

# Attacking Voice Anonymization Systems with Augmented Feature and Speaker Identity Difference

Yanzhe Zhang<sup>1</sup>, Zhonghao Bi<sup>2</sup>, Feiyang Xiao<sup>1</sup>, Xuefeng Yang<sup>1</sup>, Qiaoxi Zhu<sup>3</sup>, Jian Guan<sup>1\*</sup>

<sup>1</sup>Group of Intelligent Signal Processing, College of Computer Science and Technology, Harbin Engineering University, China

<sup>2</sup>Faculty of Computing, Harbin Institute of Technology, China

<sup>3</sup>Acoustics Lab, University of Technology Sydney, Australia

**Abstract**—This study focuses on the First VoicePrivacy Attacker Challenge within the ICASSP 2025 Signal Processing Grand Challenge, which aims to develop speaker verification systems capable of determining whether two anonymized speech signals are from the same speaker. However, differences between feature distributions of original and anonymized speech complicate this task. To address this challenge, we propose an attacker system that combines Data Augmentation enhanced feature representation and Speaker Identity Difference enhanced classifier to improve verification performance, termed DA-SID. Specifically, data augmentation strategies (i.e., data fusion and SpecAugment) are utilized to mitigate feature distribution gaps, while probabilistic linear discriminant analysis (PLDA) is employed to further enhance speaker identity difference. Our system significantly outperforms the baseline, demonstrating exceptional effectiveness and robustness against various voice anonymization systems, ultimately securing a top-5 ranking in the challenge.

**Index Terms**—Voice privacy, speaker verification, data augmentation, speaker identity difference.

## I. INTRODUCTION

The First VoicePrivacy Attacker Challenge [1], part of the ICASSP 2025 Signal Processing Grand Challenge, aims to develop speaker verification systems that determine whether two anonymized speech signals come from the same speaker, effectively attacking voice anonymization systems. The official baseline [1] adopts ECAPA-TDNN [2] as feature extractor to obtain speaker feature embeddings and a cosine similarity classifier to measure the similarity between anonymized speech for speaker verification.

However, the significant differences in feature distributions between original and anonymized speech make it challenging for an attacker system to determine whether two anonymized speech signals originate from the same speaker. Additionally, voice anonymization systems reduce the distinction between speakers in anonymized speech, further complicating the attack process for speaker verification.

In this paper, based on the official baseline [1], we propose a data augmentation (DA) and speaker identity difference (SID) based attacker system, termed DA-SID. The system utilizes data augmentation techniques, i.e., data fusion [3] and SpecAugment [4], thereby enhancing the robustness of feature representation and reducing the feature distribution gap between original and anonymized speech. Additionally,

probabilistic linear discriminant analysis (PLDA) [5] is employed as the classifier, which effectively separates speaker embeddings by modeling the differences between speakers while accounting for variations within each speaker’s speech. Our DA-SID attacker system achieves significant performance improvements over the official baseline, which was ranked among the top-5 in the First VoicePrivacy Attacker Challenge, highlighting its robustness and competitiveness against state-of-the-art anonymization systems.

## II. PROPOSED ATTACKER SYSTEM

### A. Data Augmentation Enhanced Feature Representation

Anonymization systems introduce significant distribution gaps between speaker features before and after anonymization. To mitigate these gaps, we employ data fusion [3] strategy that combines original dataset  $\mathcal{D}_{\text{orig}}$  with the corresponding anonymized dataset  $\mathcal{D}_{\text{anon}}$  to obtain the fused dataset  $\mathcal{D}_{\text{fused}}$ :

$$\mathcal{D}_{\text{fused}} = \mathcal{D}_{\text{orig}} \cup \mathcal{D}_{\text{anon}}. \quad (1)$$

By jointly learning from the fused dataset, the relation between speaker features before and after anonymization is captured, effectively mitigating distribution differences.

Meanwhile, to further enhance the robustness of the feature representation, the SpecAugment [4] strategy is also employed in the extraction of the speaker feature embedding  $\mathbf{e}$ , as

$$\mathbf{e} = \mathcal{F}(\mathbf{X} \odot \mathbf{M}_{tf}), \quad (2)$$

where  $\mathbf{X}$  is the log-Mel feature of the input speech,  $\mathbf{M}_{tf}$  is the time-frequency-masking matrix applied by SpecAugment, and  $\odot$  denotes element-wise multiplication.  $\mathcal{F}(\cdot)$  represents the ECAPA-TDNN feature extractor, which is optimized by speaker classification with additive angular margin loss [6]. For anonymization systems that retain more speaker information (e.g., B3 and B4), the contrastive loss [7] is added to further enhance the feature representation.

### B. Speaker Identity Difference Enhanced Classifier

To enhance speaker identity difference, we employ PLDA [5] trained on an anonymized dataset as the classifier. Specifically, the similarity score between two speaker feature embeddings  $\mathbf{e}_i$  and  $\mathbf{e}_j$  ( $i \neq j$ ) is calculated in PLDA based on the log-likelihood:

$$s(\mathbf{e}_i, \mathbf{e}_j) = \log p(\mathbf{e}_i, \mathbf{e}_j | H_0) - \log p(\mathbf{e}_i, \mathbf{e}_j | H_1), \quad (3)$$

\*Corresponding author

TABLE I  
PERFORMANCE COMPARISON OF OUR DA-SID ATTACKER SYSTEM AND BASELINE IN TERMS OF EER (%) AGAINST SIX ANONYMIZATION SYSTEMS.  
 $\mathcal{L}_{\text{CON}}$  REPRESENTS CONTRASTIVE LEARNING.

Anonymization System	Attacker System	SpecAugment	$\mathcal{L}_{\text{con}}$	EER on dev-clean subset			EER on test-clean subset			Total Average EER
				Female	Male	Average	Female	Male	Average	
B3	Baseline	✗	✗	28.43	22.04	25.24	27.92	26.72	27.32	26.28
	DA-SID	✓	✓	<b>25.98</b>	<b>21.12</b>	<b>23.55</b>	<b>27.33</b>	<b>21.60</b>	<b>24.47</b>	<b>24.01</b>
B4	Baseline	✗	✗	34.37	31.06	32.71	29.37	31.16	30.26	31.49
	DA-SID	✓	✓	<b>27.68</b>	<b>23.30</b>	<b>25.49</b>	<b>20.26</b>	<b>22.27</b>	<b>21.26</b>	<b>23.38</b>
B5	Baseline	✗	✗	35.82	32.92	34.37	33.95	34.73	34.34	34.36
	DA-SID	✗	✗	<b>32.53</b>	<b>27.80</b>	<b>30.16</b>	<b>28.51</b>	<b>26.45</b>	<b>27.48</b>	<b>28.82</b>
T8-5	Baseline	✗	✗	39.63	40.84	40.24	42.50	40.05	41.28	40.76
	DA-SID	✓	✗	<b>26.42</b>	<b>28.07</b>	<b>27.24</b>	<b>26.07</b>	<b>23.64</b>	<b>24.85</b>	<b>26.05</b>
T12-5	Baseline	✗	✗	43.32	44.10	43.71	43.61	41.88	42.75	43.23
	DA-SID	✗	✗	<b>32.39</b>	<b>29.18</b>	<b>30.78</b>	<b>27.55</b>	<b>26.72</b>	<b>27.14</b>	<b>28.96</b>
T25-1	Baseline	✗	✗	42.65	40.06	41.36	42.34	41.92	42.13	41.75
	DA-SID	✗	✗	<b>35.25</b>	<b>31.37</b>	<b>33.31</b>	<b>33.39</b>	<b>32.27</b>	<b>32.82</b>	<b>33.07</b>

TABLE II  
ABLATION STUDY ON THE IMPACT OF DATA AUGMENTATION (DA) AND SPEAKER IDENTITY DIFFERENCE (SID) ON EER (%).

Method	B3	B4	B5	T8-5	T12-5	T25-1
DA-SID	<b>24.04</b>	<b>23.42</b>	<b>28.82</b>	<b>26.05</b>	<b>28.96</b>	<b>33.07</b>
w/o DA	26.06	24.41	29.22	29.96	29.70	33.75
w/o SID	25.54	27.10	33.10	41.07	37.20	38.93
w/o DA & SID	26.28	31.49	34.36	40.76	43.23	41.75

where  $p(\mathbf{e}_i, \mathbf{e}_j | H_0)$  and  $p(\mathbf{e}_i, \mathbf{e}_j | H_1)$  denote the probabilities under the hypotheses  $H_0$  and  $H_1$ , which represent that  $\mathbf{e}_i$  and  $\mathbf{e}_j$  are from the same speaker or different speakers, respectively.

The optimization of the PLDA classifier increases the similarity score for speaker feature embeddings belonging to the same speaker while decreasing it for embeddings from different speakers, thereby emphasizing distinctions in speaker identity. By leveraging PLDA as the classifier, our system enhances speaker identity difference during verification, thereby improving the overall attack performance.

### III. RESULTS

**Dataset and Metric:** Following [1], our experiments are conducted on the LibriSpeech dataset [8] and the equal error rate (EER) is used as the performance metric, where a lower EER indicates better attack performance.

**Performance Comparison:** The attack performance on six anonymization systems is presented in Table I. The proposed DA-SID system consistently outperforms the official baseline across all anonymization systems. Specifically, for T8-5, DA-SID achieves a remarkable EER reduction of 14.71%, establishing it as the most effective attacker system for T8-5 among all the attacker systems in the challenge [1].

**Ablation Study:** To further evaluate the effects of DA representation and SID-enhanced classifier, we conduct an ablation study, as shown in Table II. The results show a significant degradation in EER performance when either of these components is removed (denoted as w/o), confirming that both DA representation and SID-enhanced classifier are effective in improving attack performance. Furthermore, the results indicate that SID-enhanced classifier provides a greater performance boost compared to DA representation, highlighting it as a more impactful design choice.

Note that due to the lack of details about the anonymization strategy, DA-SID cannot be applied to the T10-2 anonymization system in the challenge. Instead, we use a pretrained model TitaNet-Large [9] to extract speaker embeddings and a cosine similarity classifier for speaker verification, which still outperforms the official baseline.

### IV. CONCLUSION

We proposed a DA-SID attacker system to address the significant feature distribution differences between original and anonymized speech. By integrating data augmentation with probabilistic linear discriminant analysis, our system effectively mitigates distribution gaps and enhances speaker identity difference. Experimental results demonstrate that our system outperforms the official baseline, which was ranked among the top-5 systems in the First VoicePrivacy Attacker Challenge.

### REFERENCES

- [1] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The First VoicePrivacy Attacker Challenge."
- [2] 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.
- [3] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Inf. Fus.*, vol. 57, pp. 115–129, 2020.
- [4] 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," in *Proc. INTERSPEECH*, 2019, pp. 2613–2617.
- [5] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7649–7653.
- [6] 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. APSIPA Annu. Summit Conf.*, 2019, pp. 1652–1656.
- [7] 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.
- [8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5206–5210.
- [9] 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. Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 8102–8106.