System Description for VoicePrivacy 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]NICT, Kyoto, Japan
[4]Kuaishou Technology, Beijing, China
Email: xiaojiaoch@163.com

2022-9-18

Introduction

Background

- With 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.
- Therefore, different solutions have been proposed to protect the speaker's privacy, and one of the main approaches is speaker anonymization.

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.

Previous work

- [2] 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.
- Inspired by our previous work[3, 4, 5], 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.

Proposed Method

This section discusses the proposed methods in which we modify the x-vector based on the baseline system [2].

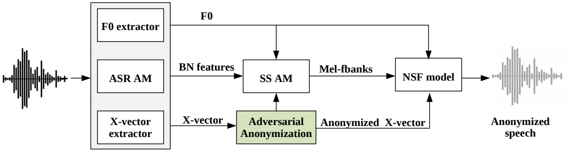

Figure 1: The flowchart of the proposed method (first approach).

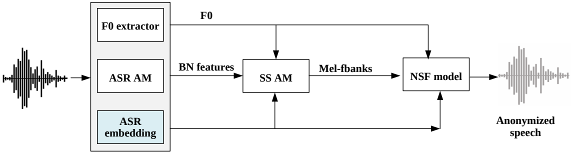

Figure 2: The flowchart of the proposed method (second approach).

The first approach is based on the concept of adversarial perturbation.

- 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:

$$Y_i = X_i + noise_{adv}$$

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. In other words, adding the adversarial noise ($noise_{adv}$) to create a fake speaker and hide the original speaker's identity.

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

- In [6, 3], 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.

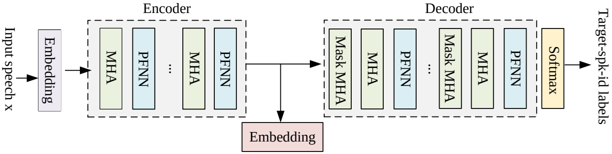

Figure 3: Proposed method to extract embedding of target Speaker

Experiments

Datasets

All datasets used in this experiment were based on the VoicePrivacy 2022 Challenge [1].

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)			

Experimental Setups

- The main part of our experiment was conducted as same as the baseline 1.a in VoicePrivacy 2022 Challenge;
- We adopted the transformer-based speech recognition model (ASR_{spk});
- The ASR_{spk} model required for embedding extraction is trained on the Librispeech train-clean-100 but based on the multitasking training method following [6, 7] with the speaker-id and label.

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

- 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.

Results

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

Results

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	$\rho^{\mathbf{F_0}}$			$\mathbf{G_{VD}}$		
			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

Table 4: WER(%) obtained by ASR_{eval} and $\mathrm{ASR}_{eval}^{anon}$ *model*

	Libri.		VCTK	
Anony. system	Dev.	Test	Dev.	Test
Ground Truth	3.82	4.15	10.79	12.82
Base.	4.34	4.75	11.54	12.82
Model 1	4.57	4.90	12.74	13.40
Model 2	4.61	4.79	12.15	12.86

- 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 baseline system.

Conclusions

- In summary, we test two methods to protect speaker privacy.
- 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.

References I

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.

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.

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.

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.

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.

References II

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.

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.