Poster: Voice-Indistinguishability – Protecting Voiceprint with Differential Privacy under an Untrusted Server

Yaowei Han
Department of Social Informatics
Kyoto University
Kyoto, Japan
yaowei@db.soc.i.kyoto-u.ac.jp

Yang Cao
Department of Social Informatics
Kyoto University
Kyoto, Japan
yang@i.kyoto-u.ac.jp

Sheng Li
National Institute of Information and Communications Technology
Kyoto, Japan
sheng.li@nict.go.jp

Qiang Ma
Department of Social Informatics
Kyoto University
Kyoto, Japan
qiang@i.kyoto-u.ac.jp

Masatoshi Yoshikawa
Department of Social Informatics
Kyoto University
Kyoto, Japan
yoshikawa@i.kyoto-u.ac.jp

ABSTRACT

With the rising adoption of advanced voice-based technology together with increasing consumer demand for smart devices, voice-controlled "virtual assistants" such as Apple's Siri and Google Assistant have been integrated into people's daily lives. However, privacy and security concerns may hinder the development of such voice-based applications since speech data contain the speaker's biometric identifier, i.e., voiceprint (as analogous to fingerprint). To alleviate privacy concerns in speech data collection, we propose a fast speech data de-identification system that allows a user to share her speech data with formal privacy guarantee to an untrusted server. Our open-sourced system can be easily integrated into other speech processing systems for collecting users' voice data in a privacy-preserving way. Experiments on public datasets verify the effectiveness and efficiency of the proposed system.

CCS CONCEPTS

• **Security and privacy → Data anonymization and sanitization**.

KEYWORDS

speaker de-identification; speech data collection; voiceprint; differential privacy

ACM Reference Format:
Yaowei Han, Yang Cao, Sheng Li, Qiang Ma, and Masatoshi Yoshikawa. 2020. Poster: Voice-Indistinguishability – Protecting Voiceprint with Differential Privacy under an Untrusted Server. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS '20), November 9–13, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3372297.3420025

CCS '20, November 9–13, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7089-9/20/11…$15.00
https://doi.org/10.1145/3372297.3420025

1 INTRODUCTION

With the rising adoption of advanced voice-based technology together with increasing consumer demand for smart devices, voice-controlled "virtual assistants" such as Apple's Siri and Google Assistant have been integrated into people's daily lives. A Statista's report shows that around 3.25 billion digital voice assistants were used around the world in 2019, and the number is even forecasted to reach 8 billion users by 2023, which is higher than the current world population [2].

However, privacy and security concerns may hinder the development of such voice-based applications since speech data contain the speaker's identifiable information, i.e., voiceprint (as analogous to fingerprint), which is a distinguishing and repeatable biometric feature of human beings [6]. Recent studies show that exposing an individual's voiceprint may cause security risks such as spoofing attacks [14] and reputation attacks [12]. With the advent of the GDPR [9] and increasing privacy concerns, the sharing of speech data is faced with significant challenges.

Our recent work [4] proposed the first formal voiceprint privacy definition called *Voice-Indistinguishability* and a privacy-preserving speech synthesis system that releases a private speech database with a trusted server. However, it remains challenging to achieve Voice-Indistinguishability without a trusted server (shown in Figure 1), where the user needs to perturb her speech utterances locally before sending it to an untrusted server such as virtual assistant service providers. A further challenge is that the service providers often need to authenticate the speaker (i.e., speaker recognition) before providing the service [1, 5].

This poster demonstrates a privacy-preserving speech synthesis system without a trusted server. As shown in Figure 1, the

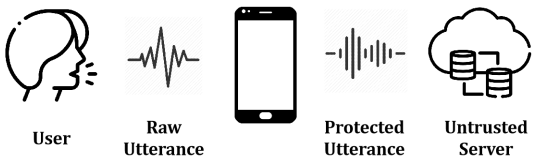

Figure 1: Voice-Indistinguishability in the local setting.

system runs locally on a user device that takes the user's raw utterances (e.g., voice commands for Siri) as inputs and outputs privacy-preserved utterances (with an anonymized voiceprint). To achieve both Voice-Indistinguishability and speaker recognition without a trusted server, we make three contributions in this poster. First, we propose a method to construct a "fake" voiceprint set (called *anonymous voiceprint set*) from publicly available speech datasets to represent the speaker's voice identity. We adopt x-vectors [11], which is the state-of-the-art model of the voiceprint, to construct the anonymous voiceprint set. Second, we design a new voiceprint anonymization method that guarantees Voice-Indistinguishability and speaker recognition. The idea is to randomly perturb a speaker's voiceprint to an anonymous voiceprint and then memorize the mapping between the original voiceprint and the anonymized voiceprint in a local "look-up table". The user can also reset her anonymized voiceprint by removing the entry from the look-up table. We also make sure that the system satisfies Voice-Indistinguishability even there are multiple users of the same device. Third, we propose a fast speech synthesis framework to synthesize the anonymized x-vector and other features in the original utterance without a trusted server. To reduce the workload of the user device, our system embeds perturbation in the local device and outsources the speech syntheses to an untrusted server.

2 METHODOLOGY

2.1 Voice-Indistinguishability

Voice-Indistinguishability is a rigorous privacy metrics for voiceprint [4]. We implement it using the state-of-the-art representation of voiceprint, i.e., the x-vector [11].

Definition 1 (Voice-Indistinguishability, i.e., Voice-Ind) [4] A mechanism K satisfies ϵ-Voice-Indistinguishability if for any output $\tilde{x}$ and any two possible voiceprints $x, x' \in \mathcal{X}$:

$$\frac{\Pr(\tilde{x}|x)}{\Pr(\tilde{x}|x')} \leq e^{\epsilon d_{\mathcal{X}}(x,x')}$$

$$d_{\mathcal{X}} = \frac{\arccos(\cos\ similarity < x, x' >)}{\pi}$$

where $\mathcal{X}$ is a set of possible voiceprints, $d_{\mathcal{X}}$ is the angular distance metric between x-vectors.

Voice-Indistinguishability guarantees that given the output x-vector $\tilde{x}$, an attacker hardly distinguishes whether the original x-vector is x or x' bounded by $\epsilon d_{\mathcal{X}}$. The privacy budget value ϵ globally influences the degree of guaranteed privacy.

2.2 Anonymous Voiceprint Set Construction

To achieve Voice-Indistinguishability in a local setting, the local device needs to obtain a set of possible x-vectors to anonymize a speaker's voice identity. We use public speech datasets to construct an x-vector database, $\mathcal{X}$. Given a public speech dataset such as the LibriSpeech [10], we extract x-vectors from utterances using the x-vector extractor, as shown in Table 1.

A challenge is that, the above extracted x-vectors cannot be directly used as the anonymous voiceprint set because a public speech dataset may contain multiple utterances of the same speaker, which causes problems for speaker recognition. To make sure that an x-vector in the anonymous voiceprint set represents a unique

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

voiceprint, a naive method can be using the mean of x-vectors from the same speaker; however, we find that such an averaged x-vector renders the synthesized utterance unnatural. To solve this problem, we design a clustering-based method to find the "representative" x-vectors in the public speech dataset. For the clustering algorithm, we use the elastic net subspace clustering (EnSC) [17] proposed for high dimensional data.

2.3 Voiceprint Anonymization Mechanism

Based on the anonymous x-vector dataset, we design an anonymization mechanism that satisfies Voice-Indistinguishability and enables speaker recognition. Given an input x-vector $x_0 \in \mathcal{X}$, the mechanism K perturbs x_0 by randomly selecting an x-vector $\tilde{x}$ in the dataset $\mathcal{X}$ according to calibrated probability distributions, thus providing plausible deniability for x_0. The perturbed x-vector serves as an anonymized voiceprint in our system.

Theorem 1 [4]. A mechanism K that randomly transforms x_0 to $\tilde{x}$ where $x_0, \tilde{x} \in \mathcal{X}$ according to the following equation, satisfies Voice-Indistinguishability.

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

In practice, the service providers often need to authenticate the speaker before providing the service; thus, we need to make sure the anonymized voiceprint can be recognized as the same speaker. To map a speaker's original x-vector to the same anonymous x-vector, our system memorizes such a mapping in a local "look-up table". Once we receive the raw utterance from a speaker, we firstly find whether it is in our "look-up table". If so, we use the anonymized x-vector that already in the look-up table without new perturbation; otherwise, we extract an x-vector from the utterance and randomly perturb it to an anonymous x-vector based on Theorem 1. We also need to remove the "used" x-vectors from the anonymous x-vectors set to make sure this anonymous x-vector will not be used by a different speaker who may use the same device. However, a potential privacy problem is that, the privacy guarantee diminishes along with the amount of the users (of the same device) because the available anonymized x-vectors will decrease. In [4], we show that a smaller size of $\mathcal{X}$ leads to weaker privacy. To solve this problem, we add a user's raw x-vector to the anonymous x-vector $\mathcal{X}$ set after "issuing" an anonymized x-vector for her. In this way, we always have the same size of $\mathcal{X}$.

2.4 System Architecture

After obtaining the anonymized x-vector, we synthesize the perturbed x-vector and other features in the original utterance and output a privacy-preserving utterance to the untrusted server. The privacy-preserving speech synthesis framework uses two modules

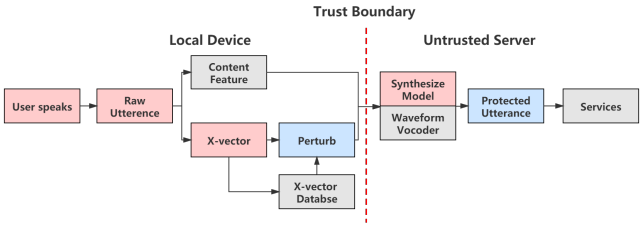

Figure 2: System architecture.

to generate the utterance: (1) an End-to-End acoustic model that generates a Mel-spectrogram (Mel-spec, used as a standard input feature by speech synthesis) [7, 13] given the two input features: content feature and x-vector. (2) a fast waveform generation model named parallel wavegan [16].

Our system architecture is shown in Figure 2. The local device transfers raw utterance into a protected one and then sends the protected utterance to an untrusted service such as a virtual assistant. To further reduce the workload of the user device, our system embeds perturbation in local devices and outsources the speech syntheses to an untrusted server.

3 EVALUATION AND DEMONSTRATION

For evaluation, we use the LibriSpeech [10] to construct the x-vector database with 500 x-vectors and use the VCTK [15] to verify the system. We employ automatic speaker verification (ASV) to show the privacy guarantee empirically and automatic speech recognition (ASR) to confirm the utility of the protected utterance.

3.1 Results

Results for the ASV objective evaluation are provided in Table 2.

#	Enroll	Trial	Gen	EER	C_{llr}^{min}	C_{llr}
1	o	o	f	2.616	0.089	0.874
2	o	a	f	47.380	0.966	159.616
3	a	a	f	3.779	0.140	4.534
4	o	o	m	1.425	0.051	1.565
5	o	a	m	49.290	0.991	160.925
6	a	a	m	4.843	0.185	5.409

Table 2: ASV results (o-original, a-anonymized speech).

When the trial utterances are anonymized, the speaker verifiability metrics, Equal Error Rate (EER) and log-likelihood-ratio cost function (Cllr), 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. It is because our algorithm always assigns the same anonymized voiceprint to the same speaker.

Table 3 shows ASR evaluation in terms of Word Error Rate (WER). Our approach with strong privacy guarantee outperforms or comes close to the performance of exiting methods [3, 8].

3.2 Demonstration Scenario

In the demonstration, we will show how the system processes a speaker's speech data with high utility and privacy. The attendees can interactively use the system with different privacy parameters and observe how it could affect privacy (e.g., ASV) and utility (e.g., ASR). Our source code is available in Github [1].

#	Data	WER(%)	
		LM_S	LM_l
1	original	14.04	10.79
2	[3]	18.92	15.38
3	[8]	30.10	25.56
4	Voice-Ind	20.35	19.05

Table 3: ASR results.

4 ACKNOWLEDGEMENT

This work is partially supported by JSPS KAKENHI Grant No. 17H06099, 18H04093, 19K20269, 19K24376, NICT tenure-track startup fund, and ROIS NII Open Collaborative Research 2020 (20FC06).

REFERENCES

[1] Andrew Boles and Paul Rad. 2017. Voice biometrics: Deep learning-based voiceprint authentication system. In *2017 12th System of Systems Engineering Conference (SoSE)*. IEEE, 1–6.
[2] Statista Research Department. 2019. Number of digital voice assistants in use worldwide from 2019 to 2023. https://www.statista.com/statistics/973815/worldwide-digital-voiceassistant-in-use/.
[3] F. Fang and et al. 2019. Speaker anonymization using X-vector and neural waveform models. In *Proc. 10th ISCA Speech Synthesis Workshop*. 155–160. https://doi.org/10.21437/SSW.2019-28
[4] Yaowei Han, Sheng Li, Yang Cao, Qiang Ma, and Masatoshi Yoshikawa. 2020. Voice-Indistinguishability: Protecting Voiceprint In Privacy-Preserving Speech Data Release. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
[5] Tencent Inc. 2015. The New WeChat Password. *https://blog.wechat.com/tag/voiceprint/* (2015).
[6] Anil Jain, Lin Hong, and Sharath Pankanti. 2000. Biometric identification. *Commun. ACM* 43, 2 (2000), 90–98.
[7] H. Kawahara and et al. 1999. 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* 27, 3–4 (1999), 187–207.
[8] Stephen Edward McAdams. 1985. Spectral fusion, spectral parsing and the formation of auditory images. (1985).
[9] Andreas Nautsch, Catherine Jasserand, Els Kindt, Massimiliano Todisco, Isabel Trancoso, and Nicholas Evans. 2019. The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding. *arXiv preprint arXiv:1907.03458* (2019).
[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5206–5210.
[11] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5329–5333.
[12] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.
[13] K. Tokuda and et al. 2013. Speech synthesis based on hidden Markov models. *Proc. IEEE* 101, 5 (2013), 1234–1252.
[14] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: A survey. *speech communication* 66 (2015), 130–153.
[15] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). (2019).
[16] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6199–6203.
[17] Chong You, Chun-Guang Li, Daniel P Robinson, and René Vidal. 2016. Oracle based active set algorithm for scalable elastic net subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3928–3937.

[1] https://github.com/iris0305/voice_ind