Attacks on Neural Networks in a Lightweight Speech Anonymization Pipeline

Daan Brugmans
Radboud University
daan.brugmans@ru.nl

Abstract

1 Introduction

As advancements in the field of Automatic Speech Recognition (ASR) have accelerated with the rise of modern End-to-End neural networks, the risks associated with using such models in ASR applications has become more evident.

Modern neural ASR models are capable of parsing and producing speech to a new level of authenticity: transformers are State-of-the-Art for ASR and word recognition, and the introduction of modern unsupervised deep neural network architectures, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), has allowed for more realistic, accurate, and easier generation of speech. These modern speech generation models are capable of learning to reproduce a person's voice, and then generating new speech using the learned voice. Such synthesized speeches are called *deepfaked* speeches, or simply *deepfakes*.

The presence and influence of deepfakes has become increasingly apparent in recent years: neurally synthesized audio and video of important persons are used to spread misinformation and manipulate. One way to counteract the repercussions of deepfakes is the removal of the personalization in the learned, and thus reproduced, speech. This is called *Speaker Anonymization*. In Speaker Anonymization, we aim to maintain the ASR quality of the audio, while applying changes that make the audio untraceable to a person's likeness.

Although modern Speaker Anonymization systems, often neural in nature, have been shown to be able to anonymize speech while maintaining ASR quality, they can also be manipulated. By attacking neural Speaker Anonymization systems, we may be able to circumvent the preventative measures they provide, and generate speech to a person's likeness regardless of their presence. This paper will focus in that topic: attacking neural Speaker Anonymization systems.

2 Related Work

Meyer et al. (2023)
Chen et al. (2024)
Yuan et al. (2019)
Goodfellow et al. (2015)
Madry et al. (2018)
Gu et al. (2019)
Liu et al. (2018)
Koffas et al. (2022)
Koffas et al. (2023)

3 Method

Kai et al. (2022)

4 Experiment

5 Results

6 Discussion

7 Conclusion

References

Shihao Chen, Liping Chen, Jie Zhang, KongAik Lee, Zhenhua Ling, and Lirong Dai. 2024. Adversarial speech for voice privacy protection from personalized speech generation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11411–11415.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. Badnets: Identifying vulnerabilities in the machine learning model supply chain.

Hiroto Kai, Shinnosuke Takamichi, Sayaka Shiota, and Hitoshi Kiya. 2022. Lightweight and irreversible

speech pseudonymization based on data-driven optimization of cascaded voice modification modules. *Computer Speech & Language*, 72:101315.

Stefanos Koffas, Luca Pajola, Stjepan Picek, and Mauro Conti. 2023. Going in style: Audio backdoors through stylistic transformations. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. 2022. Can you hear it? backdoor attacks via ultrasonic triggers. In *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning*, WiseML '22, page 57–62, New York, NY, USA. Association for Computing Machinery.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2018. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-221, 2018*. The Internet Society.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Sarina Meyer, Pascal Tilli, Pavel Denisov, Florian Lux, Julia Koch, and Ngoc Thang Vu. 2023. Anonymizing speech with generative adversarial networks to preserve speaker privacy. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 912–919.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824.