pp. 4779–4787, 2017.

[13] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[14] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-G. Kang, H.-J. Yu, N. Evans, and T. Kinnunen, "Sasv challenge 2022: A spoofing aware speaker verification challenge evaluation plan," *arXiv preprint arXiv:2201.10283*, 2022.

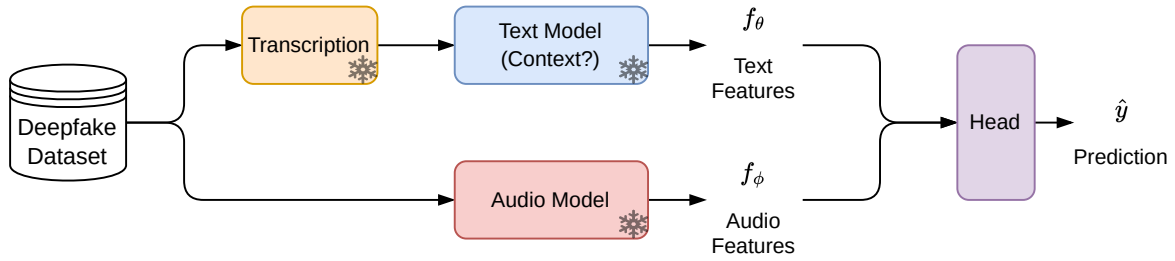

Figure 1: *Pipeline for Mid-Fusion*

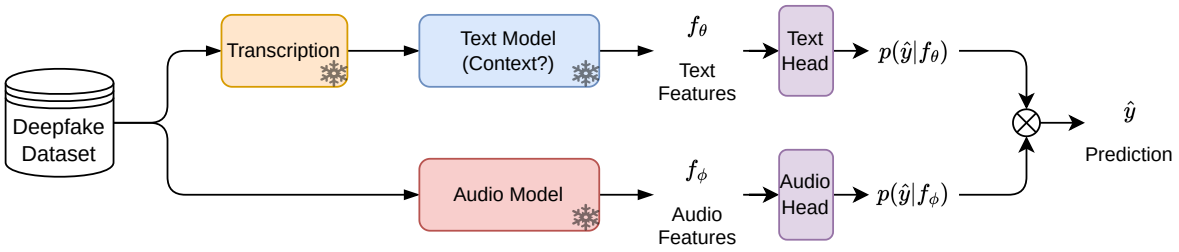

Figure 2: *Pipeline for Late-Fusion*

[15] W. Ge, H. Tak, M. Todisco, and N. Evans, "Can spoofing countermeasure and speaker verification systems be jointly optimised?" in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[17] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, S. Nie, and H. Li, "Add 2023: the second audio deepfake detection challenge," 2023.

[18] X.-M. Zeng, J.-T. Zhang, K. Li, Z.-L. Liu, W.-L. Xie, and Y. Song, "Deepfake algorithm recognition system with augmented data for add 2023 challenge," in *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*, 2023.

[19] Y. Rodríguez-Ortega, D. M. Ballesteros, and D. Renza, "A machine learning model to detect fake voice," in *Applied Informatics*, H. Florez and S. Misra, Eds. Cham: Springer International Publishing, 2020, pp. 3–13.

[20] Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in asvspoof2017." in *Interspeech*, 2017, pp. 87–91.

[21] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, and S. Tubaro, "A robust approach to multimodal deepfake detection," *Journal of Imaging 2023, Vol. 9, Page 122*, vol. 9, p. 122, 6 2023. [Online]. Available: https://www.mdpi.com/2313-433X/9/6/122/htmhttps://www.mdpi.com/2313-433X/9/6/122

[22] S. Y. Lim, D. K. Chae, and S. C. Lee, "Detecting deepfake voice using explainable deep learning techniques," *Applied Sciences 2022, Vol. 12, Page 3926*, vol. 12, p. 3926, 4 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/8/3926/htmhttps://www.mdpi.com/2076-3417/12/8/3926

[23] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, and S. Tubaro, "Deepfake speech detection through emotion recognition: a semantic approach," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8962–8966.

[24] S. Muppalla, S. Jia, and S. Lyu, "Integrating audio-visual features for multimodal deepfake detection." [Online]. Available: https://gizmodo.com/bank-robbers-in-the-middle-east-reportedly-

[25] C. Yu, P. Chen, J. Tian, J. Liu, J. Dai, X. Wang, Y. Chai, S. Jia, S. Lyu, and J. Han, "A unified framework for modality-agnostic deepfakes detection." [Online]. Available: https://www.youtube.com/shorts/j0v4UMnHn1M

[26] P. Wang, T. N. Sainath, and R. J. Weiss, "Multitask training with text data for end-to-end speech recognition," 2021.