
F24 11785 Group 56 Initial Final Report: Attacking Voice Anonymization Systems for the Future of Voice Privacy

Himansh Mulchandani,
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213
hmulchan@andrew.cmu.edu

Sarvesh Siras
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213
ssrias@andrew.cmu.edu

Dhwanil Vasani
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213
dvasani@andrew.cmu.edu

John Liu
Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
johnliu@andrew.cmu.edu

Abstract

This project proposes the usage of the deep transformer-based WavLM model to provide a stronger attacker model for modern speech anonymization methods. WavLM's speech verification capabilities are compared to the baseline ECAPA-TDNN model approach provided by the Voice Privacy Challenge 2025 hosted by the International Conference on Acoustics, Speech, and Signal Processing. This project sets up a system for evaluating other models, and it finds that the more sophisticated WavLM model out-performs the provided baseline models with further implications on speech security.

Introduction

The majority of smart devices and services available today are speech-driven. It is becoming more common for televisions, smartphones, loudspeakers, watches, intelligent digital assistants, and even smart cars to record speech for biometric verification. A speech recording can contain sensitive information, including age, gender, emotional state, and identity. It is a valuable source of personal, sensitive information that can be used for various applications, such as health profiling and biometric identification.[2] The concept of privacy preservation is classically defined as a game between users who process their utterances (known as trial utterances) with a privacy preservation system before sharing them with others, and attackers who access these processed utterances or data derived from them and wish to infer information about the users. In terms of privacy, a solution with the lowest error rate among all attackers is considered a high level of privacy. Anonymization systems are designed to disguise gender, emotional state, and identity information while maintaining speech's intelligibility and emotional content. Systems developed for this task should generate a speech waveform output; anonymize or conceal speaker identity at the utterance level; and preserve the

linguistic and emotional content without distortion.[3] The processed utterances sound like they were spoken by another speaker, known as a pseudo-speaker. In practice, many voice anonymization systems select the pseudo-speaker or adjust the prosody in a random or semi-random way using a random number generator. Developing a powerful attacker for each anonymization solution is crucial to ensuring a reliable and trustworthy privacy evaluation. So, we are developing an attacker system that attempts to break the anonymized speech data by identifying the original speaker. It is assumed that the attacker has access to one or more of the utterances—known as enrollment utterances—spoken by each speaker of interest. The developed attacker system outputs an Automatic Speaker Verification (ASV) score for each pair of trial utterances and enrollment speaker, indicating whether they belong to the same speaker. Lower ASV values imply that the trial utterance and enrollment speaker are from distinct speakers, whereas higher levels strongly suggest that they are from the same person.

Baseline Selection

For our baseline system, we chose the ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation in TDNN)[1] architecture, as employed in the VoicePrivacy 2024 Challenge attacker system. This model, originally designed for speaker verification, provides a robust baseline due to its state-of-the-art performance and recent development, aligning well with current advancements in speaker recognition. ECAPA-TDNN incorporates channel-specific attention mechanisms, which allow the model to differentiate subtle speaker-specific characteristics across audio channels. These enhancements make it more adept at capturing fine-grained speaker features, even within anonymized datasets, which is essential when considering privacy-preserving transformations. The ECAPA-TDNN model’s role in the attacker system setup, where it computes average speaker embeddings for anonymized data, provides a standardized approach to evaluating vulnerabilities in anonymization. This baseline selection, adapted from SpeechBrain’s VoxCeleb recipe to LibriSpeech, offers a tested and documented methodology, making it a reproducible and well-validated choice.

Approach

[4] introduces WavLM (and its 3 variations), which is a completely self-supervised model that learns from 94k hours of speech data. WavLM is based on HuBERT model and combines masked speech prediction and speech denoising as the pre-training objective. Additionally, it also uses convolutional layers to do feature extraction before passing the input to the transformer models. Since this is a pre-trained foundational model, it can be used for various tasks, including speaker verification. The authors have shown that it performs well on speaker recognition and verification tasks and outperforms ECAPA-TDNN on many datasets.

This project will be using the WavLM model, which is a large-scale model that is pre-trained on 83k hours of speech data. This model has already learned a strong representation of audio, and can be finetuned for any downstream task.

This project shows the results of the WavLM model’s implementation and the evaluation of its performance against the baseline.

Proposed Solution to Baseline Comparison

The Voice Privacy Challenge provided benchmarks for the performance of the baseline ECAPA-TDNN solution, and reveals that it results in an EER of 5.72 in the non-anonymized Libri-Speech clean development split, and an EER of 25.24 in the same development set with the B3 anonymization method. In our comparison, we evaluated WavLM on the same dataset, and compared it to the average of the baseline’s performance, 15.48

Model	Cosine Distance EER %	Cosine Distance EER %
ECAPPA-TDNN	15.48	N/A
WavLM-Base	0.5349	0.5239
WavLM-Base+	0.5352	0.5279
WavLM-Large	0.5606	0.5284

Table 1: ASV evaluation results for different models on a mix of the librispeech dev data split, and its B3 anonymization.

Proposed Future Work

Augmentations

To improve the model generalization capability, we propose to carefully add augmentations that would be relevant and improve the speaker recognition model’s accuracy.

1. Noise addition
2. Time and Pitch Manipulation
3. AirAbsorption - This simulates air absorption effect
4. SpecFrequencyMask - Applies Frequency Masking augmentation

Model extensions

Multiple MLPs can be added over the WavLM model and the model can be trained using triplet loss, where the positive, negative, and anchor data would be the anonymized utterance from the same speaker, utterance from a different speaker, and the input utterance from a particular speaker. It can then be trained with the Triple Loss function described below.

$$L(A, P, N) = \max(\|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + a, 0)$$

A : Anchor — the sample we’re focusing on.

P : Positive — a sample anonymized utterance from the same speaker.

N : Negative — Utterance from a different speaker.

$f(x)$: Feature embedding function, which maps the input to a latent space where distances are computed.

a : Margin — a positive constant that ensures the distance between the anchor-positive pair is sufficiently smaller than the anchor-negative pair.

This is during the train time. Once training is done, in the test time, the speaker can be determined just by comparing the cosine distance between the input utterance and enrolled utterance.

Division of Labor

(In approximate chronological order)

- Himansh Mulchandani implemented a large portion of the system to train and evaluate WavLM, including defining the dataloader, evaluation framework, and the proposed Triplet Loss Function which will be used in future iterations.
- John Liu performed the evaluations of the baseline WavLM models and formalized the results into the project presentation and video.
- Dhwanil Vasani provided administration over the team’s computing resources and helped to implement additional metrics for learning later on such as the Contrastive Loss function.
- Sarvesh Siras Assisted with the development of the system’s framework and provided technical support to the other members’ tasks.

Acknowledgments

We would like to thank Mr. Puru Samal for his mentorship and guidance thus far. His support, insights in the field of sound engineering, and his expertise in deep learning were invaluable in our accomplishments thus far. We would also like to acknowledge Professor Bhiksha Raj for suggesting this project and for sharing his expertise in the field.

References

- [1] M. Desplanques, D. D. K. Zhang, S. D. G. A. K., and A. M. K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.
- [2] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, et al. Preserving privacy in speaker and speech characterisation. *Computer Speech and Language*, 58:441–480, 2019.
- [3] N. Tomashenko, X. Miao, E. Vincent, J. Yamagishi, and VoicePrivacy Challenge. The first voiceprivacy attacker challenge evaluation plan, 2024. VoicePrivacy Challenge.
- [4] Y. Zhang, T. Xu, X. Wu, and D. Yu. Generalized voice conversion based on conditional gans. *arXiv preprint arXiv:2110.13900*, 2021.