GISP-HEU's Submission for VoicePrivacy Attacker Challenge at ICASSP 2025

Yanzhe Zhang[1], Zhonghao Bi[2], Feiyang Xiao[1], Xuefeng Yang[1], Qiaoxi Zhu[3], Jian Guan[1*]
[1]Group of Intelligent Signal Processing, College of Computer Science and Technology, Harbin Engineering University, China
[2]Faculty of Computing, Harbin Institute of Technology, China
[3]Acoustics Lab, University of Technology Sydney, Australia

***Abstract*—This technical report presents our attacker systems developed for the First VoicePrivacy Attacker Challenge, targeting seven anonymization systems: B3, B4, B5, T8-5, T10-2, T12-5, and T25-1. We designed two systems (i.e., ECAPA-PLDA-Mix and TitaNet-Cosine). For target anonymization systems (B3, B4, B5, T8-5, T12-5, T25-1), we proposed ECAPA-PLDA-Mix, which combines an ECAPA-TDNN feature extractor trained on mixed datasets with a PLDA-based scoring module trained on anonymized datasets. It also integrates SpecAugment for data augmentation and contrastive learning strategies to enhance performance. For the target anonymization system (T10-2), we observed inconsistencies in anonymization strategies: training and enrollment data used *spk-level* methods, while trial data employed *utt-level* methods. This discrepancy rendered fine-tuning on anonymized datasets ineffective. To address this, we proposed TitaNet-Cosine, which leverages a pre-trained TitaNet-Large model as the feature extractor and cosine similarity for scoring. Experimental results demonstrate the effectiveness of our systems, achieving average Equal Error Rates (EER) of 24.006%, 23.377%, 28.821%, 26.048%, 32.225%, 28.959%, and 33.068% across the seven anonymization systems (i.e., B3, B4, B5, T8-5, T10-2, T12-5, and T25-1).**

***Index Terms*—voice privacy, source speaker verification**

I. Introduction

The First VoicePrivacy Attacker Challenge [1] aimed to encourage the development of attacker systems targeting voice anonymization. The task of the challenge was to develop multiple systems to attack seven different anonymization systems (i.e., B3, B4, B5, T8-5, T10-2, T12-5, and T25-1). Since the test data did not include any speakers from the training data, the development of attacker systems was regarded as a zero-shot source speaker verification task.

The SLT 2024 Source Speaker Tracing Challenge [2] was the first to introduce the task of source speaker verification, with several approaches demonstrating promising performance during the competition [3], [4]. Wang et al. [3] proposed a contrastive learning strategy to tackle the source speaker verification task.

To tackle the zero-shot source speaker verification task, we proposed ECAPA-PLDA-Mix, which combined an ECAPA-TDNN [5] feature extractor trained on mixed datasets with a PLDA-based [6] scoring module trained on anonymized datasets. The mixed datasets consisted of the original *train-clean-360* subset of the LibriSpeech [7] dataset and the corresponding parallel anonymized dataset. Building on this foundation, we further integrated a robust feature representation strategy and adopted the aforementioned contrastive learning approach.

However, we observed that the data in T10-2 appeared to have been anonymized using *spk-level* methods for train and enrollment data, and *utt-level* methods for trial data, resulting in domain mismatches among the features. Training on datasets that included anonymized data only resulted in degraded attack performance. To address this issue and enhance the zero-shot capability of the system, we proposed TitaNet-Cosine, which utilized a pre-trained TitaNet-Large model [8] as the feature extractor and employed cosine similarity for scoring to perform targeted attacks.

In summary, our approach involved tailored strategies for different anonymization systems as follows. For B5, T12-5, and T25-1, we utilized the vanilla version of ECAPA-PLDA-Mix, which consisted of an ECAPA-TDNN feature extractor trained on a mixed dataset and a PLDA scoring module trained on an anonymized dataset. Building upon the vanilla version, we enhanced the system for T8-5 by incorporating the SpecAugment [9] data augmentation method. Further extending this enhanced version, we integrated the contrastive learning approach to improve the attack performance on B3 and B4. Finally, for T10-2, we employed the TitaNet-Cosine system to conduct the attack. Our attack systems achieved average EER of 24.006% on B3, 23.377% on B4, 28.821% on B5, 26.048% on T8-5, 32.225% on T10-2, 28.959% on T12-5, and 33.068% on T25-1 across the *dev-clean* and *test-clean* subsets.

II. System Description

A. ECAPA-PLDA-Mix

ECAPA-PLDA-Mix consisted of an ECAPA-TDNN feature extraction frontend and a PLDA scoring backend. Next, we introduce the detailed attack methods targeting different anonymization systems.

B5, T12-5, T25-1: The attacks on anonymization systems B5, T12-5, and T25-1 were performed using the vanilla version of ECAPA-PLDA-Mix. The training of the ECAPA-TDNN frontend was conducted using a mixed dataset, composed of the original and anonymized training data, resulting in a model

*Corresponding author

TABLE I

PERFORMANCE COMPARISON OF OUR ATTACKER SYSTEMS AND BASELINE IN TERMS OF EER AGAINST SEVEN ANONYMIZATION SYSTEMS. BOLD ROWS REPRESENT OUR METHODS. SA DENOTES SPECAUGMENT, AND $\mathcal{L}_{\text{CON}}$ REPRESENTS CONTRASTIVE LEARNING. F AND M REPRESENT FEMALE AND MALE RESPECTIVELY. AVG DENOTES THE AVERAGE OF FEMALE AND MALE EER RESULTS FOR EACH SUBSET, AND TOTAL AVG DENOTES THE AVERAGE EER ACROSS THE DEV-CLEAN AND TEST-CLEAN SUBSETS.

Anonymization System	Attacker System	SA	$\mathcal{L}_{\text{con}}$	EER(%)						
				dev-clean			test-clean			Total Avg
				F	M	Avg	F	M	Avg	
B3	Baseline	✗	✗	28.430	22.040	25.240	27.920	26.720	27.320	26.280
	ECAPA-PLDA-Mix	✓	✓	**25.976**	**21.117**	**23.547**	**27.329**	**21.603**	**24.466**	**24.006**
B4	Baseline	✗	✗	34.370	31.060	32.710	29.370	31.160	30.260	31.485
	ECAPA-PLDA-Mix	✓	✓	**27.680**	**23.302**	**25.491**	**20.256**	**22.270**	**21.263**	**23.377**
B5	Baseline	✗	✗	35.820	32.920	34.370	33.950	34.730	34.340	34.355
	ECAPA-PLDA-Mix	✗	✗	**32.528**	**27.796**	**30.162**	**28.507**	**26.454**	**27.481**	**28.821**
T8-5	Baseline	✗	✗	39.630	40.840	40.240	42.500	40.050	41.280	40.760
	ECAPA-PLDA-Mix	✓	✗	**26.419**	**28.069**	**27.244**	**26.070**	**23.637**	**24.854**	**26.048**
T10-2	Baseline	✗	✗	43.630	40.040	41.830	41.970	38.750	40.360	41.095
	TitaNet-Cosine	✗	✗	**34.801**	**32.902**	**33.852**	**30.290**	**30.907**	**30.599**	**32.225**
T12-5	Baseline	✗	✗	43.320	44.100	43.710	43.610	41.880	42.750	43.230
	ECAPA-PLDA-Mix	✗	✗	**32.385**	**29.175**	**30.780**	**27.550**	**26.724**	**27.137**	**28.959**
T25-1	Baseline	✗	✗	42.650	40.060	41.360	42.340	41.920	42.130	41.745
	ECAPA-PLDA-Mix	✗	✗	**35.251**	**31.368**	**33.310**	**33.386**	**32.267**	**32.827**	**33.068**

denoted as $E_{\text{mixed}}(\cdot)$. The PLDA scoring backend was trained using the anonymized training data.

To be specific, each audio segment sampled at 16,000 Hz was first split into $\lfloor l/3 \rfloor$ chunks of 3 seconds each, where l denotes the duration of the audio in seconds. These chunks were then converted into 80-dimensional mel-spectrograms and normalized at the sentence level using mean normalization. This preprocessing step was conducted in accordance with the official settings of the VoicePrivacy 2024 Challenge [10]. The spectrograms were then fed into the ECAPA-TDNN model, which extracted speaker embeddings. These embeddings were passed through a 921-dimensional linear layer to produce classification logits. The model was optimized using the Additive Angular Margin Softmax (AAM) loss [11], while the optimizer and scheduler followed the official challenge settings.

In the training process, a mixed dataset was created by combining the original *train-clean-360* subset of the LibriSpeech dataset and its corresponding anonymized parallel subset at the sample level. This mixed dataset was used to train $E_{\text{mixed}}(\cdot)$ for 10 epochs. Finally, a PLDA scoring backend was trained using the anonymized subset, enabling scoring of speaker embeddings. All training processes were conducted within the SpeechBrain [12] framework.

T8-5: Based on the vanilla version of ECAPA-PLDA-Mix, we incorporated SpecAugment after obtaining the spectrograms and subsequently fed the augmented spectrograms into the ECAPA-TDNN model to attack the anonymization system T8-5. To be specific, the spectrograms were augmented using SpecAugment with a top-K masking strategy. Specifically, for each spectrogram, $K = 2$ random time and frequency masks were applied. The lengths of the time masks t_{mask}, were sampled from $t_{\text{mask}} \sim U(0, 64)$, and the lengths of the frequency masks f_{mask}, were sampled from $f_{\text{mask}} \sim U(0, 8)$. Masked regions in the spectrogram were replaced with zero values.

B3, B4: Following the SpecAugment-enhanced version, we improved the system by incorporating contrastive loss to attack the anonymization systems B3 and B4. To be specific, we modified the one-stage training process of ECAPA-TDNN into a three-stage training process. The first stage remained consistent with the vanilla version. In the second stage, a new ECAPA-TDNN model denoted $E_{\text{orig}}(\cdot)$ was trained for 10 epochs using the original *train-clean-360* subset of the LibriSpeech dataset. The model was used to extract *utt-level* speaker embeddings, which were aggregated into *spk-level* embeddings. In the third stage, $E_{\text{mixed}}(\cdot)$ was fine-tuned using the anonymized subset for 4 additional epochs with a contrastive loss, following the method described in [3]. Specifically, the positive and negative speaker embeddings for the contrastive loss were derived from the *spk-level* embeddings extracted by $E_{\text{orig}}(\cdot)$.

B. TitaNet-Cosine

This attacker system was specifically designed for the T10-2 anonymization system. It utilized the pre-trained TitaNet-Large model as the feature extraction module and employed a cosine similarity backend for similarity computation.

III. RESULTS

Table I presents the attack results on seven anonymization systems, demonstrating that our approach outperformed the baseline across all systems and exhibited strong attack performance on certain systems. For example, on the T8-5 and T12-5 systems, the mean EER decreased from 40.240% on *dev-*

clean and 41.280% on *test-clean* to 27.244% and 24.854%, respectively, and from 43.710% on *dev-clean* and 42.750% on *test-clean* to 30.780% and 27.137%, respectively. The reductions reached 12.996% and 16.426% for T8-5, as well as 12.930% and 15.613% for T12-5. Moreover, in addressing the issue of fine-tuning being less effective for T10-2, our system achieved a reduction of 7.978% on *dev-clean* and 9.761% on *test-clean*. These results indicate that the proposed methods were more effective in attacking anonymization systems with stronger protection. Meanwhile, on systems with relatively weaker protection (e.g., B3 and B4), although the EER reductions were smaller, our approach still maintained robust attack performance.

IV. Conclusion

This report presented our attacker systems targeting seven anonymization methods for the VoicePrivacy Attacker Challenge. We proposed a generalized framework for zero-shot speaker verification, integrating robust feature representation, ECAPA-TDNN, and PLDA scoring, with enhancements from SpecAugment and contrastive learning. When fine-tuning was less effective, a pre-trained TitaNet-Large model with cosine similarity was employed. Experimental results demonstrate the superior performance of our systems compared to the baseline.

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] Z. Li, M. Li, P. Zhang, Y. Ren, Z. Cai, and H. Nishizaki, "The SLT 2024 Source Speaker Tracing Challenge (SSTC 2024) Evaluation Plan."

[3] Q. Wang, H. Guo, J. Kang, M. Du, J. Li, X.-L. Zhang, and L. Xie, "Speaker contrastive learning for source speaker tracing," *arXiv preprint arXiv:2409.10072*, 2024.

[4] Y. Du, D. Zhang, J. Deng, and R. Zheng, "The Fosafer Speaker Verification System for the Source Speaker Tracing Challenge 2024," in *Proc. of the Source Speaker Tracing Challenge 2024*, 2024. [Online]. Available: https://sstc-challenge.github.io/file/zhangdejun.pdf

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. of INTERSPEECH*, 2020, pp. 3830–3834.

[6] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7649–7653.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[8] N. R. Koluguri, T. Park, and B. Ginsburg, "TitaNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8102–8106.

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[10] 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.

[11] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.

[12] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2404.02677*, 2021.