Voice Attacker Leveraging Multi-Head Factorized Attentive Reconstructor and Gradient Reversal for Random Prosody Anonymization

Nhan Tri Do

VinBigData Joint Stock Company, Vietnam

{dotrinhan99}@gmail.com

Abstract

This is the report for Team 04-SpeechWorld in the First VoicePrivacy Attacker Challenge. The attack methods aimed to verify speakers anonymized by two main anonymization systems: STTTS-based and NAC-based. The characteristics of the original audio were reconstructed using speaker embeddings from WavLM-Ecapa and codecs for the NAC system. Additionally, gradient reversal layers were incorporated to eliminate dependencies on prosody features that were randomly simulated by the anonymization models. The results show that the proposed attackers achieved a relative improvement of 26.49% in Equal Error Rate (EER) compared to the baseline, reducing it from 43.22% to 31.77% for the T12-5 attacker system.

Index Terms: Anonymization, Gradient Reversal, WavLM

1. Introduction

The VoicePrivacy Attacker Challenge [1] focuses on developing attacker systems designed to bypass voice anonymization [2], or in other words, creating automatic speaker verification systems to authenticate anonymized speakers. The evaluation metric used in this challenge is the Equal Error Rate (EER). The training data consists of the "train-clean-360" subset of LibriSpeech, which has been anonymized by different anonymization systems, and the performance of the attacker systems is then evaluated by verifying speakers in LibriSpeech dev-clean and LibriSpeech test-clean.

In this competition, we chose two types of systems for the attack: B3, T12-5 [3], which is based on phonetic transcription, pitch and energy modification, and artificial pseudo-speaker embedding generation, and B4, which is based on neural audio codec language modeling.

The general approach to this problem is to reconstruct the original audio information before it was anonymized. For the B3 system, we aim to reconstruct the original speaker embedding, while for B4, the original audio is recreated through learning the codec mapping. The pretraining models used include WavLM large [4] and EnCodec neural codec [5].

We will present the architecture of the anonymization systems in Section 2, followed by the proposed attack strategies in Section 3, and the experimental results in Section 4.

2. Related Work

Before exploring solutions for attacks, the main methods of the two anonymization models based on ASR and NAC are outlined as follows.

System B3 and T12-5 uses a Wasserstein Generative Adversarial Network with Quadratic Transport Cost (WGAN-QC) to anonymize speaker identity in four steps. First, phonetic transcriptions are extracted using an automatic speech recognition (ASR) model with a hybrid CTC-attention architecture. Speaker embeddings are then obtained using an adapted Global Style Tokens (GST) model. The original speaker embedding is swapped with an artificial one generated by the WGAN, and the process is repeated if the cosine distance between the embeddings exceeds 0.3, with additional adjustments to pitch and energy values. Finally, the anonymized speaker embedding, modified prosody, and original phonetic transcription are used to synthesize anonymized speech with the FastSpeech2 model and HiFi-GAN vocoder.

System B4 utilizes neural audio codec (NAC) language modeling for anonymizing speech. It employs an encoder-decoder neural network, where an audio signal is first converted into discrete acoustic tokens and then decoded back into a waveform. These acoustic tokens, which capture individual speech characteristics, are extracted from a pool of pseudo-speakers and used to anonymize the input speech. The system also uses a semantic extractor to convert the input speech into a sequence of semantic tokens representing the spoken content, which is then concatenated with a randomly selected sequence of acoustic tokens. A GPT-like decoder model generates a continuation of the acoustic tokens that maintains the original semantic content while altering the speaker's identity, and the NAC decoder converts these tokens back into a waveform, preserving the semantics but changing the speaker's style.

3. Method

3.1. WavLM for Speaker Verification on Original Audios

To launch an attack by reconstructing the original audio features, we first need to optimize the Equal Error Rate (EER) for the original speaker's enrollment and trial, in order to preliminarily assess the attacker's effectiveness. Therefore, we tested speaker verification on the original audio and found that using WavLM in the embedding generation function of SpeechBrain improved the attacker's success.

Dataset	Gender	SpeechBrain	WavLM
LibriSpeech-dev	female	10.51	2.555
	male	0.93	0.008
LibriSpeech-test	female	8.76	0.181
	male	0.42	0.011

Table 1: *Speaker Verification on Original Librispeech for SpeechBrain and WavLM*

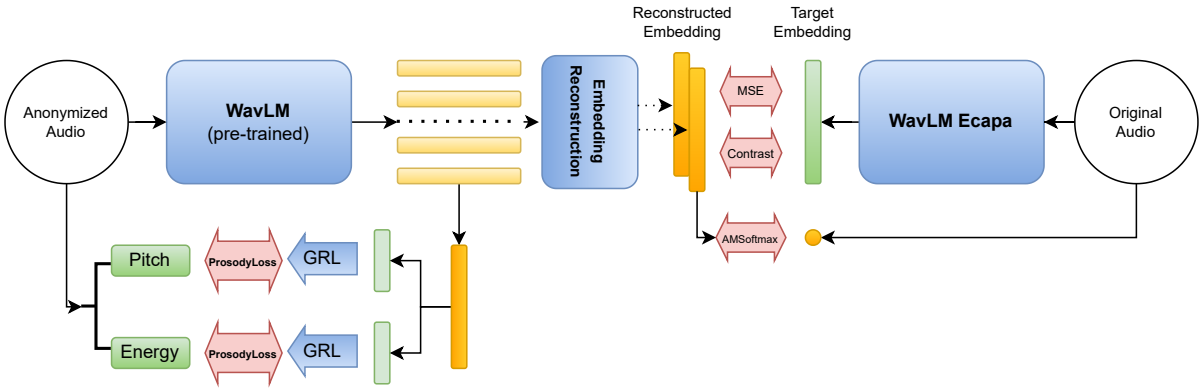

Figure 1: *Proposed Attacker Architecture for Anonymization Systems Based on Phonetic Transcriptions and GAN*

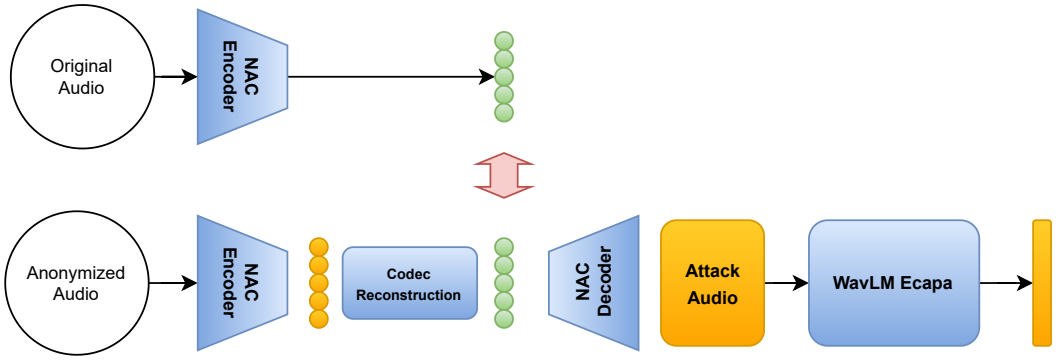

Figure 2: *Proposed Attacker Architecture for Anonymization Systems Based on Neural Audio Codec (NAC) and Language Modeling*

3.2. Embedding Reconstructor with Multi-Head Factorized Attentive WavLM

The anonymized audio is passed through a pre-trained WavLM large model, which generates a stack of output layers with an embedding dimension of 1024. The stack outputs are then passed through a Multi-Head Factorized Attentive to generate an embedding that reconstructs the original audio embedding [6]. Additionally, these stack outputs are averaged and temporally pooled to form an embedding that captures prosody information.

3.3. Gradient Reversal Layer for Random Prosody

In the anonymization system, the pitch and energy values of each phone are multiplied by random values generated uniformly and independently between 0.6 and 1.4 to remove individual prosodic patterns while preserving the overall prosody of the utterance. These random values are applied individually to each phone. To address this randomness, we propose an architecture that removes prosody information from the embedding, making it irrelevant to the speaker embedding, by using a Gradient Reversal Layer. The Gradient Reversal Layer (GRL) is a neural network layer that reverses the gradient during backpropagation, making it useful for domain adaptation and adversarial training [7]. In the forward pass, it acts as an identity function, passing input unchanged. During backpropagation, it multiplies the gradient by a negative constant, encouraging the model to learn domain-invariant features. By applying these layers to pitch and energy, we aim to remove pitch and energy information from the reconstruction of the speaker embedding, thereby eliminating dependence on randomness.

3.4. Loss Function

The target for the Attacker model in B3 and T12-5 is the speaker embedding extracted through the WavLM Ecapa, as evaluated above, and the speaker ID from the Librispeech dataset. The difference between the target and predicted embeddings is calculated using Mean Squared Error (MSE). Additionally, a contrastive embedding, distinct from the original speaker, is generated and evaluated using contrast loss. For the speaker ID label, the model is trained using AAMSoftmax to differentiate between speakers in the Librispeech dataset during the reconstruction training process.

3.5. Neural Audio Codec Reconstruction

For the attack approach that learns to reconstruct audio through a codec, both the original and anonymized audio are passed through the encoder of the NAC to generate frame-codecs, the codec bandwidth used is 24. The model is trained to reconstruct the codec of the original audio through a Reconstruction Layer, which consists of 4 layers of a Retentive Network with

	EER (%)	B3	T12-5	B4
Anonymization	dev-female	28.43	43.32	34.37
	dev-male	22.04	44.10	31.06
	test-female	27.92	43.61	29.37
	test-male	26.72	41.88	31.16
Attacker	dev-female	28.41	32.39	50.15
	dev-male	24.10	33.69	51.55
	test-female	29.33	33.40	50.91
	test-male	22.27	27.61	55.41

Table 2: *Result of Proposed Attackers*

8 heads and a dimension of 32. The reconstructed codecs are then passed through the NAC decoder to simulate the original audio. The original audio is then extracted into an embedding using WavLM Ecapa and undergoes verification through cosine similarity.

4. Experiment

The proposed attackers were implemented and tested on an NVIDIA GeForce RTX 4090 24GB. Prosody features, including energy and F0, were estimated using Parselmouth. Each experiment for the attackers requires an average of 10 epochs to learn and reconstruct the original audio and embeddings.

From the results table, it can be observed that the proposed attack method for the STTS B3 system does not significantly improve results compared to the Speechbrain baseline. However, for the T12-5 anonymization system with a similar method, the average EER decreases from 43.22% to 31.77%. Observing Table 3, we can see that combining WavLM with Multi-Head Factorized Attention and the Gradient Reversal Layer for Random Prosody significantly reduces the Equal Error Rate for the STTTS-based attack method on T12-5. Furthermore, the use of AAMSoftmax introduces an additional improvement by further lowering the error rate. For the approach of using a codec to regenerate the original audio, the result is worse than the baseline.

Method	dev-f	dev-m	test-f	test-m
Baseline	43.32	44.10	43.61	41.88
WavLM	42.044	40.026	40.328	40.533
+MHFA	37.509	35.714	38.658	35.412
+Triplet Loss	38.511	40.528	34.65	35.894
+Pitch GRL	34.801	34.001	34.853	31.627
+AAMSoftmax	32.395	33.697	33.408	27.618

Table 3: *The results of experiments to attack the T12-5 system*

5. Conclusion

The task is to develop an Attacker model to verify speakers whose voices have been anonymized and altered by anonymization systems. We proposed an approach to reconstruct the original audio features through speaker embeddings and codecs, corresponding to two types of anonymization systems: STTS and NAC. By integrating WavLM Ecapa and Encodec modules for feature extraction, along with Gradient Reversal Layers to address the random prosody issue caused by anonymization systems, the resulting EER has improved compared to anonymization and post-processing using Speechbrain.

6. Acknowledgements

We sincerely thank the organizers of the Voice Privacy competition for their invaluable efforts in advancing privacy preservation solutions. Their dedication to organizing this competition has greatly contributed to progress in the field. We also extend our gratitude to the authors and contributors of the foundational research that laid the groundwork for this work.

7. References

[1] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The first voiceprivacy attacker challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.

[2] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The voiceprivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.

[3] N. Kuzmin, H.-T. Luong, J. Yao, L. Xie, K. A. Lee, and E. S. Chng, "Ntu-npu system for voice privacy 2024 challenge," *arXiv preprint arXiv:2410.02371*, 2024.

[4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[5] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[6] J. Peng, O. Plchot, T. Stafylakis, L. Mošner, L. Burget, and J. Černockỳ, "An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 555–562.

[7] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.