Towards Proactive Protection against Unauthorized Speech Synthesis

Zhiyuan Yu
Washington University in St. Louis
St. Louis, USA
yu.zhiyuan@wustl.edu

Ning Zhang
Washington University in St. Louis
St. Louis, USA
zhang.ning@wustl.edu

ABSTRACT

The rapid advancement of artificial speech synthesis technologies, fueled by generative AI (GenAI), presents both opportunities and potential threats to society. While offering unprecedented opportunities, these technologies have been exploited to create “DeepFake” speech for fraud, impersonation, and spreading disinformation, as evidenced by recent real-world incidents. Our research aims to address such emerging threats by exploring a novel, proactive approach to disrupt unauthorized speech synthesis.

Grounded in adversarial robustness theories, the core strategy is to embed imperceptible “voice cloaks” into users’ speech. These perturbations are designed to prevent accurate voice cloning when used in synthesis processes. This concept has been realized and validated in our preliminary work, AntiFake, demonstrating the initial feasibility. Building on these foundations, we propose a line of research that seeks to understand the fundamental three-way trade-off across protection generalizability, audio quality, and computational efficiency, and further achieve balanced improvements across these dimensions.

ACM Reference Format:
Zhiyuan Yu and Ning Zhang. 2024. Towards Proactive Protection against Unauthorized Speech Synthesis. In *Proceedings of ACM Conference (Conference’17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Motivation from Real-world Threats. The advent of artificial speech synthesis presents a double-edged sword in our society. While it offers unprecedented opportunities for improving lives, it also poses significant threats in the real world. Recent reports have revealed the misuse of speech synthesis to conduct fraud [18], impersonation [6], and even disrupt election events [3]. Exacerbated by the widely available generative AI (GenAI) techniques, this issue becomes even more widespread and harmful. Therefore, it is imperative to establish protective mechanisms to safeguard the broad and diverse population.

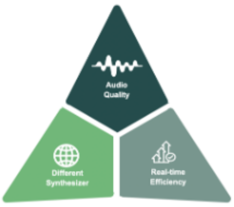

Figure 1: Inherent tradeoffs in proactive protection.

Proactive Defenses and Scientific Principles. Addressing such threats necessitates multi-faceted protection. Traditional countermeasures mostly focus on detection, leveraging diverse features such as physical characteristics [2, 17, 27], acoustic features [10, 14, 20], and non-explainable latent representations [9, 19, 28]. Complementary to such defenses, an emerging direction is to proactively hinder the synthesis process, as demonstrated in our preliminary work named AntiFake [26]. Inspired by similar solutions in the image domain [5, 15, 16], AntiFake in the audio domain works by adding imperceptible voice “cloaks" to users’ speech. When such samples are collected by the attackers and used for synthesis, the synthetic speech will not resemble the victim speaker.

The fundamental principle behind such approaches lies in adversarial examples, which are inherent vulnerabilities in the broad DNN models including speech synthesizers. To ensure generalizable protection against diverse synthesizers potentially used by the attacker, a key assumption we made is that these models share a certain level of similarity in feature space due to their common objective of mimicking speaker characteristics. This assumption is further validated in our preliminary study, thus consolidating the foundation for future improvements.

Practical Challenges and Paths Forward. Our engagement with real-world users and theoretical analysis has deepened our understanding of the remaining challenges. There exists a three-way trade-off involving generalizability across synthesizers, the quality of modified samples, and the computation efficiency for optimization. Guided by this vision, our future efforts aim to achieve balanced improvements along these dimensions, with theoretical lower bounds of the protection provided.

2 SCENARIO AND DEFENSE CONCEPT

The conceptual framework involves two parties: the *user*, who uploads their voice samples to public online platforms (e.g., streaming services and social media apps); and the *attacker*, who collects the target user’s voice samples to create DeepFake speech.

Conference’17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

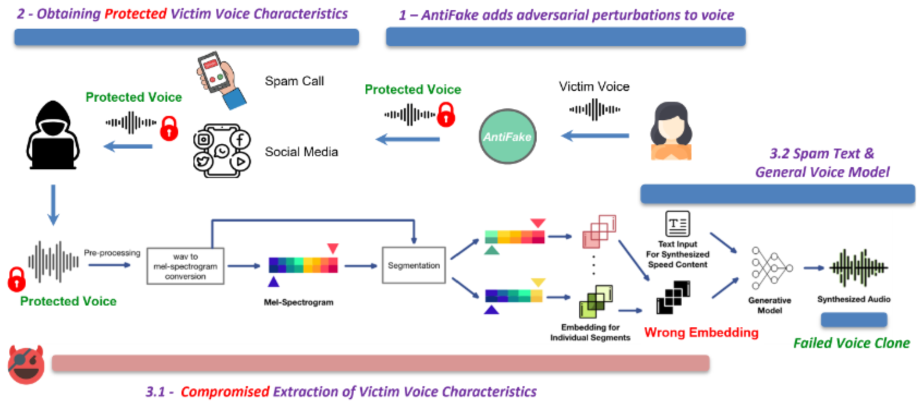
2 - Obtaining Protected Victim Voice Characteristics
1 – AntiFake adds adversarial perturbations to voice
Protected Voice
Spam Call
Social Media
Protected Voice
AntiFake
Victim Voice
3.2 Spam Text & General Voice Model
Pre-processing
wav to mel-spectrogram conversion
Mel-Spectrogram
Segmentation
Text Input For Synthesized Speed Content
Generative Model
Synthesized Audio
Protected Voice
Embedding for Individual Segments
Wrong Embedding
Failed Voice Clone
3.1 - Compromised Extraction of Victim Voice Characteristics

Figure 2: Workflow of AntiFake for proactive protection.

The attacker aims to generate DeepFake speech in the target speaker's voice. These synthetic voice samples can be exploited for various malicious purposes, such as financial scams, stealing sensitive information, spreading misinformation, and bypassing voice authentication systems. The attacker is assumed to have no direct/-physical access to the target's speech. Instead, such samples will be collected from online sources such as social media and streaming services. After that, the attacker could leverage commercial or open-source models for zero-shot synthesis using these samples.

3 PRELIMINARY WORK

The initial concept was realized in our preliminary work named AntiFake [26], as outlined in Figure 2.

The key component that enables synthesizers to mimic speaker identities lies in speaker embeddings extracted via encoders. These DNN-based models are commonly trained to produce highly similar embeddings for the same speaker across different speech content, while distinctly differentiating between different speakers. Guided by this observation, the ultimate goal of adding perturbations is to deviate the extracted speaker embedding, such that the subsequent synthesis conditioned on this representation will be disrupted. Such perturbations are identified through an iterative optimization process, guided by gradients from a set of known encoders.

The developed AntiFake was evaluated across diverse synthesizers and datasets. For speech synthesis engines, we tested four open-source synthesizers, *SV2TTS* [22], *YourTTS* [4], *TorToiSe* [1], *Adaptive Voice Conversion (AdaptVC)* [13], and one commercial product named *ElevenLabs* [7]. Four speech datasets were involved, including VCTK [24], LibriSpeech [12], Speech Accent Archive (SAA) [23], and TIMIT [8]. The results showed that it achieved over 95% rate against a total of 600 synthetic speech samples, with a mean mean opinion score (MOS) of 3.38 that quantified the modified audio quality. For reference, the mean MOS for the clean TIMIT corpus was measured at 3.45±0.52. Therefore, it indicates that AntiFake can reliably shift the speaker identity in the synthesized speech while preserving the audio quality to a reasonable extent.

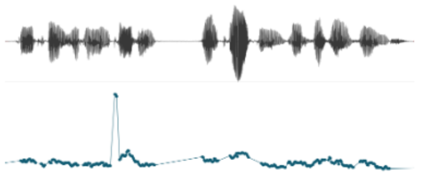

Figure 3: Improved approach for injecting modifications.

As a further investigation of the impacts brought by encoders used by AntiFake, we also examined the effectiveness when ensembling different encoders during optimization. We found that protection effectiveness generally grows when more encoders are used. More importantly, the protection retained a certain level even when completely different encoders were used for optimization and synthesis respectively. This validates our previous hypothesis that speaker encoders generally share similarities in their feature spaces. For the detailed table please refer to the original paper.

Our preliminary study has gained interest from a wide community ranging from the general public and professional voice actors [21]. Engaging in such communication has helped us gain a better understanding of practical needs, which inspire and guide our current ongoing efforts and future directions.

4 ONGOING EFFORTS AND DIRECTIONS

Recognizing the remaining challenges, we propose a line of research with each individually addressing a type of trade-off.

Improve Audio Quality. Inspired by another preliminary work named SMACK [25], one of the ongoing efforts involves more natural manipulation of speech, such as pitch contours, to maintain the natural quality of the audio while still introducing effective perturbations. More specifically, such manipulation involves dynamically adjusting pitch contours based on the linguistic content and emotional tone of the speech, optimized with the dual objectives of protection effectiveness and naturalness of the modified speech. To address the challenge of natural pitch manipulation, we adapted the state-of-the-art generative model [11] that ingests pitch embeddings and output adjusted audio. An example of the resulting speech audio is illustrated in Figure 3.

Improve Generalizability. To ensure the protection can generalize well to other unknown synthesizers, we are working to leverage knowledge distillation techniques to cover a wider range of speaker feature spaces. Specifically, we utilize a set of well-established teacher encoders trained on a vast and diverse dataset of speaker voices, under different acoustic environments and noise conditions. Subsequently, a student model is trained on the outputs and intermediate representations of the teacher models. During this process, linear projection is used to align embeddings with varying lengths. As such, this distilled model learns to generalize better across various embedding spaces and transmission conditions.

Improve Efficiency. Our approach focuses on creating a universal perturbation method that leverages insights from explainable AI to target critical acoustic features. Specifically, XAI techniques such as saliency maps are used to identify which features within the speech signal most influence the speaker identification process. These features are then prioritized in the perturbation process to maximize efficiency and impact.

Automatic Benchmarking. To facilitate the evaluation of these novel approaches, we also plan to develop a testing platform that fully automates the process of optimizing protection, synthesizing DeepFake speech, and benchmarking protection strengths. During this process, humans can opt to listen to voice samples and give scores that quantify protection strengths and audio quality. This platform could also be used as a component that guides the optimization direction.

REFERENCES

[1] James Betker. Tortoise tts. https://github.com/neonbjb/tortoise-tts, May 2022.

[2] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. Who are you (i really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2691–2708, 2022.

[3] Shannon Bond. A political consultant faces charges and fines for biden deepfake robocalls. https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative, May 2024.

[4] Edresson Casanova et al. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

[5] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[6] Joseph Cox. How i broke into a bank account with an ai-generated voice. https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice, Feb 2023.

[7] ElevenLabs. Prime voice ai. https://beta.elevenlabs.io/, Jan 2023.

[8] John S Garofolo et al. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report*, 1993.

[9] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6367–6371. IEEE, 2022.

[10] Piotr Kawa, Marcin Plata, Michal Czuba, Piotr Szymanski, and Piotr Syga. Improved deepfake detection using whisper features. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 4009–4013. ISCA, 2023.

[11] Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE, 2021.

[12] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015.

[13] Kaizhi Qian et al. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019.

[14] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. 2015.

[15] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.

[16] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pages 1589–1604, 2020.

[17] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *Sixteenth annual conference of the international speech communication association*, 2015.

[18] Catherine Stupp. Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, Aug 2019.

[19] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386. IEEE, 2022.

[20] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Odyssey*, volume 2016, pages 283–290, 2016.

[21] Chloe Veltman. Worried about ai hijacking your voice for a deepfake? this tool could help. https://www.npr.org/2023/11/13/1211679937/ai-deepfake, Nov 2023.

[22] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez-Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, April 2018*, pages 4879–4883. IEEE, 2018.

[23] Steven H Weinberger and Stephen A Kunath. The speech accent archive: towards a typology of english accents. In *Corpus-based studies in language use, language learning, and language documentation*, pages 265–281. Brill, 2011.

[24] Junichi Yamagishi, Christophe Veaux, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[25] Zhiyuan Yu, Yuanhaur Chang, Ning Zhang, and Chaowei Xiao. {SMACK}: Semantically meaningful adversarial audio attack. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3799–3816, 2023.

[26] Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 460–474, 2023.

[27] Linghan Zhang, Sheng Tan, and Jie Yang. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 57–71, 2017.

[28] Xiaohui Zhang, Jiangyan Yi, Jianhua Tao, Chenglong Wang, and Chu Yuan Zhang. Do you remember? overcoming catastrophic forgetting for fake audio detection. In *International Conference on Machine Learning*, pages 41819–41831. PMLR, 2023.