System Description for Voice Privacy Challenge 2022

Xiaojiao Chen[1], *Guangxing Li*[1], *Hao Huang*[1], *Wangjin Zhou*[2], *Sheng Li*[3], *Yang Cao*[2], *Yi Zhao*[4,*]

[1]Xinjiang University, Urumqi, China
[2]Kyoto University, Kyoto, Japan
[3]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[4]Kuaishou Technology, Beijing, China

xiaojiaoch@163.com, ligx2022@gmail.com, hwanghao@gmail.com, zhou.wangjin.54r@st.kyoto-u.ac.jp, sheng.li@nict.go.jp, yang@i.kyoto-u.ac.jp, zhaoyi07@kuaishou.com

Abstract

This paper introduces our system submitted to Voice Privacy Challenge 2022. We adopted the following methods to improve the traditional methods. Firstly, the adversarial anonymization method was used, further hiding speaker information. Then, we extracted the embedding from the encoder of the transformer-based ASR systems because ASR has rich speaker information, so we do not have to train an individual speaker recognition/verification system for speaker embedding extraction. Experimental results prove that the proposed methods can be used for speaker anonymization tasks.

Index Terms: speech recognition, speaker anonymization, adversarial example, transformer

1. Introduction

With the rapid development of data mining, machine learning, deep learning, and the widespread application of web pages and mobile apps, privacy in processing and storing data has also attracted great attention. Although no clear privacy law is established, the security of speech data has received many concerns from researchers. A piece of speech not only conveys speech content information but also contains many personal identity information, e.g., gender, age, health status, emotion, and accent. One of the most prominent applications is the voice assistant, which authenticates the user's identity to log in and access many applications and accounts. While bringing convenience, applications also allow lawbreakers to get useful information. Therefore, different solutions have been proposed to protect the speaker's privacy, and one of the main approaches is speaker anonymization.

Speaker anonymization technology, also known as speaker de-identification, aims to suppress speaker identity information in the speech signal. Specifically, according to the VoicePrivacy 2022 Challenge [1], the speaker anonymity system needs to satisfy: (i) output a speech waveform, (ii) conceal the speaker identity, (iii) the linguistic content and paralinguistic attributes should be preserved, and (iv) ensure a one-to-one correspondence between speakers and pseudo-speakers.

Several approaches [2, 3, 4] have been proposed to protect speaker identity. Previous research focused on adding noise, speech synthesis, and voice conversion. By contrast, anonymization technology is capable of selectively preserving speech information. Fang et al. [4] proposed an anonymization method, which modified the x-vectors by selecting an x-vector from an x-vector pool as the pseudo-x-vector. This method is the first baseline system in the VoicePrivacy 2022 Challenge. However, this method needs a large speakers pool. Inspired by our previous work[5, 6, 7], this paper proposes two modifications to improve the x-vector-based baseline: (i) adding the adversarial noise and (ii) eliminating speaker information in a transformer-based ASR system.

The remainder of this paper is organized as follows. Section 2 describes our proposed methods in detail. Section 3 represents the experiment setup and experiment results. Conclusions and future plans are presented in Section 4.

2. Proposed Method

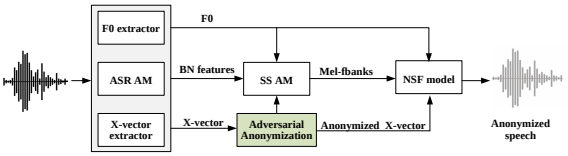

Figure 1: *The flowchart of the proposed adversarial anonymization method.*

This section discusses the proposed methods in which we modify the x-vector based on the baseline system [4]. Fig.1 and Fig. 2 illustrate the flowchart of the proposed methods.

The proposed methods and frameworks are mainly based on the first baseline 1 in VoicePrivacy 2022 Challenge. The framework is performed in three steps: feature extraction, X-vector anonymization, and speech synthesis. Feature extraction includes extraction of the speaker x-vector [8], which is based on the Time-delay neural network (TDNN) [9, 10], the fundamental frequency (F0), and bottleneck (BN) feature from the input speech waveform. The x-vector anonymization module is an essential part of the anonymization system. Moreover, it uses an external pool of speakers to anonymize the source-speaker x-vector. The speech synthesis uses an acoustic and neural waveform model to synthesize a speech waveform from the anonymized x-vector and the original BN and F0 features. In our approach, we modified these modules one by one. The detailed description is in the following subsections.

2.1. Proposed Adversarial Anonymization Method

The first approach is based on the concept of adversarial perturbation. Previous studies have shown DNN vulnerable to ad-

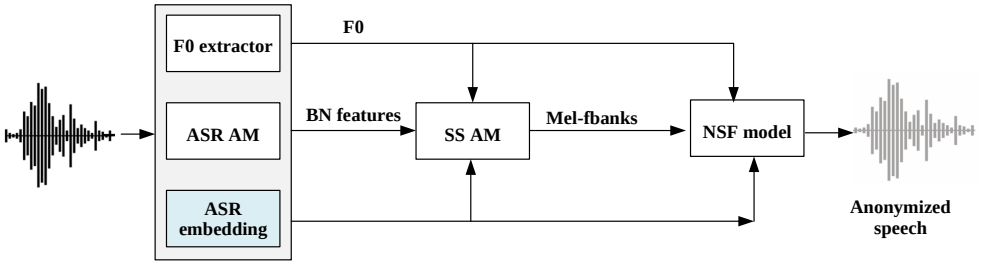

Figure 2: *The flowchart of the proposed eliminate Speaker Information by ASR System method.*

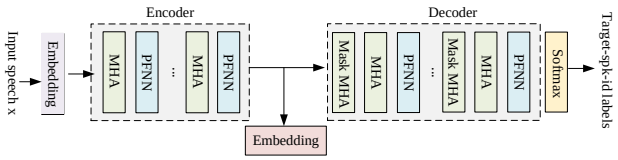

Figure 3: *Proposed method to extract embedding of target Speaker. The whole model is a speaker adaptive ASR, which includes multi-head self-attention (MHA), positional-encoding (PE), and position-wise feed-forward networks (PFFN).*

versarial perturbations [11, 12, 13, 14]. These works show that adding some small perturbations to the original input can mislead the classified system to get erroneous results. Our previous work [6, 7] can apply adversarial perturbations to TDNN models. The essence of the idea of adversarial perturbation is consistent with the idea that we want to modify the speaker anonymization method based on the x-vector. Therefore, we use the method of adding perturbation to anonymize the speaker. As shown in Fig.1, we proposed a new anonymization method based on adversarial perturbation.

The process of our proposed adversarial anonymization method can be formulated as follows:

$$Y_i = X_i + noise_{adv} \tag{1}$$

where the X_i denotes the original x-vectors of speaker i, and the anonymized x-vector of speaker i is Y_i. Considering the amount of computation required in the anonymization process, we borrow the method of non-targeted attack and add the same size of normally distributed tiny noise for each speaker. In other words, adding the adversarial noise ($noise_{adv}$) creates a fake speaker and hides the original speaker's identity.

2.2. Eliminate Speaker Information by ASR Systems

Figure 2 shows the second method of our anonymization systems. The detail information is described as follow:

ASR embedding extraction: Transformer-based seq2seq speech recognition architecture [15] generally includes an encoder and a decoder, where the encoder is responsible for encoding the input speech feature sequence. In [16, 5], it is shown that the output of the acoustic features by the encoder of the transformer can effectively show the classification characteristics of the speaker. Therefore, to some degree, the ASR embedding can represent speaker identity. And we replace the X-vector extractor in baseline with the transformer-based ASR system. Fig.3 shows the flowchart for extracting embedding.

3. Experiments

3.1. Datasets

All datasets used in this experiment were based on the VoicePrivacy 2022 Challenge[1]. As shown in Table 1, the train-clean-360 of Librispeech was used to train the anonymized automatic speaker verification (ASR_{eval}) and automatic speech recognition systems (ASV_{eval}). We also anonymized the train-clean-360 of Librispeech to train the ASR_{eval}^{anon} and ASV_{eval}^{anon}. Moreover, the development and test sets comprise Librispeech dev-clean and a subset of the VCTK corpus to evaluate ASR_{eval}, ASR_{eval}^{anon}, ASV_{eval} and ASV_{eval}^{anon}.

Table 1: *Number of speaker and utterances in the development and evaluation sets*

Dataset			Female	Male	Total
Train.	Librispeech-train-clean-360		430	482	921
Dev. &Eval.	Librispeech	Enrollment	15	14	29
		Trial	20	20	40
	VCTK	Enrollment	15	15	30
		Trial(different)			
		Trial(common)			

3.2. Experimental Setups

The main part of our experiment was conducted as same as the baseline 1.a in VoicePrivacy 2022 Challenge. In Subsection 2.2, we extracted the embedding from the encoder of the transformer-based ASR systems. We adopted the transformer-based speech recognition model (ASR$_{spk}$). In this paper, the ASR$_{spk}$ model required for embedding extraction is trained on the Librispeech train-clean-100 but based on the multitasking training method following [16, 17] with the speaker-id and label. The WER% was approximately 9.0%. The speaker-id was explicitly added as the label during training. The training labels are organized as "<SOS> <speaker-id> labels <EOS>". We extracted the encoder output of the ASR$_{spk}$ model as the speaker embedding.

For the evaluation, attackers were assumed to have access to the un-anonymized speech and anonymized speech utterances. Therefore, there are three attack scenarios:

Table 2: *Primary privacy evaluation: EER% achieved by ASV_{eval}^{anon} on data processed by Baseline, Model 1, or Model 2 vs. EER achieved by ASV_{eval} on the original (Orig.) unprocessed data*

Dataset	Gender	Weight	EER%			
			Orig.	Baseline	Model 1	Model 2
LibriSpeech-dev	fmale	0.25	8.67	17.76	30.40	20.45
	male	0.25	1.24	6.37	12.58	13.35
VCTK-dev(different)	fmale	0.20	2.86	12.46	23.98	12.97
	male	0.20	1.44	9.33	16.77	9.23
VCTK-dev(common)	fmale	0.05	2.62	13.95	25.00	11.05
	male	0.05	1.43	13.11	13.11	11.97
Weighted average dev			3.54	11.74	20.80	13.17
LibriSpeech-test	fmale	0.25	7.66	12.04	18.25	14.78
	male	0.25	1.11	8.91	20.04	11.14
VCTK-test(different)	fmale	0.20	7.66	12.04	24.85	17.18
	male	0.20	1.11	8.91	15.84	15.90
VCTK-test(common)	fmale	0.05	2.89	17.34	19.36	13.83
	male	0.05	1.13	9.89	17.23	11.58
Weighted average dev			3.79	11.81	19.54	14.07

Table 3: *Pitch correlation ρ^{F_0} and gain of voice distinctiveness $\mathbf{G_{VD}}$ achieved on data processed by Baseline, Model 1, or Model 2.*

Dataset	Gender	Weight	$\rho^{\mathbf{Fo}}$			$\mathbf{G_{VD}}$		
			Baseline	Model 1	Model 2	Baseline	Model 1	Model 2
LibriSpeech-dev	female	0.25	0.77	0.83	0.81	-9.15	-7.24	-12.93
	male	0.25	0.73	0.79	0.72	-8.94	-6.88	-11.47
VCTK-dev(different)	female	0.20	0.84	0.87	0.85	-8.82	-8.02	-9.65
	male	0.20	0.78	0.79	0.69	-12.61	-11.12	-11.08
VCTK-dev(common)	female	0.05	0.79	0.85	0.83	-7.56	-5.43	-6.82
	male	0.05	0.72	0.77	0.66	-10.37	-7.64	-8.05
Weighted average dev			0.77	0.82	0.77	-9.71	-8.01	-10.99
LibriSpeech-test	female	0.25	0.77	0.85	0.82	-10.04	-6.12	-12.17
	male	0.25	0.69	0.74	0.67	-9.01	-6.36	-10.79
VCTK-test(different)	female	0.20	0.84	0.87	0.85	-10.29	-9.56	-11.78
	male	0.20	0.79	0.80	0.69	-11.69	-10.43	-11.79
VCTK-test(common)	female	0.05	0.79	0.85	0.84	-9.31	-7.51	-10.57
	male	0.05	0.70	0.75	0.65	-10.43	-6.47	-8.88
Weighted average test			0.77	0.82	0.76	-10.15	-7.82	-11.43

Table 4: *WER% achieved by ASR_{eval}^{anon} on data processed by Baseline, model 1, or model 2 vs. WER achieved by ASR_{eval} on the original (Orig.) unprocessed data*

	Libri.		VCTK	
Anonymization system	Dev.	Test	Dev.	Test
Orig.	3.82	4.15	10.79	12.82
Baseline	4.34	4.75	11.54	12.82
Model 1	4.57	4.90	12.74	13.40
Model 2	5.39	5.60	14.49	15.00

- One or more anonymized trial utterances are exposed to the attacker;
- Original or anonymized enrollment utterances for each speaker are available to the attacker;
- Anonymized training data, which can retrain an ASV system, can be accessed by the attacker.

3.3. Results

We evaluated the anonymized speech on ASR_{eval}, ASR_{eval}^{anon}, ASV_{eval} and ASV_{eval}^{anon}. Table 2 shows the speaker verification performance. The equal error rate (EER) is the main objective metric. Model 1 is described in Subsection 2.1, which utilized adversarial anonymization, and model 1 is the primary model. Model 2 is described in Subsection 2.2 and is the contrastive model.

As shown in Table 2, the anonymized speaker has a lower performance on the same speaker. The robustness of Model 1 and 2 can be further improved. Model 1 performs better on the test datasets. Moreover, in Model 2, the performance on all development and test datasets approaches the baseline. Compared with the Baseline, a small calculation ability is needed for Models 1 and 2.

Table 3 shows the results of the pitch correlation ρ^{F_0} and gain of voice distinctiveness $\mathbf{G_{VD}}$ achieved on data processed by proposed methods and baseline. Model 1 performs well on the pitch correlation ρ^{F_0}, which is better than the baseline. Moreover, the pitch correlation of Model 2 almost equals the baseline scores. For the gain of voice distinctiveness $\mathbf{G_{VD}}$, proposed methods still have room for improvement.

Table 4 shows that the ability of the proposed anonymization system to preserve linguistic information is no less weak than the baseline system. The results show that the speech content after proposed anonymity has relatively complete preservation. Moreover, our proposed M2 system has simplified the pipeline of the baseline system.

4. Conclusions

In summary, we test two methods to protect speaker privacy. We use adversarial perturbation for speaker anonymization. Moreover, we extract speaker embedding from the End-to-End ASR system. Experimental results prove that both methods can be used for speaker anonymization tasks.

5. References

[1] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J. F. Bonastre, "The voiceprivacy 2022 challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[2] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5500–5504.

[3] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 82–94.

[4] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[5] D. Liu, L. Wang, S. Li, H. Li, C. Ding, J. Zhang, and J. Dang, "Exploring effective speech representation via asr for high-quality end-to-end multispeaker tts," in *International Conference on Neural Information Processing*. Springer, 2021, pp. 110–118.

[6] H. Zhang, S. Li, X. Ma, Y. Zhao, Y. Cao, and T. Kawahara, "Phantom in the opera: Effective adversarial music attack on keyword spotting systems," in *Proc. IEEE-SLT (demo)*, 2021.

[7] S. Li, J. Li, Q. Liu, and Z. Gong, "Adversarial speech generation and natural speech recovery for speech content protection," in *Proc. LREC (Language Resources and Evaluation Conference)*, 2022.

[8] D. Snyder and et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE-ICASSP*, 2018, pp. 5329–5333.

[9] A. Waibel and et al., "Phoneme recognition using time-delay neural networks," *IEEE/ACM Trans. ASLP*, vol. 37, no. 3, pp. 328–339, 1989.

[10] V. Peddinti and et al., "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[12] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," *in Proc. IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, 2018.

[13] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," *in Proc. 27th {USENIX} Security Symposium ({USENIX} Security 18)*, pp. 49–64, 2018.

[14] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," *arXiv preprint arXiv:1911.01840*, 2019.

[15] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE-ICASSP*, 2018, pp. 5884–5888.

[16] S. Li, R. Dabre, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation," in *Proc. INTERSPEECH*, 2019.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *CoRR abs/1706.03762*, 2017.