Review of Existing Methods for Generating and Detecting Fake and Partially Fake Audio

Abdulazeez AlAli
School of Computer Science and Informatics
Cardiff University
Cardiff, United Kingdom
aliaa8@cardiff.ac.uk

George Theodorakopoulos School of Computer Science and Informatics
Cardiff University
Cardiff, United Kingdom
theodorakopoulosg@cardiff.ac.uk

ABSTRACT

Using deep-learning technologies, both text-to-speech (TTS) and voice conversion (VC) methods can generate fake speech effectively, making it challenging to differentiate between real and fake speech. Accordingly, researchers have employed deepfake detection solutions to distinguish them. These solutions can achieve high detection accuracy and exhibit robustness against unseen data, which are data that differ from those used in initial model training. The emergence of partially fake (PF) audio, which combines real and fake speech, presents a new challenge for deepfake detection. This tutorial presents a comprehensive overview of TTS, VC, and PF generation and detection methods and analyses the characteristics of publicly available datasets for each type. Furthermore, it highlights directions for PF detection that can pave the way for valuable research in fake speech detection.

CCS CONCEPTS

• **Security and privacy→ Spoofing attacks;** Privacy protections simulations.

KEYWORDS

Fake speech; text-to-speech; deepfake audio; deepfake detection; partial fake speech.

ACM REFERENCE FORMAT:

Abdulazeez AlAli and George Theodorakopoulos. 2024. Review of Existing Methods for Generating and Detecting Fake and Partially Fake Audio. In Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics (IWSPA '24), June 21, 2024, Porto, Portugal. ACM, New York, NY, USA, 2 pages.

1. INTRODUCTION

Deepfake is a technique that employs artificial intelligence (AI) and machine learning algorithms to alter or replicate media convincingly, typically images or videos; 'deepfake' is derived from the words 'deep learning' and 'fake'. Deepfake audio generated using artificial neural networks (ANNs) impressively resembles the voices of real people. Although this advancement is useful for multiple applications, attackers can exploit it.

Recently, deepfake videos and images have garnered significant interest. Deepfake videos are directly connected to deepfake audio, with several successful detection techniques for image and video classification having achieved excellent accuracy results in audio classification [1, 2]. Audio deepfake research can be classified into two main categories: generation and detection.

Generation: A deep learning model is trained using a large dataset of audio samples from the target individual and learns to recognise the distinctive qualities and subtleties of their voice. Typically, the procedure consists of the following steps:

Dataset Collection: The voice of the target individual is featured in a large audio-recording dataset, which contains speeches, interviews, or any other content that demonstrates the vocal expressions of the target in different contexts.

Preprocessing: Researchers analyse the audio data to extract essential features and prepare them for deep learning model training; this includes transforming audio into spectrograms or other model-appropriate representations.

Model Training: A deep learning model such as a recurrent neural network (RNN) or convolutional neural network (CNN) is trained on a pre-processed audio dataset and learns to match the input audio features with the vocal characteristics of the target individual.

Fine-tuning: Following the initial training, the model can be fine-tuned to improve the performance and accuracy of the generated audio. This step requires additional training on a smaller scale.

Audio Generation: Finally, the model can generate new audio samples replicating the target voice by inputting text.

The two main ways of generating fake speech can be classified as follows:

Text-to-Speech Generation: A TTS system converts natural language text inputs into artificial human-like spoken outputs. Text normalisation, aligned linguistic featurisation, mel-spectrogram synthesis, and raw audio waveform synthesis are examples of the phases included in typical TTS pipelines that are independently trained or created [13]. A TTS system has become essential in many devices and applications, including speech-enabled gadgets, navigational systems, screen readers, and telephone information systems.

Voice Conversion Generation: Voice Conversion aims to alter the speech of a source speaker such that the created output resembles a sentence uttered by the target speaker [8]. Significant work has been performed to improve the quality of converted voices, with higher-quality voices having been achieved through deep learning

IWSPA '24, June 21, 2024, Porto, Portugal

ACM ISBN 979-8-4007-0556-4/24/06.
https://doi.org/10.1145/3643651.3659894

[14]. These improvements have led to effective attacks against automatic speaker verification systems as well as privacy and authentication concerns [15]. Notably, VC can be used in different applications such as creating a new personalised voice, games and entertainment, voice recovery, vocal pathology, and identity concealment [16].

Detection: Detecting audio deepfakes is difficult because of the complexity of the techniques used to generate them. Researchers are actively working to simplify this process. Currently, they are targeting three types of audio: text-to-speech (TTS) audio, voice conversion (VC) audio, and partially fake (PF) audio, which is a new type of fake audio that combines both real and fake speech (TTS or VC) with real audio. This new audio has created cause for concern in deepfake audio detection because most existing detection solutions only distinguish entirely real audio from entirely fake audio.

Most recent surveys on deepfake generation and detection only focus on images and videos, neglecting deepfake audio [3]; even if these surveys consider audio, images, and videos, they neglect PF generation and detection [4, 5]. Currently, there are a few comprehensive deepfake audio surveys, each of which focuses on a specific problem. The researchers in [6, 7] and [8, 9] focused exclusively on TTS and VC generation, respectively. Recent surveys have targeted the detection of TTS and VC, neglecting or only partially considering PF detection [10, 11]. Although the authors of [12] considered all three types of deepfake media, they barely considered audio deepfakes. Additionally, none of these surveys considered PF generation, detection, or datasets.

Our review aims to clarify PF creation and detection techniques, including determining whether existing detection tools can detect the combinations of PF audio that can be produced. Additionally, it describes existing PF datasets and discusses the need to create more challenging PF datasets.

2. CONCLUSION

This tutorial provides comprehensive methods for generating and detecting TTS, VC, and PF. Audio generation techniques have undergone significant advancements in recent years, making it challenging to differentiate between real and fake audio. Current advancements in TTS systems are particularly evident when generating TTS voices with specific emotions. Recognising a converted voice is challenging because it follows all the rhythms and emotions of the source speaker. However, detection techniques must exhibit resilience to adversarial attacks to ensure the efficacy of detection algorithms. Publicly available fake audio datasets can be categorised according to the generation type; the limitations of these datasets and their key characteristics are highlighted. Clearly, there are significant challenges in detecting PF audio recordings. Therefore, there is an urgent need to develop a new defence method that can effectively identify fake audio types, including entirely fake audio, such as TTS, VC, and PF audio.

REFERENCES

[1] Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromański, Mario Lučić, Yi Tay, Adrian Weller, and Mostafa Dehghani. 2021. PolyVIT: Co-training vision transformers on images, videos and audio. *arXiv Preprint arXiv:2111.12993* (2021).

[2] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. of INTERSPEECH.*

[3] Tao Zhang. 2022. Deepfake generation and detection, a survey. Multimedia Tools and Applications 81, 5: 6259–6276. https://doi.org/10.1007/s11042-021-11733-y.

[4] Enes Altuncu, Virginia N. L. Franqueira, and Shujun Li. 2022. Deepfake: definitions, performance metrics and standards, datasets and benchmarks, and a Meta-Review. *arXiv Preprint arXiv:2208.10913* (2022).

[5] Abhishek Dixit, Nirmal Kaur, and Staffy Kingra. 2023. Review of audio deepfake detection techniques: Issues and prospects. Expert Systems 40, 8. https://doi.org/10.1111/exsy.13322.

[6] Chenshuang Zhang, Chaoning Zhang, Shusen Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. 2023. A survey on Audio Diffusion models: text to speech synthesis and Enhancement in Generative AI. *arXiv Preprint arXiv:2303.13336* (2023).

[7] Yogesh Kumar, Apeksha Koul, and Chamkaur Singh. 2022. A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. Multimedia Tools and Applications 82, 10: 15171–15197. https://doi.org/10.1007/s11042-022-13943-4.

[8] Seyed Hamidreza Mohammadi and Alexander Kain. 2017. An overview of voice conversion systems. Speech Communication 88: 65–82. https://doi.org/10.1016/j.specom.2017.01.008.

[9] Tomasz Walczyna and Zbigniew Piotrowski. 2023. Overview of voice conversion methods based on deep learning. Applied Sciences 13, 5: 3100. https://doi.org/10.3390/app13053100.

[10] Zaynab Almutairi and Hebah ElGibreen. 2022. A review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. Algorithms 15, 5: 155. https://doi.org/10.3390/a15050155.

[11] Zahra Khanjani, Gabrielle Watson, and Vandana P. Janeja. 2023. Audio deepfakes: A survey. Frontiers in Big Data 5. https://doi.org/10.3389/fdata.2022.1001063.

[12] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. 2022. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence 53, 4: 3974–4026. https://doi.org/10.1007/s10489-022-03766-z.

[13] Paul Taylor. 2009. Text-to-Speech synthesis. https://doi.org/10.1017/cbo9780511816338.

[14] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2019. Cyclegan-VC2: Improved Cyclegan-based Non-parallel Voice Conversion. International Conference on Acoustics, Speech, and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2019.8682897.

[15] Chien-Yu Huang, Yist Y. Lin, Hung-Yi Lee, and Lin-Shan Lee. 2021. Defending Your Voice: Adversarial Attack on Voice Conversion. IEEE Spoken Language Technology Workshop. https://doi.org/10.1109/slt48900.2021.9383529.

[16] S. Ramakrishnan. 2012. Speech Enhancement, Modeling and Recognition-Algorithms and applications. BoD – Books on Demand.