Privacy Measurement of Physical Attributes on Voice Anonymity

Shaohu Zhang
University of North Carolina at Pembroke
Pembroke, NC, USA
shaohu.zhang@uncp.edu

Zhouyu Li
North Carolina State University
Raleigh, NC, USA
zli85@ncsu.edu

Anupam Das
North Carolina State University
Raleigh, NC, USA
anupam.das@ncsu.edu

ABSTRACT

Various methods have been proposed for protecting the speaker's identity while preserving speech intelligibility. However, existing studies fail to consider the overall tradeoff between speech utility, speaker verification, and inference of voice physical attributes, such as emotion, age, accent, and gender. we propose a tradeoff metric to encapsulate voice biometrics as well as different voice attributes, to study the feasibility of applying cutting-edge voice anonymization solutions to achieve the optimum tradeoff between privacy protection and speech utility.

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

Voice anonymity; Privacy Measurement; Physical Attributes

ACM Reference Format:
Shaohu Zhang, Zhouyu Li, and Anupam Das. 2024. Privacy Measurement of Physical Attributes on Voice Anonymity . In *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '24), November 18–22, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3636534.3697448

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM MobiCom '24, November 18–22, 2024, Washington D.C., DC, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0489-5/24/11…$15.00
https://doi.org/10.1145/3636534.3697448

1 INTRODUCTION

Voice interfaces such as Siri, Alexa, and Google Assistant have become increasingly pervasive in our daily lives, offering numerous conveniences from music streaming to hands-free home appliance control. However, this convenience often results in security and privacy risks as voice data, including personally identifiable information like physical attributes, is often stored and processed by vendors to improve speech recognition engines and for commercial purposes.

The security issue intensifies with advances in voice cloning and speech synthesis technologies. With a few audio samples, it is possible to clone a victim's voice. Additionally, the privacy risk arises from the potential use of stored audio data for linkage attacks, where voiceprint uniqueness can be exploited to identify speakers from unlabeled speech data. Moreover, raw audio data can be used to infer age, gender, accent, and emotional state.

There has been extensive research on voice anonymization techniques, with initiatives like the VoicePrivacy challenge encouraging participants to design systems that preserve as much linguistic content while minimizing speaker identity. However, these techniques mainly focus on speaker recognition and speech recognition to evaluate privacy and utility tradeoffs, and do not consider how anonymization limits the inference of physical attributes. This study introduces *VoicePM* , a Voice Privacy Measurement tool for evaluating state-of-the-art voice anonymization solutions.

2 SYSTEM DESIGN

In a typical voice interaction system, the microphone records the audio input and uploads it to a cloud service maintained by the manufacturer or some third party for further processing. While speech-to-text is a typical processing that takes place, vendors have also been known to extract other forms of voice attributes (e.g., emotion, age, accent, and gender) for commercial purposes.

VoicePM bridges the communication between the user input, the cloud, and third-party apps. *VoicePM* accesses the raw audio, perturbs it, and produces sanitized audio via the anonymization engine. The sanitized audio is then sent to the

cloud, which provides automatic speech recognition (ASR) to send back the corresponding transcript. *VoicePM* can be integrated into the operating system and offer customizable controls to ensure input anonymity.

2.1 Speech Utility

As different voice anonymization systems might impact ASRs differently, we normalize the word error rate (WER) to perform a comparative analysis. Eq. 1 presents our used utility metric, where $WER_{baseline}$ is the WER for the original speech in a database and WER_{model} is for the anonymized speech. That is, U is equal to 1 for the original audio dataset while $U \in [0, 1)$ for the anonymized audio.

$$U = \frac{1 - WER_{model}}{1 - WER_{baseline}} \tag{1}$$

2.2 Speech Privacy

Speaker Verification. Speaker verification is the process of identifying a person from the characteristics of the voice. The Equal Error Rate (EER) is the rate at which a false reject rate equals a false acceptance rate to measure the optimum performance of the speaker verification system. Eq. 2 represents the normalized speaker verification accuracy.

$$S = \frac{EER_{model} - EER_{baseline}}{EER_{model}} \tag{2}$$

where $EER_{baseline}$ is the EER for the original database and EER_{model} is the overall EER between clean speech and sanitized speech generated by the anonymization model. Thus, theoretically, S is equal to 0 for the original audio dataset while $S \in (0, 1]$ is for the anonymization model.

Attributes Inference. Speaker identity is one of many potential paralinguistic attributes. In addition, voice attributes, including gender, age, accent, and emotion, are also important paralinguistic attributes.

Privacy Metric. We use Jaccard similarity to measure the similarity between two sets of voice attributes to see which attributes are shared among the two sets, as shown below.

$$J(A, A') = \frac{A \cap A'}{A \cup A'} \tag{3}$$

where A represents the set of voice attributes (i.e., gender, age, accent, and emotional state) of the original speaker, and A' represents the inferred voice attributes from the recorded audio. For simplicity, we assign equal weight to all attributes, but *VoicePM* can easily incorporate different weights for the different attributes when computing Jaccard similarity.

To compare the effectiveness of different voice anonymization techniques, we normalize the Jaccard index as shown in Eq. 4, where $J_{baseline}$ and J_{model} refers to the Jaccard index of the original and anonymized speech, respectively. That is, J is equal to 1 for the unaltered audio dataset while $J \in [0, 1)$ for the sanitized audio. A higher J means the adversary has more chance to infer the speaker's voice attributes.

$$J = \frac{J_{model}(A, A')}{J_{baseline}(A, A')} \tag{4}$$

We use the normalized EER (S from Eq. 2) and Jaccard index ratio (J from Eq. 4) to represent the privacy metric (P). As P monotonically increases with S and monotonically decreases with J, we use Eq. 5 to represent P. To this end, privacy accounts for both speaker verification and voice attribute inference as shown below:

$$P = \gamma S + (1 - \gamma)(1 - J) \tag{5}$$

where $\gamma \in (0, 1]$ and signifies to what extent we want to prioritize the individual components within P.

2.3 Privacy vs. Utility Tradeoff

For a given anonymization model, speech privacy increases (P) while the speech utility (U) decreases. Therefore, there exists an optimum tradeoff between privacy and utility. The ideal relationship between P and U forms an arc of an eclipse. For the original audio, $ERR_{model} = ERR_{baseline}$, $J = 1$, and $WER_{model} = WER_{baseline}$, so $P = 0$ and $U = 1$. As the anonymization model perturbs the audio signal, P increases ($\uparrow$) while U decreases ($\downarrow$). The ideal anonymization solution would be for both privacy and utility to be at their maximum possible levels. Therefore, there exists a point (U, P) where the P and U form a rectangle with the highest area ($P \times U$); we define this area measurement as the tradeoff between privacy and utility, which is represented by Eq. 6.

$$T(S, J, U) = P \times U = [\gamma S + (1 - \gamma)(1 - J)] \times U \tag{6}$$

where S, J and $U \in [0, 1]$, $\gamma \in (0, 1)$, and $T \in [0, 1]$. T equals 0 for the original speech, and higher values of T mean a better tradeoff of privacy and utility for a given voice anonymization technique.

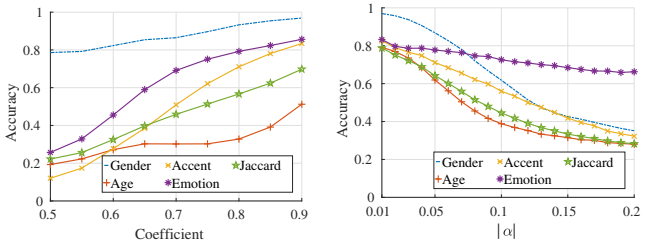
3 INFERENCE & ANONYMIZATION MODELS

Datasets. We use three datasets including Mozilla Common Voice (CV) English Corpus 10.0 (released on July 4, 2022), AISHELL-1 (Mandarin Chinese), and IEMOCAP (emotion dataset), to evaluate the feasibility and transferability of the inference attack models as well as the voice anonymization models. All datasets are resampled to 16kHz WAV files.

As shown in detail in Table 1, we follow the original accents categories labeled in the CV dataset and consider the top 10 accents with the highest number of speech utterances for our analysis. We adopt the `wav2vec2 + CTC` as our speech-to-text engine to perform transcription evaluation and a pre-trained `ECAPA-TDNN` model [3] for speaker verification.

Table 1: Common Voice English dataset summary.

Accents	Alias	# of samples	# of speakers	Length (hrs)
United States	US	10000	2683	13.78
England	EN	10000	1343	13.17
India and South Asia	INSA	10000	1450	13.26
Canadian	CA	10000	649	13.28
Australian	AU	10000	534	12.98
New Zealand	NZ	8514	138	10.80
Scottish	SC	7995	141	11.13
Ireland	IE	6052	164	7.93
Southern African	SA	5794	112	3.26
Chinese	CN	4887	285	10.74



(a) McAdams **(b) VoiceMask$_\alpha$.**

Figure 1: Inference of different voice attributes using different voice anonymization techniques.

Attribute Inference Models. We use the BASE `wav2vec2` model [2] to train and infer the physical attributes.We use the trained `wav2vec2` model on the IEMOCAP data to infer the emotional state as the *baseline*, which will be used to compare the emotional state after applying the voice anonymizers.

Voice Anonymization Models. We implemented five state-of-the-art privacy-preserving models. These models modify a source speaker's voice so that it sounds like another target speaker without changing the language contents. We consider four types of voice anonymization methods, including voice signal processing (SP) [6, 7], voice synthesis (VS) [5], voice conversion (VC) [4], and adversarial example [1].

4 RESULT AND CONCLUSION

Figure 1a and 1b show the accuracy of inferring gender, age, accent, and emotional state, along with the J metric. With the increase of the McAdams coefficient from 0.5 to 0.9, McAdams does not change gender inference significantly, while the inference accuracy for accent, age, and emotion varies significantly. Specifically, the gender inference accuracy changes from 78.60% to 96.89%, the accent inference increases from 12.02% to 83.41%, the age inference increases from 19.29% to 51.23%, and the emotional state inference rises from 25.69% to 85.61%. With the increase of the warping coefficient, VoiceMask$_\alpha$ does not significantly decrease the accuracy of emotional state inference (<20%). Fig. 2 plots the U, P, and T for all five models. For the signal processing-based approaches, McAdams (U = 0.86, P = 0.59, T = 0.50) slightly performs worse than VoiceMask. VoiceMask$_\alpha$ (U = 0.84, P = 0.67, T = 0.56) has overall better performance in U and P than VoiceMask$_\beta$ (U = 0.84, P = 0.64, T = 0.53). HiFi-GAN performs with the best tradeoff (T = 0.86) among all five anonymizers, followed by V-CLOAK with a tradeoff of 0.66. MaskCycleGAN preserves the highest privacy (P = 0.85), but its utility (U = 0.33) has the worst performance, resulting in the worst tradeoff (T = 0.28). In this study, we build and evaluate voice attribute inference models, including emotion, age, accent, and gender, We develop a novel voice privacy measurement tool *VoicePM* , to first explore and evaluate the tradeoff of privacy and utility for state-of-the-art voice anonymizers. Our experiments show that *VoicePM* can effectively measure the tradeoff of different anonymization models for a larger set of voice attributes. In the future, we plan to expand voice anonymizers and develop an open-source library for voice anonymity.

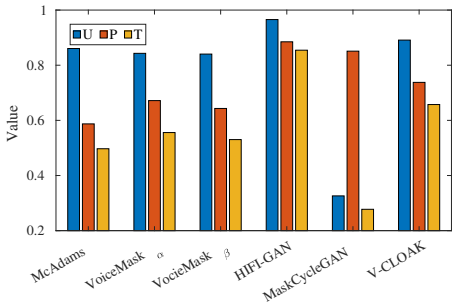


Figure 2: Overall performance for voice anonymizers.

REFERENCES

[1] 2023. V-Cloak: Intelligibility-, Naturalness- & Timbre-Preserving Real-Time Voice Anonymization. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security)*. USENIX Association, Anaheim, CA.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143* (2020).

[4] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2021. Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5919–5923.

[5] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 17022–17033.

[6] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans. 2021. Speaker Anonymisation Using the McAdams Coefficient. In *Interspeech 2021*. ISCA, 1099–1103.

[7] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. 2018. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 82–94.