A Novel Method to Evaluate the Privacy Protection in Speaker Anonymization

Wei Liu[1], Jiakang Li[1], Chunyu Wei[1], Meng Sun[1](✉), Xiongwei Zhang[1], and Yongqiang Li[2]

[1] Lab of Intelligent Information Processing, Army Engineering University, Nanjing, China
sunmengccjs@163.com
[2] World Future Leaders Organization, Melbourne, Australia

Abstract. The technique to hide the real identity of speakers is called speaker anonymization. Aiming at deceiving automatic speaker verification (ASV) systems, speaker anonymization is usually conducted by modifying the temporal or spectral properties of original voices, e.g., by pitch scaling, by vocal tract length normalization (VTLN) or by voice conversion (VC). However, the real identity of anonymized speech can be recovered with a careful re-training of ASVs, e.g., data augmentation by anonymizing voices of the same speaker. In order to evaluate the effectiveness of speaker anonymization, a pre-restoration method for both enrollment and testing data is proposed, investigated and compared for the de-anonymization of anonymized voices. Experimental results show that the pre-restoration method is effective to speaker de-anonymization. Moreover, it is also found that the pre-restoration for testing data performs better than that for enrollment data, which would also be useful to other decision-making tasks involving enrollment and testing stages.

Keywords: Automatic speaker verification · Pre-restoration · Anonymization · De-anonymization

1 Introduction

With the rapid development of artificial intelligence and mobile computing, voice interface has been widely used in voice-based searching engine, smart phones/speakers, social communications and personal assistants. In order to improve the performance of voice interface, voice data is usually collected, transmitted and analyzed in clouds by service providers [1], e.g., speech enhancement [2], speech recognition [3], speech emotion recognition [4] and health detection from speech [5]. However, voice contains private information of the users, e.g., identity, health, emotion, etc.. The leakage of voice data would bring risks to the users. For example, recent advances of text-to-speech [6] and voice cloning [7] make it possible to easily synthesize the target speaker's voice given his/her small number of voices. The synthesized voices might be used for malicious purposes, together with the real identity of the speaker.

In order to protect privacy and promote security in voice communications, miscellaneous methods towards speaker anonymization or de-identification are proposed

X. Sun et al. (Eds.): ICAIS 2022, LNCS 13339, pp. 627–636, 2022.
https://doi.org/10.1007/978-3-031-06788-4_51

to hide the speaker's identity while remaining the intelligibility of voices, so that the anonymized voice is unlinkable to its real identity [8]. Those methods can generally be divided into physical and logical ones given the way how they confuse the identities [9]. Physical anonymization performs identity confusion by adding disturbance signals to the original waveform (e.g., adding noises), while logical anonymization [9] performs voice transformation (VT) on the original voices to change its temporal or spectral properties. However, physical anonymization also degrades the intelligibility of voices [10]. With the using of carefully designed algorithms to change the personality characteristics of voices, logical anonymization can yield highly intelligible anonymized voices [11]. Currently, the more commonly used anonymous algorithms include adjusting the pitch frequency and modifying spectral parameters at the signal processing level [12], and signal reconstruction at the voiceprint feature level [13].

For evaluating speaker anonymization in an objective way, equal error rates (EER) of automatic speaker verification (ASV) and word error rates (WER) of automatic speech recognition (ASR) are introduced to measure privacy protection and voice quality respectively [14]. Based on the prior knowledge an attacker has, three groups of de-anonymizations are investigated in [14]: an Ignorant case where the attacker knows nothing about the anonymization, an Informed case where the attacker knows both the anonymization algorithm and its parameters, and a Semi-Informed case where the attacker only knows the algorithm but does not know the specific parameters. The assumption of an Informed attacker is too strong to be practical. In this paper, we focus on our evaluation on the cases of Ignorant and Semi-Informed.

Data augmentation is utilized to enrich enrollment data in [14] with which the real identity of anonymized voices may be recovered. In [15], data augmentation is applied on testing data to recover the real identity of anonymized voices. In this paper, we unify the data augmentation in [14] and the pre-restoration in [15] for both enrollment and testing data. The experimental results demonstrate the usefulness of pre-restoration for de-anonymization. Furthermore, we find that the pre-restoration for testing data performs better than that for enrollment data, for which intuitive illustration is also provide. Given the great improvement of deep learning on speech processing [16], the backend classifiers usually take the variants of deep neural networks. Commonly used deep learning network structures include ResNet [17], VGG [18], RNN [19], etc.

The rest of this paper is organized as follows: In Sect. 2, three methods for voice anonymization are introduced using voice transformation. Section 3 describes the recipe of pre-restoration to identify anonymized voices. Experimental settings are presented in Sect. 4 and the results and discussions are given in Sect. 5. Section 6 is the conclusion of this paper.

2 Methods for Speaker Anonymization

A practical software for speaker anonymization should hold at least three properties: (a) Real-time processing to generate voices without latency, (b) Convenient usage without expert knowledge, (c) High quality of the generated voices. Therefore, three algorithms with open-source codes are chosen for conducting speaker anonymization: pitch scaling [20], vocal tract length normalization (VTLN) [21] and voice conversion (VC) [22].

However, VC seems not a good choice to speaker anonymization by failing to fulfill (c) as reported in [20]. Specifically, experimental results in [20] have proved that the voices generated by state-of-the-art VC algorithms performed poorly on ASR, which indicates that VC deteriorates the quality of voices. Let x and y are the original and anonymized voices respectively, the anonymization may be represented by a transform

$$y = f(x, \alpha) \tag{1}$$

where α is a group of parameters to characterize f.

2.1 Pitch Scaling

Pitch scaling increases or decreases the pitch by linearly compressing and stretching the speech spectrum. In phonetics, pitch is usually described by 12 semitones, which means that the pitch can be raised or lowered by up to 12 semitones [20]. Let p_0 and p denote the pitch values of the original and anonymized voices respectively, the processing of anonymization can be expressed by,

$$p = 2^{\alpha/12} p_0 \tag{2}$$

where α is a parameter measured in semitones with $\alpha \in [-11, 11]$ as analyzed in [13].

Because pitch scaling performs a linear transformation on the spectrum, it uses the same parameter to distort the value of the frequency spectrum for all the frames in an utterance. As has been vastly used in electronic devices, this algorithm suffices the properties (a) to (c) above. However, given its simplicity, the algorithm would be relatively weak on the anonymization of speakers, as has been studied for clean conditions in [11, 14] and will be investigated for noisy conditions in Sect. 4 of this paper.

2.2 Vocal Tract Length Normalization

Vocal tract length contains the personality characteristics of a speaker. VTLN adjusts the frequency axis of the spectrum through the warping function to change the position and bandwidth of formants, thereby hiding the personality characteristics [21]. Towards high quality of generated voices, VTLN first cuts an utterance into pseudo-periodic segments, and then applies a warping function to the Fourier indices of each segment. The warping function can be any mapping from [0, π] to [0, π]. In this paper, the warping functions are (3) from [21], and α is the parameter of each warping function.

2.3 Voice Conversion

Pitch scaling and VTLN are actually voice transform, where the generated voice does not target for a specific speaker, just for anonymization purpose. Recently, VC or deep fake of voices has been extensively studied for generating the voices of some target speaker. This technique can also be used for the anonymization of the source speaker. A representative work is the competitive baseline provided in VCC 2018 [23], which utilizes a vocoder to analyze speech and then synthesize speech with the converted spectral features. When treating VC as some kind of speaker anonymization, α represents its spectral mapping model, e.g., Gaussian mixture models or deep neural networks.

3 Evaluation of Speaker Anonymization

EER of ASV has been introduced as a numerical metric to evaluate the privacy protection in speaker anonymization, e.g., in [14] and [17]. A successful speaker anonymization algorithm should completely fool the ASV by yielding high EERs.

When computing EER, a list of trials is considered for evaluation, where each trial consists of a pair of utterances $\{x, y\}$ with x for enrollment and y for testing. A decision on if x and y are from the same speaker is made by introducing a threshold η,

$$d(g(x), g(y)) \leq \eta, \tag{3}$$

where $g(x)$ can be taken as the state-of-the-art speaker features, e.g., x-vector [24] and d can be some kind of negative similarity, e.g., negative scores of probabilistic linear discriminative analyses (PLDA) or cosine, which may be computed by some standard recipe in ASV [25].

Attackers can use various methods to improve the EER on anonymized voices, i.e., speaker de-anonymization. Pre-restoration on enrollment data has been studied for speaker de-anonymization on clean voices from LibriSpeech in [14]. In this paper, we contrastively study the pre-restoration for enrollment data and testing data.

3.1 Pre-restoration for Enrollment Voices

In the pre-restoration conducted on enrollment voices, the following steps are performed for each trial $\{x, y\}$ towards making a decision on if x and y are from the same speaker.

1) The enrollment utterance x is firstly transformed by algorithms of speaker anonymization, e.g. pitch scaling, VTLN or VC in Sect. 2 to yield many versions of x, say $\{x^{(1)},\ldots, x^{(r)},\ldots, x^{(R)}\}$, by choosing various α's for each anonymization function.
2) The testing utterance y and all the versions of x are fed into TDNN [26] to extract x-vectors, i.e., $g(x^{(r)})$ and $g(y)$, respectively.
3) The lowest distance $\widehat{d}$ is obtained and used for making a decision in (3), by ranging over r,

$$\hat{d} = \min_r d\left(g\left(x^{(r)}\right), g(y)\right). \tag{4}$$

3.2 Pre-restoration for Testing Voices

Being symmetric to the pre-restoration for enrollment data, the following steps are taken for pre-restoration for testing voices.

1) The testing voice y, possibly anonymized, is transformed by algorithms of speaker anonymization, e.g. pitch scaling, VTLN or VC in Sect. 2 to yield many versions of y, say $\{y^{(1)},\ldots,y^{(r)},\ldots,y^{(R)}\}$, by choosing various α's for each anonymization function.
2) The enrollment utterance x and all the versions of y are fed into TDNN to extract x-vectors, i.e., $g(x)$ and $g(y^{(r)})$, respectively.

3) The lowest distance $\widehat{d}$ is obtained and used for making a decision, by ranging over r,

$$\hat{d} = \min_r d\left(g(x), g\left(y^{(r)}\right)\right). \tag{5}$$

One may argue the motivation on the anonymization of anonymized voice y in step 1). This works as f with some α performs as an inverse of f with some other α, given the symmetric property of the warping function f w.r.t. α, e.g., for pitch scaling and VTLN in [20] and [21].

In the following experiments, only pitch scaling and VTLN are taken to perform (4) and (5) given its simplicity on usage. The de-anonymization is semi-informed when speaker anonymization shares the same function f with pre-restoration. Otherwise, it is ignorant. Both cases will be experimented and discussed in Sect. 4 and Sect. 5.

4 Experiments

4.1 Datasets and ASV Settings

Voxceleb1 and Voxceleb2 are taken as the datasets given their real-world noisy recording conditions, in contract to the clean datasets LibriSpeech used in [14]. State-of-the-art ASV system based on x-vector is chosen to evaluate speaker anonymization by following the recipe provided by Kaldi[1]. In our experiments, 21,820 segments of speech from 1,211 speakers of Voxceleb1(Train and Dev.) and 150,480 segments of 6,112 speakers from Voxceleb2 are used for training. 10,000 trials are randomly generated from 678 segments of 40 speakers from Voxceleb1(Test) for evaluation purpose. The EER of the trained ASV system on the 10,000 trials is as low as 2.06%.

The speaker anonymization using VC is evaluated on a clean dataset from VCC2018, where 280 pieces of speech from 8 speakers are taken as enrollment data, 11,200 pieces of converted voices are used as testing data. For the sake of briefness, 1,600 trials are randomly selected in the following experiments when evaluating VC.

4.2 Settings of Speaker Anonymization

In order to make the experimental results reproducible, open source software available online are used through all our experiments. Pitch scaling is conducted by SoundStretch from [27], which is able to modify the pitch, rate and tempo of an audio file. VTLN is performed by the open-source toolbox in [28], which provides four choices of nonlinear frequency warping functions, bilinear, quadratic, power and piecewise-linear. Algorithms for VC in VCC 2018 are not re-conducted but the source and converted samples are used directly in our experiments.

The anonymization parameters involved in pitch scaling and the four warping functions of VTLN are carefully tuned to ensure both intelligibility and anonymization of the generated voices. The discretized values of the parameters are shown in Table 1. These ranges also bound the r's in (4) and (5).

[1] https://github.com/kaldi-asr/kaldi.

Table 1. Proper ranges of anonymization parameters.

Anonymization function	Range	Step
Pitch scaling	[−11,11]	1
Bilinear	[−0.3, 0.3]	0.02
Quadratic	[−2, 2]	0.2
Power	[−0.5, 0.5]	0.05
Piecewise-linear	[0.5, 1.5]	0.05

5 Results and Discussion

The results are shown in Table 2. By comparing Table 2 and Table 1 of [14], several conclusions can be drawn.

5.1 Speaker Anonymization on Noisy Speech Still Works

As shown in the first row of Table 2, the EERs of anonymized voices are higher than 30%, which means speaker anonymization has bring considerable confusion to ASV on noisy speech from Voxceleb.

5.2 Pre-restoration for Testing Data Works Better Than that of Enrollment Data

By contrasting the first row and the second/third row of Table 2, it is seen that pre-restoration works as a means of speaker de-anonymization. Specifically, we find that the pre-restoration for testing data generally works better than that for enrollment data by comparing the left and right values of '/' in the second and third rows of Table 2. It is not straightforward given the symmetric configurations of (4) and (5). Intuitive analysis would be presented in the next subsection.

5.3 Semi-informed De-anonymization Outperforms the Ignorant Ones, as Expected

As explained at the end of Sect. 3, the bold figures in Table 2 correspond to the semi-informed cases, where the attackers know how the voices are anonymized, while the reaming figures correspond to the ignorant cases. It is clearly seen that the semi-informed de-anonymizations outperform the ignorant ones, as also reported in [14].

Table 2. EERs (%) of speaker de-anonymization by pre-restoration for enrollment/testing data.

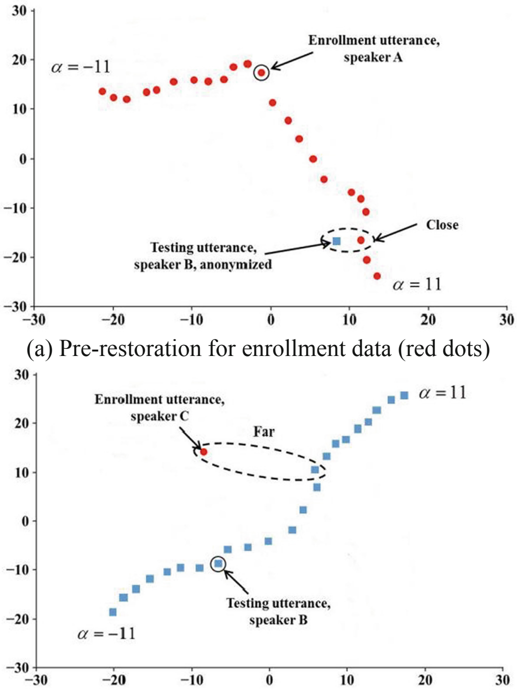
De-anonymization	Anonymization		
	Pitch scaling	VTLN	VC
None	39.00	34.34	44.17
Pitch scaling	**27.16/7.10**	33.48/21.54	42.41/43.53
VTLN	30.88/33.40	**25.84/18.54**	43.66

5.4 Visualization and Analysis

Towards a careful inspecting on the conclusion in Sect. 5.2, t-SNE is utilized to visualize the x-vectors of the enrollment and testing utterances for the speaker anon-ymization by pitch scaling. A wrong trial from pre-restoration for enrollment data is show in Fig. 1 (a), while its correct result obtained by pre-restoration for testing data is shown in Fig. 1 (b).

For a testing utterance anonymized by a large $|\alpha|$ (say $\alpha = 9$ in Fig. 1 (a), denoted by the blue square), an enrollment utterance from a different speaker (i.e., a non-target trial in ASV, denoted by the circled red dot in Fig. 1 (a)) may move close to the testing x-vector with the increasing of $|\alpha|$, which misleads ASV to wrong decisions, as shown in Fig. 1 (a). It is not strange that for some voice morphing softwares, severely modified voices sound similar even they are from different speakers. On the other hand, pre-restoration for a testing utterance (whether anonymized or not, denoted by the circled blue square in Fig. 1 (b) has little chance to get close to the x-vector of an enrollment voice from a non-target speaker without anonymization (denoted by the red dot in Fig. 1 (b), given the good performance of ASV on datasets without speaker anonymization. An example is shown in Fig. 1 (b).

In order to examine the analysis quantitively, the error rates per anonymization parameter α are shown in Fig. 2. It is observed that the poor performance of the pre-restoration for enrollment data comes from large $|\alpha|$'s, while the pre-restoration for testing data yields relatively smooth performance w.r.t. α. These numerical results are consistent with the visualizations in Fig. 1.



(a) Pre-restoration for enrollment data (red dots)

(b) Pre-restoration for testing data (blue squares)

Fig. 1. t-SNE visualization of x-vectors for a non-target trial. (a) illustrates how a wrong decision is made when pre-restoration for enrollment voices, while (b) shows the correct one by pre-restoration for testing voices. (Color figure onine)

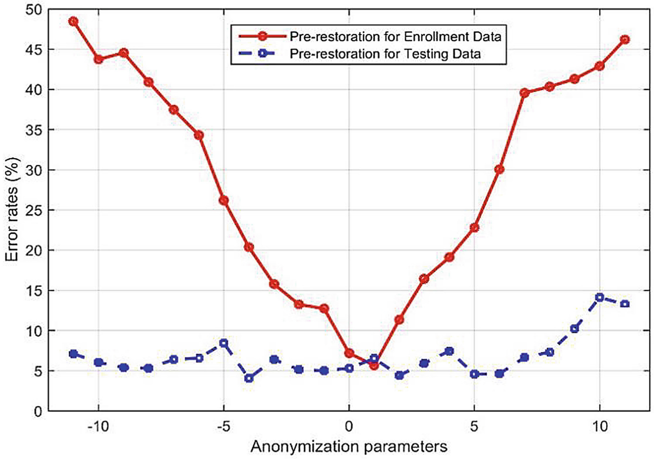


Fig. 2. Error rates (%) of speaker anonymization by pitch scaling w.r.t.α.

6 Conclusions

In this paper, pre-restoration for both enrollment and testing voices was studied and compared to evaluate speaker anonymization using pitch scaling, VTLN and voice conversion on a real-world noisy speech dataset. The results showed that pre-restoration for testing voices performed better than that for enrollment voices. Visualization and analysis explained the reasons for this outcome. Future work would be investigating more advanced methods of speaker anonymization to avoid attacks of speaker de-anonymization, e.g. improving the quality and speed of voice conversion to any target person.

Acknowledgement. The authors would like to thank Linlin Zheng and Yujun Wang for helpful discussions.

Funding Statement. This work was supported by the Natural Science Foundation of Jiangsu Province (BK20180080) and the National Natural Science Foundation of China (62071484).

References

1. Qian, J., Du, H., Hou, J., Chen, L., Jung, T., Li, X.: Hidebehind: enjoy voice input with voiceprint unclonability and anonymity. In: Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, pp. 82–94 (2018)
2. Zhou, L., Zhong, Q., Wang, T., Lu, S., Hu, H.: Speech enhancement via residual dense generative adversarial network. Comput. Syst. Sci. Eng. **38**, 279–289 (2021)
3. Nisar, S., Khan, M.A., Algarni, F., Wakeel, A., Uddin, M.I.: Speech recognition-based automated visual acuity testing with adaptive mel filter bank. Comput. Syst. Sci. Eng. **70**, 2991–3004 (2022)
4. Kwon, M.S.: 1D-CNN: speech emotion recognition system using a stacked network with dilated CNN features. Comput. Mater. Continua **67**, 4039–4059 (2021)
5. Lalitha, S., Gupta, D., Zakariah, M., Alotaibi, Y.A.: Mental illness disorder diagnosis using emotion variation detection from continuous English speech. Comput. Mater. Continua **69**, 3217–3238 (2021)
6. Székely, E., Henter, G.E., Beskow, J., Gustafson, J.: Spontaneous conversational speech synthesis from found data. In: Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 4435–4439 (2019)
7. Arik, S.O., Chen, J., Peng, K., Ping, W., Zhou, Y.: Neural voice cloning with a few samples. arXiv:1802.06006 (2018)
8. Gomez-Barrero, M., Galbally, J., Rathgeb, C., Busch, C.: General framework to evaluate unlinkability in biometric template protection systems. IEEE Trans. Inf. Forensics **13**(6), 1406–1420 (2017)
9. Fang, F., et al.: Speaker anonymization using x-vector and neural waveform models. In: Proceedings of 10th ISCA Speech Synthesis Workshop, pp. 155–160 (2019)
10. Hashimoto, K., Yamagishi, J., Echizen, I.: Privacy-preserving sound to degrade automatic speaker verification performance. In: Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5500–5504 (2016)
11. Jin, Q., Toth, A.R., Schultz, T., Black, A.W.: Speaker de-identification via voice transformation. In: Proceedings of 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 529–533 (2009)

12. Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., Evans, N.: Speaker anonymisation using the McAdams coefficient. In: Proceedings of Interspeech 2021, pp. 1099–1103. ISCA (2021)
13. Perero-Codosero, J.M., Espinoza-Cuadros, F.M., Hernández-Gómez, L.A.: X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. Comput. Speech Lang. **2022**, 10135 (2022)
14. Srivastava, B.M.L., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M., Vincent, E.: Evaluating voice conversion-based privacy protection against informed attackers. In: Proceedings of 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2802–2806 (2020)
15. Zheng, L., Li, J., Sun, M., Zhang, X., Zheng, T.F.: When automatic voice disguise meets automatic speaker verification. IEEE Trans. Inf. Forensics Secur. **16**, 823–837 (2021)
16. Changrampadi, M.H., Shahina, A., Narayanan, M.B., Khan, A.: End-to-end speech recognition of Tamil language. Intell. Autom. Soft Comput. **32**, 1309–1323 (2022)
17. Wu, Z., Shen, C., Den, A.V.: Hengel: wider or deeper: revisiting the ResNet model for visual recognition. Pattern Recognit. **90**, 119–133 (2019)
18. Mateen, M., Wen, J., Song, S.: Fundus image classification using VGG-19 architecture with PCA and SVD. Symmetry **11**(1), 1 (2018)
19. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D **404**, 132306 (2020)
20. Wang, Y., Wu, H., Huang, J.: Verification of hidden speaker behind transformation disguised voices. Digit. Signal Process. **45**, 84–95 (2015)
21. Sundermann, D., Ney, H.: VTLN-based voice conversion. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, pp. 556–559 (2003)
22. Kobayashi, K., Toda, T.: Sprocket: open-source voice conversion software. In: Proceedings of Odyssey 2018 The Speaker and Language Recognition Workshop, pp. 203–210 (2018)
23. Sprocket. https://github.com/k2kobayashi/sprocket
24. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 20–24 (2017)
25. Povey, D., et al.: The kaldi speech recognition toolkit. In: Proceedings of IEEE Workshop Automatic Speech Recognition and Understanding (ASRU), pp. 11–15 (2011)
26. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 6–10 (2015)
27. SoundTouch audio processing library. http://www.surina.net/soundtouch
28. Voice-Conversion. https://github.com/DenisStad/Voice-Conversion