ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888822

Fine-tuning TitaNet-Large Model for Speaker Anonymization Attacker Systems

Candy Olivia Mawalim, Aulia Adila, Masashi Unoki
Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{candylim, adila, unoki}@jaist.ac.jp

***Abstract*—Speaker anonymization techniques are crucial for safeguarding user privacy in voice-based applications. However, these methods are susceptible to adversarial attacks that can compromise their effectiveness. This paper proposes attacker systems that leverage the power of fine-tuned TitaNet-Large and ECAPA-TDNN models to identify the original speaker from anonymized speech generated by various anonymization methods. Both pre-trained models are renowned for their state-of-the-art ability to extract robust speaker embeddings. Fine-tuning these models with anonymized speech enables them to identify underlying patterns in anonymized speech. We evaluated the proposed attacker systems against multiple anonymization techniques that performed effectively in a series of voice privacy challenges. Our experimental results underscore the effectiveness of the fine-tuned TitaNet-Large model in breaking through these anonymization methods, as indicated by the reduced equal error rate (EER). This highlights the importance of robust and adaptive anonymization strategies to counter such emerging semi-informed threats.**

***Index Terms*—speaker anonymization, attacker model, semi-informed, TitaNet-Large, voice privacy**

I. Introduction

Speaker anonymization techniques are essential for safeguarding user privacy in voice-based applications [1], [2]. However, the effectiveness of these techniques is constantly challenged by the emergence of sophisticated adversarial attacks [3]. The First Attacker Challenge has highlighted the growing threat to speaker anonymization systems [4]. This challenge focuses on developing attacker systems in the form of automatic speaker verification (ASV) systems, which are capable of identifying speakers even after anonymization.

This paper introduces an attack strategy that compares the power of fine-tuned TitaNet-Large (TitaNet-L) [5] and ECAPA-TDNN models [6]. Both models are renowned for their state-of-the-art performance in the speaker verification task. By fine-tuning these models on a diverse dataset of anonymized and original speech samples, we enable them to identify underlying patterns in anonymized speech, thereby accurately recognizing the original speaker.

Our work focuses on the semi-informed attack scenario, where the attacker has access to the speaker anonymization system and anonymized speech samples. We evaluated our proposed systems against the state-of-the-art anonymization techniques submitted to the VoicePrivacy 2024 Challenge [7]. Our experimental results underscore the effectiveness of the fine-tuned TitaNet-L models in breaking through these defenses, as indicated by the reduced equal error rate (EER). These results highlight the limitations of current anonymization techniques to counter emerging semi-informed threats.

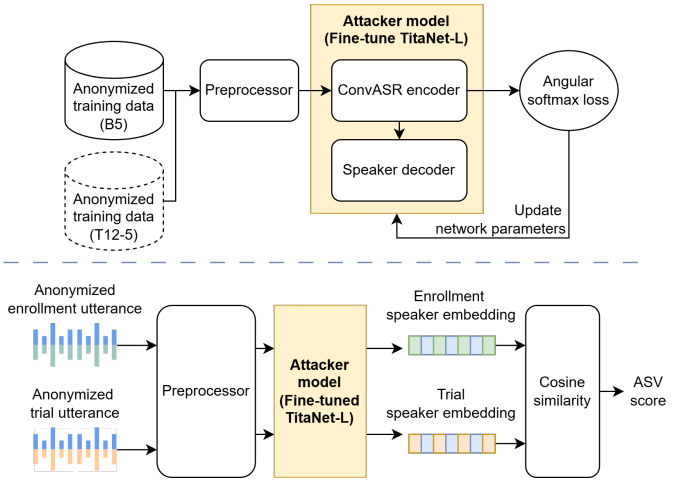

Fig. 1. Fine-tuning phase (Top) and Inference phase (Bottom) of our proposed attacker systems based on fine-tuned TitaNet-L model.

II. Proposed Attacker Systems

Figure 1 shows our proposed system which is based on fine-tuning the TitaNet-Large (TitaNet-L) model [5]. TitaNet-L is the largest variant of the TitaNet architecture with 25.3 million parameters. TitaNet has an encoder-decoder structure. The ConvASR encoder acts as a high-level feature extractor, processing input audio from normalized mel spectrograms. It combines local features extracted through 1D depth-wise separable convolutions with global context information obtained via global average pooling in Squeeze-and-Excitation (SE) layers. This process improves the ability of the network to distinguish speaker-specific characteristics from input audio.

The speaker decoder incorporates an attentive statistics pooling layer to compute attention features across channel dimensions, creating time-independent, utterance-level speaker representations. This layer calculates weighted statistics, allowing the network to focus on the most relevant information for speaker verification. The features are then passed through linear layers to reduce dimensionality and map the resulting 192-dimensional features to the final number of classes, representing the different speakers in the training set.

The model generates fixed-length speaker embeddings, known as t-vectors, as its final output from the decoder. These embeddings encapsulate speaker-specific information and are used in speaker verification tasks, where the cosine similarity between t-vectors serves as the scoring backend.

To fine-tune the TitaNet-L model, we used a 9:1 training-validation ratio. We only fine-tuned the final decoder layer of

Authorized licensed use limited to: Japan Advanced Institute of Science and Tech. Downloaded on April 24,2025 at 04:40:09 UTC from IEEE Xplore. Restrictions apply.

TABLE I
EER (%) COMPARISON OF TITANET-LARGE AND ECAPA-TDNN ATTACKER MODELS ACROSS VARIOUS EXPERIMENTAL CONDITIONS

Method	Training data	Gender	Development							Test							Global avg.
			B3	B4	B5	T8-5	T10-2	T12-5	T25-1	B3	B4	B5	T8-5	T10-2	T12-5	T25-1	
ECAPA-TDNN	Original	F	**28.43**	34.37	35.82	**39.63**	43.63	43.32	42.65	**27.92**	**29.37**	33.95	42.50	41.97	43.61	42.34	37.82
		M	**22.04**	**31.06**	32.92	**40.84**	40.04	44.10	40.06	**26.72**	**31.16**	34.73	**40.05**	38.75	41.88	41.92	36.16
		Avg.	**25.23**	**32.71**	34.37	**40.23**	41.84	43.71	41.36	**27.32**	**30.27**	34.34	41.27	40.36	42.75	42.13	36.99
Fine-tuned ECAPA-TDNN	Anonymized (B5 and T12-5)	F	41.51	42.80	37.97	46.57	39.48	39.18	42.57	38.47	41.04	35.37	46.97	34.61	35.73	41.13	40.24
		M	38.08	42.16	36.34	47.06	34.66	36.15	41.59	37.98	42.16	35.80	45.94	31.91	36.27	40.82	39.07
		Avg.	39.80	42.48	37.15	46.82	37.07	37.66	42.08	38.23	41.60	35.58	46.45	33.26	36.00	40.97	39.65
TitaNet-L	VoxCeleb 1,2; SRE; LibriSpeech; RIR noise; Fisher; Switchboard	F	43.18	46.59	48.27	45.20	33.65	49.29	48.46	42.34	44.53	46.70	45.44	27.14	46.21	48.87	43.99
		M	37.60	44.38	49.35	42.24	30.43	49.40	47.20	40.32	43.43	47.66	44.77	29.40	50.78	46.55	43.11
		Avg.	40.39	45.48	48.81	43.72	32.04	49.34	47.83	41.33	43.98	47.18	45.10	28.27	48.49	47.71	43.55
Fine-tuned TitaNet-L*	Anonymized (B5)	F	36.09	32.67	34.66	43.18	33.40	36.19	38.07	34.13	34.13	33.21	43.79	33.03	32.71	36.32	35.83
		M	34.01	33.23	34.00	44.24	31.95	33.51	**35.44**	33.63	32.50	31.64	43.43	33.85	32.25	**34.08**	34.84
		Avg.	35.05	32.95	34.33	43.71	32.68	34.85	36.75	33.88	33.31	32.42	43.61	33.44	32.48	35.20	35.33
Fine-tuned TitaNet-L**	Anonymized (B5 and T12-5)	F	34.66	**34.23**	**31.96**	41.73	**33.10**	**31.25**	**36.54**	32.15	32.12	**26.24**	**41.42**	**26.66**	**26.82**	**33.75**	**33.04**
		M	33.56	32.92	**27.02**	41.93	**27.64**	**28.42**	36.37	30.72	33.80	**28.06**	40.92	**28.29**	**29.16**	36.07	**32.49**
		Avg.	34.11	33.57	**29.49**	41.83	**30.37**	**29.83**	**36.46**	31.44	32.96	**27.15**	**41.17**	**27.47**	**27.99**	**34.91**	**32.77**
Fine-tuned TitaNet-L	Anonymized (all)	F	45.57	46.57	49.00	48.30	48.72	49.15	48.44	46.35	49.60	49.45	51.82	49.45	49.79	49.77	48.71
		M	49.84	51.41	50.93	51.54	51.06	51.53	51.09	47.89	48.78	49.89	45.88	49.00	49.95	48.56	49.81
		Avg.	47.71	48.99	49.97	49.92	49.89	50.34	49.77	47.12	49.19	49.67	48.85	49.23	49.87	49.17	49.26

The abbreviations F, M, and Avg. under Gender correspond to female, male, and average, respectively. "Original" training data refers to the original, unanonymized training data used for building speaker anonymization models. In contrast, "Anonymized" training data consists of speech samples that have been processed by a specific speaker anonymization model. * and ** indicate the submitted system 1 and 2. Bold font represents the EER obtained by the most effective attacker system in each category.

the model for a maximum of 10 epochs with a batch size of 8. We also performed speed perturbation as a data augmentation technique to enhance the model's robustness to variations in speech speed. We used the angular softmax loss function to improve the discriminative power of embeddings by increasing the angular margin between different classes.

The fine-tuned model was optimized using the AdamW optimizer with a learning rate of 0.0001 and a weight decay of 0.0002. A cosine annealing learning rate scheduler with warmup was used to gradually decrease the learning rate over time, improving convergence and generalization.

III. EXPERIMENTS

We utilized fine-tuned TitaNet-L models[1] [5] as our primary attacker models. We fine-tuned our attacker models on the anonymized speech dataset. Specifically, we focused on anonymized speech generated by the B5 and T12-5 systems, which exhibited high EER scores and shared similar feature extraction and modification techniques.

We compared our proposed attacker model with a baseline ECAPA-TDNN model[2] [4] and its fine-tuned version trained on B5 and T12-5 anonymized speech. Our results, summarized in Table I, demonstrate that our fine-tuned TitaNet-L model outperforms the baseline model overall on anonymized B5 and T12-5 data. Fine-tuning an ASV system on anonymized speech generated by similar methods proves to be an effective strategy for creating a robust attacker model for those particular methods. This approach leads to a significant reduction in the EER, exceeding 10% in both attacks on B5 and T12-5 speaker anonymization methods. However, we also observed a reduction in performance on systems with significantly different anonymization techniques, such as B3, B4, and T8-5.

As a straightforward approach, we also fine-tuned the TitaNet-L model on all anonymized data. However, this approach was ineffective and in fact led to a decrease in performance as the model became confused by the diverse and potentially misleading information present in the anonymized data. This is evident in the increase in EER observed when training on all anonymized data (last row of Table I) compared to the original pretrained model.

[1]https://huggingface.co/nvidia/speakerverification_en_titanet_large

[2]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

IV. CONCLUSION

This paper presents attacker systems based on fine-tuning the TitaNet-L model to compromise speaker anonymization systems. Our results demonstrate the effectiveness of this approach, particularly against similar techniques such as B5 and T12-5. However, its performance degrades against fundamentally different techniques. The ineffectiveness of directly fine-tuning TitaNet-L on all anonymized data underscores the importance of researching training data quality, as well as the sources and anonymization methods applied to datasets, for developing robust attacker models.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI (20KK0233, 23K18491) and the SCAT Foundation.

REFERENCES

[1] N. A. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P. Noé, A. Nautsch, N. W. D. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J. Bonastre, M. Todisco, and M. Maouche, "The voiceprivacy 2020 challenge: Results and findings," *Comput. Speech Lang.*, vol. 74, p. 101362, 2022.

[2] M. Panariello, N. A. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. W. D. Evans, E. Vincent, and J. Yamagishi, "The VoicePrivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 3477–3491, 2024.

[3] T. Yang, L. S. Cang, M. Iqbal, and D. J. Almakhles, "Attack risk analysis in data anonymization in internet of things," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 4, pp. 4986–4993, 2024.

[4] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The First VoicePrivacy Attacker Challenge Evaluation Plan," 2024. [Online]. Available: https://www.voiceprivacychallenge.org/attacker/docs/Attacker_Challenge_Eval_Plan_v2.2.pdf

[5] 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. ICASSP 2022*. IEEE, 2022, pp. 8102–8106.

[6] 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*. ISCA, 2020, pp. 3830–3834.

[7] N. A. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. W. D. Evans, J. Yamagishi, and M. Todisco, "The VoicePrivacy 2024 Challenge Evaluation Plan," *CoRR*, vol. abs/2404.02677, 2024.

Authorized licensed use limited to: Japan Advanced Institute of Science and Tech. Downloaded on April 24,2025 at 04:40:09 UTC from IEEE Xplore. Restrictions apply.