System Description for Voice Privacy Challenge

Kyoto Team: Yaowei Han[1], Sheng Li[2], Yang Cao[3], Masatoshi Yoshikawa[3]

[13]Department of Social Informatics, Kyoto University, Kyoto, Japan
[2]National Institute of Information and Communications Technology, Kyoto, Japan
[1]yaowei@db.soc.i.kyoto-u.ac.jp,[2]sheng.li@nict.go.jp,[3]{yang,yoshikawa}@i.kyoto-u.ac.jp

1. System Description

Our system is based on Voice-Indistinguishablity, voiceprint perturbation mechanism, and privacy-preserving speech synthesis framework proposed in [1].

1.1. Voice-Indistinguishability

Voice-Indistinguishability is a rigorous privacy metrics for voiceprint (i.e., speaker identity) privacy by extending differential privacy [2] for privacy-preserving speech data release. We use the state-of-the-art representation of voiceprint, i.e., the x-vector [3].

We follow the definition proposed for a speech database but with an additional constraint for applying it to VoicePrivacy challenge: In the x-vector database, each x-vector refers to one speaker. That is, different x-vectors can be seen as different speaker identities.

Definition 1 (Speech data release under Voice-Ind) For every two neighboring x-vector databases $\mathcal{D}, \mathcal{D}'$, only differing in the ith x-vector, which are x, and x', a mechanism K satisfies ϵ-voice-indistinguishability if for all possible perturbed x-vector databases $\tilde{\mathcal{D}}$

$$\begin{aligned} \frac{\Pr(\tilde{\mathcal{D}}|\mathcal{D})}{\Pr(\tilde{\mathcal{D}}|\mathcal{D}')} &\leq e^{\epsilon d(\mathcal{D},\mathcal{D}')} \\ d(\mathcal{D},\mathcal{D}') &= d_{\mathcal{X}}(x,x') \\ d_{\mathcal{X}} &= \frac{\arccos(\cos\ similarity < x, x' >)}{\pi} \end{aligned} \tag{1}$$

where $\mathcal{X}$ is a set of possible voiceprints, $d_{\mathcal{X}}$ is the angular distance metric, cos $similarity$ is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.

Voice-Indistinguishability guarantees that given the output x-vector database $\tilde{\mathcal{D}}$, an attacker hardly distinguishes whether the original x-vector database is $\mathcal{D}$ or $\mathcal{D}'$ bounded by $\epsilon d_{\mathcal{X}}$. In other words, a lower $\epsilon d_{\mathcal{X}}$ indicates higher indistinguishability, hence a higher level of privacy. The privacy budget value ϵ globally influences the degree of guaranteed privacy.

1.2. Voiceprint Perturbation Mechanism

According to the definition and constraint that each x-vector refers to one speaker, we provide a perturbation mechanism satisfying Voice-Indistinguishability with the following two steps.

(1) X-vector database construction. Given a speech database, for each speaker with numerous utterances, we use one extracted x-vector (using the mean of x-vectors extracted from these utterances) to represent this speaker's identity permanently. Thus, we obtain an x-vector database, $\mathcal{D}$, where each x-vector refers to one speaker.

(2) Perturbation. Given an input x-vector $x_0 \in \mathcal{D}$, the mechanism K perturbs x_0 by randomly selecting an x-vector $\tilde{x}$ in the dataset $\mathcal{D}$ according to caliborated probability distributions, thus providing plausible deniability for x_0.

Theorem 1. A mechanism K that randomly transforms x_0 to $\tilde{x}$ where $x_0, \tilde{x} \in \mathcal{D}$ according to the following equation, satisfies voice-indistinguishability

$$\Pr(\tilde{x}|x_0) \propto e^{-\epsilon d_{\mathcal{X}}(x_0,\tilde{x})}$$

To satisfy the requirement "hide speaker identity as much as possible" [4], we set $\Pr(x_0|x_0) = 0$ to guarantee the minimum anonymization level.

For example, if our x-vector database has three items: A, B, C. The angular distance between A and B, between A and C are 1 and 3, respectively. Assume that $\epsilon = 1$, then we have $\Pr(B|A) \propto e^{-1}$; $\Pr(B|A) \propto e^{-3}$, thus $\Pr(A|A) = 0$, $\Pr(B|A) = 0.88$, and $\Pr(C|A) = 0.12$.

1.3. Privacy-preserving Speech synthesis framework

After obtaining the perturbed x-vector database, we should synthesize the perturbed x-vector and original speech characteristics without the original x-vector. In the perturbed x-vector database, each x-vector still refers to one speaker. Thus for utterances of one speaker, we use the same x-vector given by perturbed x-vector database so that we can naturally satisfy the fourth requirement. We use a privacy-preserving speech synthesis framework to synthesis the perturbed x-vector and original speech characteristics other than x-vector.

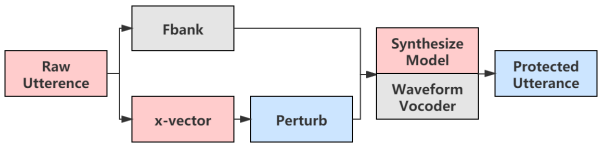

Figure 1: *Proposed System*

The privacy-preserving speech synthesis framework is shown in Figure 1. It uses two modules to generate the speech data: (1) an End-to-End acoustic model that produces a Mel-spectrogram (Mel-spec, used as a standard input feature by speech synthesis) [5, 6] given the two input features: filter-bank (Fbank, a commonly used feature for speech recognition) and x-vector [7], and has been proved robustness since adopted to neural network based speech recognition after 2011 [8]. Since our voice-indistinguishablity is defined on x-vector, the post-processing (e.g., speech synthesis) of the perturbed (protected) x-vector does not affect the defined privacy as long as the original x-vector is not used in the post-processing. Here we follow the similar setting using Fbank (lmfb) as acoustic model

input [9]. It is an interesting future study to investigate whether other types of features could improve the synthesized speech. (2) a waveform vocoder based on Griffin-Lim algorithm [10] that produces a speech waveform given the Mel-spectrogram after converting Mel-spectrogram to linear scale spectrogram using an inverse matrix. To protect voiceprint, after obtaining the x-vectors, we perturb them according to the voiceprint perturbation mechanism stated in section 1.2.

1.4. Privacy Guarantee

Privacy guarantee of the released private speech database. Figure 2 shows an example of the speech database before and after transformation using the proposed mechanism. Voice-indistinguishability guarantees that an attacker can hardly distinguish whether the original voiceprint is from A, B, or C.

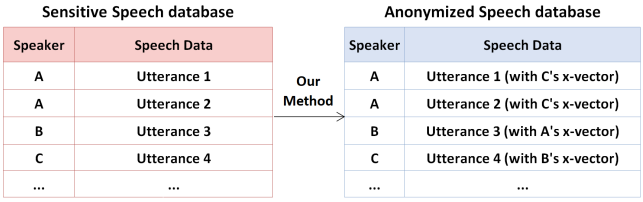

Figure 2: *Speech database before and after perturbation*

Privacy guarantee of voice-indistinguishability. We further explain the privacy guarantee provided by voice-indistinguishability by comparing the prior and posterior distributions of information obtained by an adversary. We prove that the prior and posterior distributions are bounded by $\epsilon d_{\mathcal{X}}$. In other words, voice-indistinguishability does not impose that an adversary gains no information but limits the increase of information that an adversary can obtain.

Let $\Pr(x)$ and $\Pr(x \mid \tilde{x})$ be the prior and posterior distributions of information obtained by an adversary, respectively, then for two indistinguishable x-vectors x, x':

$$\lg \frac{\Pr(\tilde{x}|x)}{\Pr(\tilde{x}|x')} = \lg \frac{\Pr(x \mid \tilde{x})}{\Pr(x' \mid \tilde{x})} - \lg \frac{\Pr(x)}{\Pr(x')} \leq \epsilon d_{\mathcal{X}}(x, x')$$

2. Results

The system is built using the End-to-End speech synthesis toolkit [11] on default settings[1] but trained using *train-clean-100*. For each "sensitive" database, we use the x-vector database constructed using itself.

Table 2 shows the architecture of the x-vector extractor. It consists of the context-aggregating time-delay neural network (TDNN) [12] layers operating at frame level (with the final context window of ± 7 frames), a statistics pooling layer which computes the mean and standard deviation of all the frames, effectively changing the variable-length sequence of frame-level activations into a fixed-length vector, and an utterance-level part consisting of two fully connected bottleneck layers which extract more sophisticated features and compress the information into a lower-dimensional space, and an additional softmax output layer.

Because ϵ represents our privacy budget, modified speech data with a larger ϵ has a weaker capacity for mitigating speaker verification attacks but better utility. After several experiments, we choose $\epsilon = 20$, which seems a good choice of balancing between privacy and utility.

[1] https://github.com/espnet/espnet/tree/master/egs/librispeech/tts1

2.1. ASV results

Results for the ASV objective evaluation using this system are provided in Table 1 for the development and evaluation datasets.

We can see that when the trial utterances are anonymized, the results for speaker verifiability metrics are significantly higher than the case when both the enrollment and trial utterances are original. When both the enrollment and trial utterances are anonymized, the results show evident speaker verifiability, which meets the fourth requirement.

Compared with baseline-1, although the speaker verifiability results when only the trial utterances are anonymized is a little bit worse, our system has significantly better results for meeting the fourth requirement. Compared with baseline-2, our system has overall better performance.

2.2. ASR results

Results for ASR evaluation using this system are presented in Table 3 in terms of WER. Compared with baseline-1 and baseline-2, our results for WER for *vctk_dev* and *vctk_dev* have similar even better performance. However, our results for WER for *libri_dev* and *libri_dev* don't perform very well. It seems that since our x-vector database are contructed using the "sensitive" dataset itself, it contains less x-vectors in *libri_dev* and *libri_dev* subset. An intuitive explanation is that the distances between the original x-vector and numerous possible transferred x-vectors are so big that it influences the WER results.

3. References

[1] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[2] C. Dwork and et al., "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*, 2006, pp. 265–284.

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

[4] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé *et al.*, "Introducing the voiceprivacy initiative," *arXiv preprint arXiv:2005.01387*, 2020.

[5] H. Kawahara and et al., "Re-structuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[6] K. Tokuda and et al., "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[7] V. Panayotov and et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*. ISCA, 2017, pp. 4006–4010.

[8] D. Povey and et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE-ASRU*, 2011.

[9] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Multi-speaker sequence-to-sequence speech synthesis for data augmentation in acoustic-to-word speech recognition," in *ICASSP 2019-2019*

#	Dev. set	EER, %	C_{llr}^{min}	C_{llr}	Enroll	Trial	Gen	Test set	EER, %	C_{llr}^{min}	C_{llr}
1	libri_dev	8.665	0.305	42.931	o	o	f	libri_test	7.664	0.184	26.799
2	libri_dev	43.470	0.944	148.637	o	a	f	libri_test	42.520	0.936	155.794
3	libri_dev	3.693	0.138	5.103	a	a	f	libri_test	0.730	0.025	2.380
4	libri_dev	1.242	0.035	14.275	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	42.080	0.913	148.332	o	a	m	libri_test	45.210	0.899	157.034
6	libri_dev	2.174	0.086	1.684	a	a	m	libri_test	4.232	0.133	4.774
7	vctk_dev_com	2.616	0.089	0.874	o	o	f	vctk_test_com	2.890	0.092	0.858
8	vctk_dev_com	47.380	0.966	159.616	o	a	f	vctk_test_com	50.000	0.996	170.682
9	vctk_dev_com	3.779	0.140	4.534	a	a	f	vctk_test_com	2.890	0.095	3.009
10	vctk_dev_com	1.425	0.051	1.565	o	o	m	vctk_test_com	1.130	0.035	1.029
11	vctk_dev_com	49.290	0.991	160.925	o	a	m	vctk_test_com	57.060	0.974	156.263
12	vctk_dev_com	4.843	0.185	5.409	a	a	m	vctk_test_com	5.650	0.202	7.388
13	vctk_dev_dif	2.920	0.102	1.152	o	o	f	vctk_test_dif	4.990	0.170	1.501
14	vctk_dev_dif	54.690	1.000	181.446	o	a	f	vctk_test_dif	60.390	1.000	171.734
15	vctk_dev_dif	4.323	0.166	2.366	a	a	f	vctk_test_dif	3.035	0.114	1.648
16	vctk_dev_dif	1.439	0.052	1.164	o	o	m	vctk_test_dif	2.067	0.071	1.819
17	vctk_dev_dif	45.010	0.979	138.723	o	a	m	vctk_test_dif	58.900	0.990	162.631
18	vctk_dev_dif	9.082	0.304	7.011	a	a	m	vctk_test_dif	6.028	0.225	4.345

Table 1: *ASV results for both development and test partitions (o-original, a-anonymized speech).*

Layers	Layer context	#context	#units
time-delay 1	$[t-2,\ t+2]$	5	512
time-delay 2	$\{t-2,\ t,\ t+2\}$	9	512
time-delay 3	$\{t-3,\ t,\ t+3\}$	15	512
time-delay 4	$\{t\}$	15	512
time-delay 5	$\{t\}$	15	1500
statistics pooling	$[0, T)$	T	3000
bottleneck 1	$\{0\}$	T	512
bottleneck 2	$\{0\}$	T	512
softmax	$\{0\}$	T	L

Table 2: *The x-vector TDNN. T is the number of frames in a given utterance. L is the number of speakers.*

#	Dev. set	WER, %		Data	Test set	WER, %	
		LM_s	LM_l			LM_s	LM_l
1	libri_dev	5.25	3.82	o	libri_test	5.55	4.15
2	libri_dev	16.52	14.72	a	libri_test	15.00	13.12
3	vctk_dev	14.04	10.79	o	vctk_test	16.39	12.81
4	vctk_dev	20.35	19.05	a	vctk_test	19.03	17.77

Table 3: *ASR results for both development and test partitions (o-original, a-anonymized speech).*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6161–6165.

[10] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. ASSP*, vol. 32, no. 2, pp. 236–243, 1984.

[11] S. Watanabe and et al., "Espnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

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