VoicePM: A Robust Privacy Measurement on Voice Anonymity

Shaohu Zhang
North Carolina State University
Raleigh, NC, USA
szhang42@ncsu.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

Voice-based human-computer interaction has become pervasive in laptops, smartphones, home voice assistants, and Internet of Thing (IoT) devices. However, voice interaction comes with security and privacy risks. Numerous privacy-preserving measures have been proposed for hiding the speaker's identity while maintaining speech intelligibility. However, existing works do not consider the overall tradeoff between speech utility, speaker verification, and inference of voice attributes, including emotional state, age, accent, and gender. In this study, we first develop a tradeoff metric to capture voice biometrics as well as different voice attributes. We then propose *VoicePM*, a robust Voice Privacy Measurement framework, to study the feasibility of applying different state-of-the-art voice anonymization solutions to achieve the optimum tradeoff between privacy and utility. We conduct extensive experiments using anonymization approaches covering signal processing, voice synthesis, voice conversion, and adversarial techniques on three speech datasets that include both English and Chinese speakers to showcase the effectiveness and feasibility of *VoicePM*.

CCS CONCEPTS

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

KEYWORDS

Voice assistant; Voice anonymity; Privacy control

ACM Reference Format:
Shaohu Zhang, Zhouyu Li, and Anupam Das. 2023. VoicePM: A Robust Privacy Measurement on Voice Anonymity . In *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '23), May 29-June 1, 2023, Guildford, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3558482.3590175

1 INTRODUCTION

In recent years, voice interfaces (e.g., voice assistants) like Apple Siri, Amazon Alexa, and Google Assistant have become increasingly pervasive in our daily lives. Voice assistants (VAs) provide great conveniences, such as searching the web, listening to music, and hands-free control of home appliances. These VAs are not only integrated into laptops, smartphones, and smart speakers but also are prevalent in kid's toys, smart TVs, smart cars, and other appliances. Moreover, the Covid-19 pandemic accelerated the adoption of voice interfaces to avoid in-person interactions [70].

The convenient and pervasive usage of voice interfaces often comes at the cost of security and privacy risks. A voice interface usually sends the raw audio signal to the cloud for further processing, which can lead to the extraction of personally identifiable information (PII), such as physical attributes (e.g., emotion, gender, and accent). In fact, there have been numerous reports of vendors storing and processing users' voice data to improve the speech recognition engine [2, 3] as well as for financial gains [1, 74]. In addition, recent lawsuits claim voice data is being used to serve targeted ads [28, 66]. For example, both Apple [3], and Microsoft [2] store the audio clips generated when people use their voice-enabled products (i.e., Microsoft Cortana and Apple Siri) for up to two years, and Microsoft claims that it shares voice data with third parties [2]. Furthermore, McDonald's was reported to extract the voiceprint of a customer's speech to identify repeating customers at drive-throughs [4].

The security concern with voice data captured through voice interfaces is further exacerbated by the rapid development in voice cloning and speech synthesis technology, as with only a few audio samples from a victim, one can easily clone the victim's voice input [9, 72]. As voiceprint (i.e., voice-based biometric) is widely used in emerging authentication systems to unlock smart devices and activates voice assistants like Amazon Alexa and Google Assistant, recording and storing voice data imposes new attack vectors. The privacy concern stems from the possibility of stored audio data being used to conduct linkage attacks. While many service providers remove the IDs associated with the speech data to anonymize the collected speech data, as each person's voiceprint is unique, it is still feasible to identify the speakers of unlabeled speech data via speaker recognition/verification. For example, suppose one can collect speech samples of the target person from other sources. In that case, one can easily identify the records belonging to the target from the database, which can lead to an identity breach.

Another privacy risk with the collection of raw audio data is that it is possible to infer the age, gender, accent, and emotional state of a speaker from speech signals, which leads to building applications such as advertisements based on customer age, gender, and accent. Recently, Amazon filed a patent to determine the physical (e.g., accent, gender, age, etc.) and emotional characteristics of users based on voice audio input [27]. However, people usually have low awareness regarding what type of information can be inferred through voice data. For example, a recent study conducted a nationally representative survey in the UK (683 individuals, 18–69 years old) to investigate people's awareness of the inferential power of voice and speech analysis [36]. The results show that most participants have rarely or never thought about the possibility of personal information being inferred from speech

WiSec '23, May 29-June 1, 2023, Guildford, United Kingdom

ACM ISBN 978-1-4503-9859-6/23/05…$15.00
https://doi.org/10.1145/3558482.3590175

data, and only 18.7% of participants are at least somewhat aware that physical and mental health information can be inferred from voice recordings [36]. Lack of user awareness can lead to unwanted information leakage through voice interfaces. Therefore, there is a critical need for voice anonymization to preserve the privacy of recorded voice data.

In the last decade, there has been extensive research on voice anonymization techniques [7, 19, 49, 53, 67]. In the 2020 and 2022 VoicePrivacy challenge [64, 65], participants designed systems to anonymize the speaker's voice to hide the speaker's identity as much as possible while at the same time limiting the distortion of other speech characteristics to retain as much of the linguistic content as the original voice. By doing so, the solution with a higher speaker verification error rate (i.e., a better privacy guarantee) and a lower word error rate in speech content transcription (i.e., higher utility) would be the top candidates. However, the challenge is limited to only the tradeoff between speaker verification and speech intelligibility. Other inferrable attributes, such as physical attributes, are not thoroughly analyzed. For example, we applied our voice attribute inference on 50 original utterances from a fifty-year-old Irish (IE) woman, speaking with a neutral emotion (using Common Voice dataset [8]). The state-of-the-art inference models accurately identified the user's age, gender, emotional state, and accent. Privacy-preserving anonymization techniques such as McAdams [49] only hides the age information, while V-CLOAK [5] perturbs the gender as male and age as a senior but not accent. Thus, the majority of existing works [19, 49, 53] focus on speaker recognition/verification and speech recognition alone to evaluate the privacy and utility tradeoff of voice anonymity techniques; however, it is unclear how the anonymization solution limits the inference of physical attributes, including gender, age, accent, and emotion. In addition, there is no robust systematic tool to measure the tradeoff between privacy and utility for various anonymity schemes while considering attribute inference.

We propose, **VoicePM**, a robust **Voice** **P**rivacy **M**easurement on the state-of-the-art of voice anonymization solutions for a larger set of sensitive attributes, including the user's voice biometric and physical attributes. We make the following contributions:

- We propose a novel privacy measurement that leverages the speech utility, voice biometric, and physical attributes to systematically study the tradeoff for different voice anonymization solutions.
- We implement *VoicePM* and thoroughly evaluate it on three datasets (i.e., Common Voice [8], IEMOCAP emotion dataset [14], and AISHELL Chinese Mandarin dataset [13]) by applying five state-of-the-art anonymization models to highlight the tradeoff between speech utility, speaker verification, and physical attribute inference (i.e., emotion, age, accent, and gender). Existing works lack a comprehensive tradeoff analysis for different voice attributes. We have open-sourced our code base. [1]
- We perform a comprehensive feasibility analysis, studying the impact of various anonymization models, tradeoffs for different voice attributes, generalizability, and transferability across datasets. Our extensive experiments highlight *VoicePM* 's ability to better design a privacy mode for emerging voice interfaces.

[1] https://github.com/zhangshaohu/VoicePM

2 RELATED WORKS

Voice Synthesis. Spoofing attack on speaker verification systems has received considerable attention over the past decade. Studies have shown that voice-based authentication systems are vulnerable to impersonation [23, 24] and replay attacks [34, 69]. More powerful techniques include speech synthesis [35, 37, 46] and voice conversion [9, 29–31] techniques. Kumar et al. [37] proposed MelGAN, which adopted generative adversarial networks (GANs) to reliably generate high-quality coherent waveforms to enable real-time synthesis on the CPU. To achieve efficient and high-fidelity speech synthesis, HiFi-GAN [35] developed GANs with a discriminator consisting of small sub-discriminators, each of which obtains only specific periodic parts of raw waveforms. By doing so, HiFi-GAN generates samples with comparable speed and quality. MaskCycleGAN-VC [31] is a non-parallel Voice conversion (VC) technique that applies a temporal mask to the input Mel spectrogram and fills in the missing frames based on the surrounding frames to train voice converters without a parallel corpus. MaskCycleGAN-VC has shown its advantage in speech naturalness and speaker similarity compared with the state-of-art voice conversion approach such as CycleGAN-VC2 [29], and CycleGAN-VC3 [30].

Sensitive Data Inference. Voice recordings are typically a rich source of personally sensitive information. The voice signal contains linguistic and paralinguistic information, whereas the latter is rich with inferrable details such as age, gender, accent, body size, and health status [55]. Studies have shown that speech recordings can reveal an individual's gender with almost certainty [18, 38, 47] or be used to estimate a speaker's age [12, 25, 62]. Speech recordings can also be used to determine the speaker's spoken language [20] or accent [10], even to profile a person's health condition, feeling or emotional state [15, 51].

Privacy-preserving Voice Conversion. Many privacy-preserving approaches have been explored to protect various degrees of privacy to voice data [7, 42–44, 50, 53, 73]. Nautsch et al. [44] investigate the importance of developing privacy-preserving technologies to protect speech signals and highlight the importance of applying these technologies to protect speakers and speech characterization in recordings. Recent works have sought to protect speaker identity [49, 53], gender identity [7], and emotional state [6, 73].

Qian et al. implemented VoiceMask [53] on Android smartphones to convert voice based on vocal tract length normalization (VTLN). As a result, the speaker identification decreased to 16% based on 50 speakers while reducing voice input accuracy by no more than 14.2%. Patino et al. [50] use the McAdams coefficient to transform the spectral envelope of speech signals to level up the equal error rate (EER) of speaker verification as much as 30% while keeping the word error rate (WER) as low as 9%, compared with the WER of 8.4% in the original dataset. V-CLOAK [5] explores adding imperceptible noises to audio to generate adversarial examples so that the Automatic Speaker Verification (ASV) cannot recognize the speaker. Zhu et al.[73] designed an emotion privacy protection mechanism to filter users' emotions across multiple emotion states. Alofi et al. [7] adopted disentangled representation learning to prevent speaker verification implemented on different real-world datasets and show that the proposed approach can effectively defend against

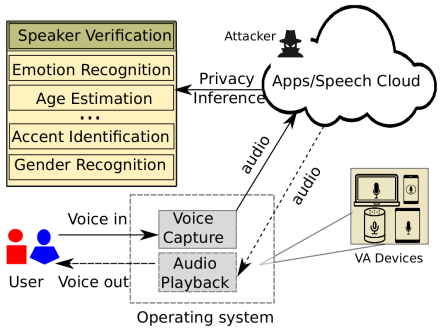


Figure 1: Typical data flow in a voice assistant.

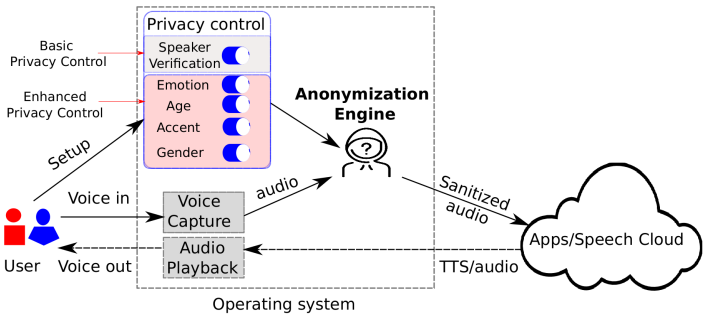


Figure 2: Our proposed privacy control for voice interfaces.

this inference attack, including gender recognition and emotion recognition.

Distinction with related work. Existing works are limited to analyzing only one or two voice-based attributes and lack a comprehensive tradeoff analysis for different voice anonymization techniques. Performing a comparative analysis of the different voice anonymization techniques and how their parameters impact the tradeoff between utility and privacy is critical for realistic deployment. In this paper, we not only compare various anonymization techniques but also perform a tradeoff analysis for a larger number of voice attributes (considering both voice biometrics as well as physical attributes). We believe our analysis will help shape the design of different *privacy settings* for voice interfaces.

3 DESIGN OF VOICEPM

In this section, we provide an overview of *VoicePM* and the threat model. We define speech utility and privacy, and formalize the tradeoff measurement model.

3.1 Overview

Fig. 1 shows the typical system architecture of a 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 [1, 74].

In Fig. 2, we present the privacy-preserving voice input scenario where *VoicePM* can be used. *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. The user can utilize *VoicePM* as an additional feature to adjust privacy settings for various applications. For example, trusted apps may have access to the authentic voice, while untrusted apps can only access sanitized voice data through privacy controls. *VoicePM* can be deployed either on the devices or from a reliable cloud service provider to prevent the cloud from collecting extra private information from the user's device.

3.2 Threat Model

As depicted in Fig.2, the anonymization step takes the user's voice input along with configuration parameters to produce an anonymized version. We assume the adversary accesses the anonymized audio whose speaker is unknown in dataset X' and the adversary has collected clean utterances of a pool of potential speakers in dataset X (i.e., from some auxiliary source) to train various inference algorithms. An adversary attempts to deanonymize a given anonymized test sample by inferring the speaker. For this, the adversary designs a linkage function that outputs a score for any utterance from X and X'. For instance, the attacker usually employs the cosine similarity score (shown in Eq. 1) to assess the similarity of speech representations between the target sample x_{target} and the test sample x'_{test}, determining if the utterances originate from the same speaker or not. In addition, the attacker has trained the voice attribute inference models based on the clean dataset X and then applies the inference models to identify the speaker's accent, emotion, age, gender, etc.

$$Score(x_{target}, x'_{test}) = \frac{x_{target} \cdot x'_{test}}{||x_{target}||||x'_{test}||} \tag{1}$$

3.3 Speech Utility

The utility of speech recognition systems is usually evaluated with word error rate (WER), which measures the differences between the transcription given by the ASR system and the ground truth as captured using the following function:

$$WER = \frac{N_{sub} + N_{del} + N_{ins}}{N_{ref}} \tag{2}$$

where N_{sub}, N_{del}, and N_{ins} are the numbers of substitution, deletion, and insertion errors of words, respectively. N_{ref} is the ground-truth number of words. We set WER to 1 if WER is greater than 1. As different voice anonymization systems might impact ASRs differently, we normalize the WER to perform a comparative analysis. Eq. 3 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.

$$U = \frac{1 - WER_{model}}{1 - WER_{baseline}} \tag{3}$$

3.4 Speech Privacy

Voice signals are a rich source of personal information which includes speaker identification and inferred voice attributes, including gender, age, accent, and emotional states.

3.4.1 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. 4 represents the normalized speaker verification accuracy.

$$S = \frac{EER_{model} - EER_{baseline}}{EER_{model}} \tag{4}$$

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.

3.4.2 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.

Gender. Sexual dimorphism in the vocal apparatus of male and female adults affects both the source and filter aspects of voice production [63]. Humans can easily identify and perceive the fundamental frequency (related to the perceived pitch). Adult males generally tend to have voices with a low fundamental frequency of phonation (F0) or low pitch, while adult females tend to have voices with a high F0 or high pitch. Researchers noted that the voice pitch of males and females, on average, is 100–200 Hz and 120–350 Hz, respectively [52, 63]. In addition, adult females have shorter vocal tract lengths, and their formant frequencies are 15% higher than adult males [57].

Age. The changes caused by age to voice are called in medical terms presbyphonia [39]. The vocal tract and its components are vital to producing sound, where the vocal folds in the larynx change continuously with age as a person grows from childhood to adulthood. The vocal folds are short at birth and grow through childhood and early adulthood. The pitches of children's voices are much higher than those of adults. From early adulthood until the age of about 55 years, voice pitch remains relatively constant, and then changes in senior as the tissue structure within the vocal tract begins to undergo deteriorating changes [57].

Accent. Language learning and exposure in early childhood cause the formation of tenacious speaking habits [57]. These habits determine how a person coordinates and moves their articulators during the speech and what sounds they are able to form or even perceive. The speaking habits of early childhood remain in a person's speech throughout their life unless they make extreme efforts to "unlearn" or mask them. These habits translate to specific patterns of signal characteristics known as accents which encode information about the speaker's nationality and geographical origin [57].

Emotion. Emotions affect the physiology of a person. The effects also extend to the speech production mechanism and the nervous system, including the brain. The process of phonation is affected by emotions, and it is possible to infer emotion through voice data [51].

3.4.3 Formalizing Privacy Metric. We use Jaccard similarity [45] to measure the similarity between two sets of voice attributes to see which attributes are shared among the two sets, as shown below.

$$J(A, A') = \frac{A \cap A'}{A \cup A'} \tag{5}$$

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 (such weights can come in the form of privacy settings selected by the user).

To compare the effectiveness of different voice anonymization techniques, we normalize the Jaccard index as shown in Eq. 6, 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.

$$J = \frac{J_{model}(A, A')}{J_{baseline}(A, A')} \tag{6}$$

We use the normalized EER (S from Eq. 4) and Jaccard index ratio (J from Eq. 6) to represent the privacy metric (P). As P monotonically increases with S and monotonically decreases with J, we use Eq. 7 to represent P. To this end, privacy accounts for both speaker verification and voice attribute inference as shown below:

$$P = \gamma S + (1 - \gamma)(1 - J) \tag{7}$$

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

3.5 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 [40] as shown in Fig. 3. 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 (↑) while U decreases (↓). 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. 8.

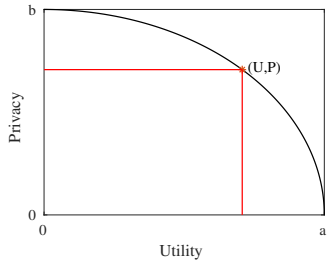


Figure 3: The typical relationship between privacy and utility.

$$T(S, J, U) = P \times U = [\gamma S + (1 - \gamma)(1 - J)] \times U \tag{8}$$

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.

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

4 INFERENCE & ANONYMIZATION MODELS

4.1 Datasets

We use three datasets 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.

Mozilla Common Voice (CV). The CV corpus [8] is a multilingual collection of transcribed speech, which employs crowdsourcing for data collection and validation. In this study, we use the CV English Corpus 10.0, released on July 4, 2022. The metadata has more than 1.59 million validated utterances, including fields such as client ID, audio transcription, upvotes, downvotes, age, gender, and accent. We first filter out audio files missing annotation for age, gender, and accent. Next, we drop samples with audio segments greater than 8 seconds to reduce the computation overhead and filter samples where the gender annotation is labeled as 'other'. 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. The only exception is we relabel the Singapore and Hong Kong accents as the Chinese (short for CN) accent as most people in Singapore and Hong Kong are descendent of southern Chinese. India and South Asia accents (INSA) include speakers from India, Pakistan, and Sri Lanka. Southern African accents (SA) consist of South Africa, Zimbabwe, and Namibia. We randomly select 10,000 utterances from the top five majorities of accents (i.e., US, EN, INSA, CA, and AU) while keeping all the utterances for the remaining five accents (i.e., NZ, SC, IE, SA, and CN). To this end, we reduce the dataset to 110.3 hours consisting of 83,242 samples (65,569 male utterances and 17,673 female utterances), as shown in detail in Table 1. We label six classes of ages, including teens (<20 years old), twenties (20-29), thirties(30-39), forties (40-49), fifties (50-59), and seniors ($\geq$ 60). We randomly split the dataset into the following portions 70:20:10 as train, validation, and test sets for the age, accent, and gender inference models (as described in Section 4.4).

IEMOCAP. The IEMOCAP corpus [14] comprises 5,531 utterances from 10 speakers (5 male and 5 female). The actors performed selected emotional scripts and improvised hypothetical scenarios designed to elicit nine specific forms of emotions (e.g., happiness, anger, sadness, and neutral state). In this study, we relabel excitement samples as happiness (similar to what existing work has done [51]) and used four emotional classes, including anger, happiness, sadness, and neutral, for our inference model. The dataset was randomly split into the following portions 80:10:10 as train, validation, and test set for the emotion inference model.

Table 2: Performance of different ASR systems.

Model	Source	Language	Dataset	WER(%)
wav2vec2+CTC	SpeechBrain	English	CV	14.50
CRDNN + CTC/Attention	SpeechBrain	English	CV	25.90
DeepSpeech	DeepSpeech	English	CV	27.09
Google Speech2Text	Google Cloud	English	CV	28.19
wav2vec2+CTC	SpeechBrain	English	IEMOCAP	24.57
CRDNN + CTC/Attention	SpeechBrain	English	IEMOCAP	37.15
Google Speech2Text	Google Cloud	English	IEMOCAP	37.76
wav2vec2+CTC	SpeechBrain	Mandarin Chinese	AISHELL1-test	5.04
Transformer	SpeechBrain	Mandarin Chinese	AISHELL1-test	6.04
Google speech2text	Google Cloud	Mandarin Chinese	AISHELL1-test	7.69

AISHELL-1 (Mandarin Chinese). The AISHELL-1 database [13] includes 400 people from different areas in China. The utterance was recorded in a quiet indoor environment using high fidelity microphone. We conduct the experiment using the default test set containing 7,176 utterances from 20 speakers with gender information. We use this dataset to evaluate model transferability across English and Chinese speakers.

4.2 Transcription Utility

ASR performance is assessed using the test set from all three datasets. As shown in Table 2, for the CV English dataset, we test four Speech-To-Text engines, including the SpeechBrain's [54] `wav2vec2 + CTC` and `CRDNN + CTC/Attention` models, DeepSpeech model [22], and the commercial Google Speech-To-Text engine using US English model. The `wav2vec2 + CTC` model [11] is trained on the audio of Librispeech (LS-960), which uses a pretrained wav2vec 2.0 model (wav2vec2-large-960h-lv60-self) combined with two DNN layers. The obtained final acoustic representation is given to the Connectionist Temporal Classification (CTC). The `wav2vec2 + CTC` model obtained the lowest WER of 14.50% among all the four tested ASRs, while Google Speech-To-Text performed the worst, with a WER of 28.19%. However, since the CV dataset contains samples with different accents, the WER is relatively higher. The IEMOCAP test set has a higher WER of 24.57% using the `wav2vec2 + CTC` model. We observe that emotional speech usually contains more modal particles, which confuses the ASRs. The AISHELL1-test set in Mandarin Chinese has a lower WER of 5.04% using the `wav2vec2 + CTC` model trained on the AISHELL1-train set as the utterance was recorded in a quiet indoor environment using high fidelity microphone. Overall, we observe that the `wav2vec2 + CTC` model has the lowest WER. Therefore, we adopt the `wav2vec2 + CTC` as our speech-to-text engine to perform transcription evaluation for all three datasets.

4.3 Speaker Verification

Current speaker verification systems (SVS) rely on a neural network to extract speaker representations. The `x-vector` [59] architecture is a Time Delay Neural Network (TDNN) that applies statistical pooling to project variable-length utterances into fixed-length speaker-characterizing embeddings. The `ECAPA-TDNN` architecture, subsequent improvement over the traditional TDNN model, outperforms state-of-the-art TDNN-based systems on the VoxCeleb [43] test sets and the 2019 VoxCeleb Speaker Recognition Challenge test sets [16]. We adopt a pre-trained `ECAPA-TDNN` model [17] for speaker verification from the SpeechBrain Library [15]. This model was trained using the Voxceleb1 [43], and Voxceleb2 [42] datasets.

Table 3: Attributes inference performance. The emotion inference model is trained and tested on IEMOCAP dataset while the other three models including age, accent, and gender are trained and tested on CV data.

Attributes	Test set (# of utterances)	wav2vec2 Base	ECAPA-TDNN
Emotion	happiness (167), anger (122) sadness (113), neutral (149)	77.31%	65.15%
Age	teens (876), twenties (2,799) thirties (1,703), forties (1,601) fifties (783), senior (563)	85.36%	80.95%
Accent	AU (969), NZ (872), CN (480), SA (609) INSA (1,006), CA (1,005), EN (1,013) IE (630), SC (797), US (944)	87.72%	82.10%
Gender	male (6,562), female (1,763)	99.06%	97.87%

Speaker verification itself is performed using cosine similarity between extracted speaker embeddings. We measure two metrics for speaker verification, including speaker verification accuracy and EER. Here, we adopt the default similarity score of 0.25 from the original model as the threshold to decide if two utterances are the same or not, while EER is the measurement with a threshold score when the false positive rate equals to false reject rate. For each dataset, we randomly generate 10,000 pairs of utterances (half pairs belong to the same speaker and half are not) to evaluate the speaker verification.

4.4 Attribute Inference Models

Inference model selection. Most recently, speech representation learning networks such as `wav2vec2` [11] and `ECAPA-TDNN` [17] have shown their advantage in speech recognition [71] and speaker recognition [68] over existing approaches like `i-vector` [21] and `x-vector` [59]. Furthermore, these representation learning models have also been applied in other domains, including speech anonymization [19], language detection [58], and emotion identification [41]. We, therefore, apply both `wav2vec2` [11] and `ECAPA-TDNN` [17] architecture to train the emotion, age, accent, and gender inference models. We use the BASE `wav2vec2` structure, in which the convolutional layer has a kernel size of 128 and 16 blocks. The model input dimension is 768 with the inner dimension 3,072 [11]. We train 60 epochs for each `wav2vec2` model with a batch size of 32. The `ECAPA-TDNN` architecture consists of blocks of TDNNs and Squeeze-and-Excite (SE) layers unified with blocks of Res2Block layers, and each convolutional frame layer has 512 channels. We train 300 epochs for each `ECAPA-TDNN` model with a batch size of 32.

Table 3 shows that the accuracy of the `wav2vec2` model is 77.31%, 85.36%, and 87.94% and 99.06% (female 98.30%, male 99.27%), for emotion, age, accent, gender prediction, respectively. In contrast, the corresponding accuracy of the `ECAPA-TDNN` model is 65.15%, 80.95%, 82.10%, and 97.87% (female 96.60%, male 98.25%) for emotion, age, accent, gender prediction, respectively. Thus, the `wav2vec2` model shows its overall advantage. However, the `wav2vec2` model has around 90.2 million trainable parameters, whereas the `ECAPA-TDNN` only has around 5.5 million trainable parameters. Thus, the `wav2vec2` model requires significantly more computing resources than the `ECAPA-TDNN` model. The inference accuracy for age and gender is slightly better (<5%) for `wav2vec2` compared to `ECAPA-TDNN`, while the inference accuracy of emotion and accent is significantly different (>5%). Therefore to reduce the computing burden while maintaining reasonable accuracy for inferring emotion and accent, we use the `wav2vec2` model, while for gender and age prediction, we use the `ECAPA-TDNN` model for our evaluation.

Emotion labeling for the CV Dataset. Due to the lack of emotional states in the CV dataset, 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 anonymization models. Even though the emotional inference model is imperfect, it will still enable us to understand the general trend after applying the different voice anonymization techniques to determine the optimal operating region.

Table 4: State-of-the-art of voice anonymization models.

Model	Type	Pre-training	Overhead
McAdams [49]	Signal processing	×	Low
VoiceMask [53]	Signal processing	×	Low
HiFi-GAN [35]	Voice synthesis	×	High
MaskCycleGAN [31]	Voice conversion	✓	Medium
V-CLOAK [5]	Voice adversarial example	✓	Low

4.5 Voice Anonymization Models

To protect the identity of the users of voice input, we implemented five state-of-the-art privacy-preserving models shown in Table 4, many of which have been used as baselines in the 2022 Voice Privacy Challenge [64]. 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) [49, 53], voice synthesis (VS) [35], voice conversion (VC) [31], and adversarial example [5]. McAdams [49], and VoiceMask [53] are SP-based methods that directly apply signal processing techniques to modify speaker-related features in the audio signals to obscure voiceprints. HiFi-GAN [35] is a VS-based GAN network that converts the transcript to a target speaker's voice. MaskCycleGAN-VC [31] translates one voice into another target. V-CLOAK [5] transfers the speaker's voice to an adversarial voice example.

McAdams. McAdams coefficient [49] can be used to adjust the frequency of each harmonic, as shown below:

$$X(t) = \sum_{k=1}^{K} A_k(t) \cos\left(2\pi (k f_0)^{\alpha} + \phi_k\right) \tag{9}$$

where k is the harmonic index, $A_k(t)$ is signal amplitude, ϕ_k is the phase, and α is the McAdams coefficient, which is usually in the range of [0.5, 1]. The speech loses intelligibility when $\alpha < 0.5$, while the anonymization is significantly low when $\alpha > 1$. Adjustments to the distribution of harmonics act to modify the resulting timbre.

VoiceMask. VoiceMask [53] is a Vocal Tract Length Normalization (VTLN) based approach that aims to compensate for the effects of different vocal tract lengths by warping the frequency spectrum in the filter bank analysis. Given a source utterance, VTLN-based voice conversion processes it in six steps: pitch marking, frame segmentation, FFT (fast Fourier transform) to the frequency domain, VTLN, IFFT (inverse fast Fourier transform) to the time domain, PSOLA (pitch-synchronous overlap and add). Pitch marking and frame segmentation aim to split the speech signal into frames that

match the pseudo-periodicity of the voiced sounds as determined by the fundamental frequency of the voice to output a synthetic voice with the best audio quality. VTLN modifies the spectrum of each frame using frequency warping, that is, stretching or compressing the spectrum with respect to the frequency axis according to a warping function. VoiceMask implemented two popular warping functions including bilinear function [53, 61] (noted as VoiceMask$_\alpha$) and quadratic function [61] (noted as VoiceMask$_\beta$). The bilinear warping function is shown below:

$$\varphi_\alpha = \omega + 2\arctan^{-1}\left(\frac{(1-\alpha)\sin\omega}{1-(1-\alpha)\cos\omega}\right) \tag{10}$$

where $\omega \in [0, \pi]$ is the normalized frequency, and $\alpha \in (-1, 1)$ is a warping factor used to tune the strength of voice conversion. The quadratic function [61] is represented by the equation given below:

$$\varphi_\beta = \omega + \beta\left(\frac{\omega}{\pi} - \left(\frac{\omega}{\pi}\right)^2\right) \tag{11}$$

where $\beta \in (-1, 1)$. To design a mechanism that ensures the attacker cannot reverse or reduce voice conversion, we propose randomizing the warping coefficients α and β within a specific range.

HiFi-GAN. HiFi-GAN [35] is a generative adversarial network (GAN) to synthesize high-fidelity waveforms from Mel-spectrograms. First, the Tacotron2 [56] model is used to generate the Mel-spectrogram based on the speech text, then a HiFI-GAN vocoder trained with LJSpeech [26] takes the Mel-spectrogram and produces a waveform in output. As most VAs have a female voice by default, we convert all the utterances using a female vocoder.

MaskCycleGAN-VC. Voice conversion (VC) is a technique for translating one voice into another without changing the linguistic content and has been extensively studied. MaskCycleGAN-VC [31] is a non-parallel VC technique that applies a temporal mask to the input mel-spectrogram and fills in the missing frames based on the surrounding frames to train voice converters without a parallel corpus.

We follow the default training setting of MaskCycleGAN-VC [31] to train the VC models. We randomly selected 20 speakers from the CV data, with 240 utterances for each speaker. We used combinations of 10 pairs of source-target for the evaluation. For each speaker, we used 80 sentences for training and 140 sentences for the evaluation. Note that the training set contains no overlapping utterances between the source and target speakers, and the pair between the source and target speakers have different gender, age, and accent.

V-CLOAK: V-CLOAK [5] is a generative model-based anonymizer based on Wave-U-Net [60]. By adding imperceptible noises to audio, V-CLOAK generates adversarial examples so that an ASV system cannot recognize the speaker. V-CLOAK optimizes the intelligibility, naturalness, and timbre of the audio without retraining the anonymizer. This means V-CLOAK can be trained on data from one language and be applied to samples from another language without losing much intelligibility. V-CLOAK was trained on train-clean-100, and train-other-500 datasets in LibriSpeech [48] with the adding noises level of 0.1. We trained 80 epochs with a batch size of 16 to achieve 16.3% WER on the test-clean of the LibriSpeech dataset, which is close to the 13.92% reported in the paper [5]. We use this trained model to transfer our test data in the evaluation.

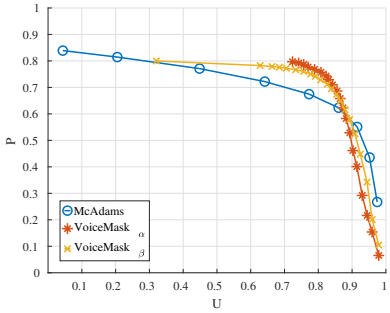


Figure 4: Privacy and utility tradeoff relationship for different voice anonymization models.

5 EVALUATION

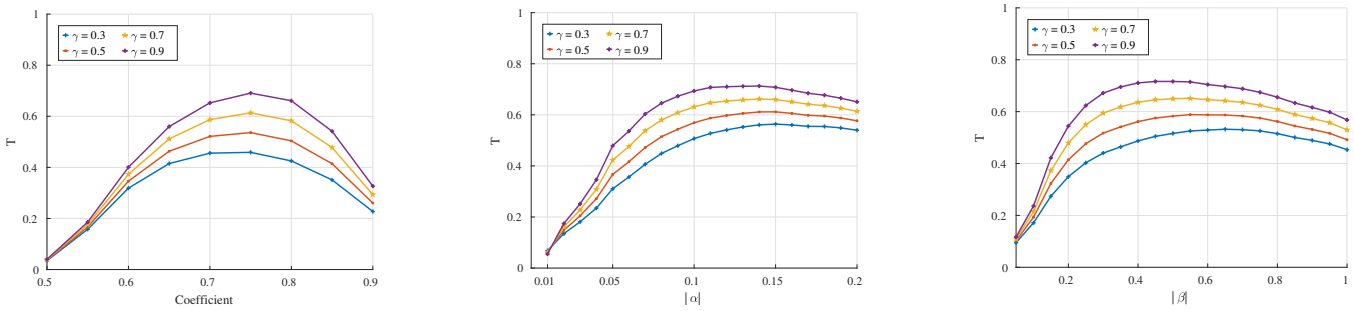
In this section, we comprehensively analyze *VoicePM* under various settings to evaluate how well *VoicePM* meets three design objectives: **preserve transcription utility**, **hinder speakers' voice biometrics**, and **thwart voice attribute inference**. First, we empirically validate the privacy and utility tradeoff model in §5.1. We then evaluate how the tradeoff model parameter (γ) impacts the optimum tradeoff point in §5.2. Next, we analyze the attribute inference accuracy for different models in §5.3 and determine the optimum coefficients for voice anonymization models in §5.4. We evaluate the impact of stochastic anonymization in §5.5 and provide the overall performance in §5.6. We also evaluate the run-time of the different voice anonymization models in §5.7. Finally, we examine the sensitivity of our system by analyzing the impact of the following factors: generalizability across English datasets (§5.8), and transferability between English and Chinese speakers (§5.9). We use the Common Voice (CV) dataset by default unless mentioned otherwise.

5.1 Privacy vs. Utility Tradeoff Relationship

In this section, we determine how our proposed privacy and utility metrics vary to justify our choice of the tradeoff metric (Eq. 8) as we discussed in Sec. 3.4. We plot the privacy (P) and utility (U) metrics for McAdams, VoiceMask$_\alpha$, and VoiceMask$_\beta$ with varying coefficients in Fig.4. The figure demonstrates that P and U approximately form a non-linear pattern like an arc, as shown in Fig. 3. This also implies that there exists a point (U, P) on the arc where the tradeoff is optimum.

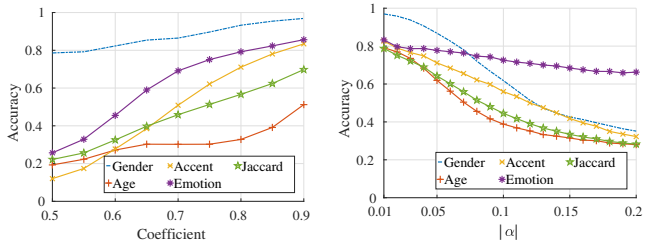
5.2 Determining the Impact of γ on Tradeoff

The weight of S and J is determined by the γ value, and we examine how it affects the optimal tradeoff in this section. As shown in Eq. 7, we calculate the privacy (P) metric as the weighted average of S and J. To determine the proper value of γ, we plot the tradeoff metric (T) for McAdams, VoiceMask$_\alpha$, and VoiceMask$_\beta$ for varying different coefficient value for each model as shown in Fig. 5. The plots show that the tradeoff metric increases with the value of γ. However, the warping coefficient for the three models' optimum tradeoff stays mostly the same when γ equals 0.3, 0.5, 0.7, and 0.9. Specifically, McAdams has an optimum tradeoff with a coefficient of 0.75. VTLN linear warping (VoiceMask$_\alpha$) has an optimum tradeoff with a warping coefficient of 0.14, while VTLN quadratic warping (VoiceMask$_\beta$) has a peak tradeoff with a warping coefficient of



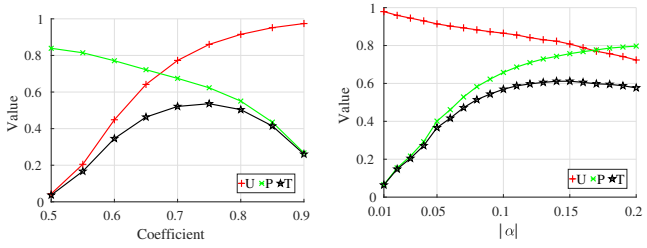
(a) McAdams (b) VoiceMask$_\alpha$ (c) VoiceMask$_\beta$

Figure 5: Tradeoff measurement with varying γ. The warping coefficient for the three models' optimum tradeoff stays mostly the same.



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

Figure 6: Inference of different voice attributes using different voice anonymization techniques.



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

Figure 7: Distribution of U, P, and T with varying coefficient for McAdams and VoiceMask approach.

around 0.55. We see the optimum point does not change much when $\gamma \in [0.3, 0.9]$. In this study, we consider that S and J have the same weight and, thus, pick $\gamma = 0.5$ for the following evaluation in the paper.

5.3 Voice Attribute Inference

We now determine to what extent the inference capabilities are impacted by varying the parameters of McAdams and VoiceMask models. Figure 6a and 6b 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 VTLN warping coefficient, VoiceMask$_\alpha$ does not significantly decrease the accuracy of emotional state inference (<20%). We also see that J consistently follows the overall change in inference accuracy. From a privacy point of view, a smaller value of J is desirable.

5.4 Tradeoff and Optimum Coefficient

In this section, we determine if our proposed tradeoff metric can obtain an optimum co-efficient consistent with existing results. Fig. 7a and 7b shows the plot between U, P and T for McAdams and VoiceMask$_\alpha$ (VoiceMask$_\beta$ shows similar pattern as VoiceMask$_\alpha$). McAdams achieves the peak tradeoff of 0.5704 (EER = 14.99%, J = 0.6014, WER = 26.45%, U = 0.8602, and P = 0.6232) with the coefficient of 0.75. The tradeoff plot has a plateau when the McAdams coefficient lies in the range of 0.7 and 0.8, as shown in Fig. 5a. The VoicePrivacy challenge 2020 [65] set a fixed coefficient of 0.8 for McAdams, and existing work by Patino et al. [49] experimented with different ranges of McAdams coefficient (e.g., $\alpha \in$[0.7, 0.9] and $\alpha \in$[0.5, 0.9]) but do not conclude which range performs the best for both EER and WER.

Fig. 7b shows that the optimum tradeoff of 0.6114 for VoiceMask$_\alpha$ (EER = 22.11%, J = 0.4110, WER = 29.63%, U = 0.8230, and P = 0.7429) with $|\alpha|$ = 0.14. We also found the optimum tradeoff of 0.5884 for VoiceMask$_\beta$ (EER = 23.39%, J = 0.4788, WER = 29.32%, U = 0.8266, and P = 0.7118) with a $|\beta|$ = 0.55. VoiceMask [53] set the range of $|\alpha| \in [0.08, 0.10]$ in the bilinear function to maintain a speech recognition accuracy in the range of [0.72, 0.75] while $|\beta| \in [0.4, 0.6]$ in the quadratic function to maintain speech recognition accuracy in the range of [0.77, 0.81] by leveraging the LibriSpeech dataset [48]. We see our optimum coefficients differ from the existing works, which determine a range based on EER and WER. However, *VoicePM* considers different voice attributes and can automatically detect the optimum coefficient and determine the optimum range based on the tradeoff analysis.

5.5 Stochastic Anonymization

If an adversary knows the anonymizer but does not know the exact coefficient, a fixed and deterministic coefficient could be reverse-engineered to recover the original voice. We adopt a stochastic approach to randomize the coefficient from a uniform distribution, i.e., $\alpha \in [\alpha_{min}, \alpha_{max}]$ per session. As shown in Fig. 7a and 7b, we observe the tradeoff settles on a plateau when the coefficient of McAdams and $VoiceMask_\alpha$ in the range of [0.7, 0.8] and [0.13, 0.15], respectively. Therefore, we randomize the coefficient in these

Table 5: Performance of different voice anonymization models.

Model	Emotion (%)	Age (%)	Accent (%)	Gender(%)	Jaccard	Speaker Acc (%)	EER (%)	WER (%)	S	J	U	P	T
Baseline	100	80.95	87.94	97.87	0.8534	97.45	2.28	14.50	0.0000	1.0000	1.0000	0.0000	0.0000
McAdams	76.07	35.24	62.96	90.14	0.5386	71.04	18.39	26.42	0.8055	0.6311	0.8606	0.5872	0.4971
VocieMask$_\alpha$	71.97	37.25	49.89	50.67	0.4038	62.53	20.58	27.92	0.8166	0.4731	0.8431	0.6717	0.5558
VocieMask$_\beta$	71.70	36.34	54.20	67.45	0.4534	66.14	20.85	28.15	0.8178	0.5313	0.8404	0.6432	0.5303
HiFi-GAN	40.50	19.39	12.13	24.28	0.1561	50.06	48.32	17.44	0.9528	0.1829	0.9656	0.8849	0.8545
MaskCycleGAN	36.18	24.21	19.32	40.25	0.2056	51.21	39.95	72.12	0.9429	0.2410	0.3261	0.8510	0.2775
V-CLOAK	60.54	25.13	51.08	81.26	0.4107	50.02	52.79	23.81	0.9568	0.4812	0.8911	0.7378	0.6574

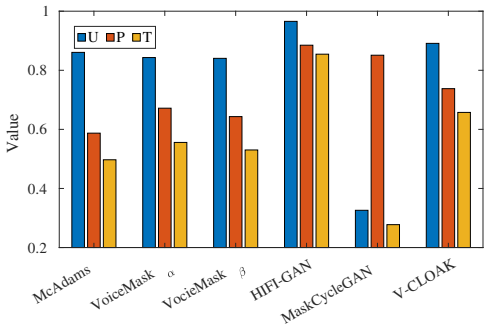


Figure 8: Overall performance for different voice anonymization techniques.

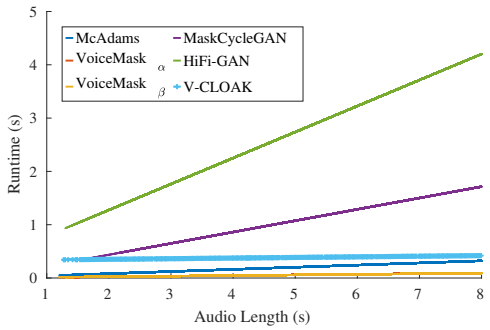


Figure 9: Run time required by different voice anonymization techniques for different length of utterance.

ranges for each utterance. By doing so, the adversary would have to know the exact coefficient used to anonymize the speech of any particular session to revert the conversion. To evaluate the stability of the proposed stochastic anonymization, we compute the tradeoff metric across ten runs using CV test set. The result shows that the standard deviation of the tradeoff metric is 0.0010, 0.0005, and 0.0011 for McAdams, $VoiceMask_\alpha$, and $VoiceMask_\beta$, respectively. We also observe similar small standard deviations for other metrics such as S, J, U, and P. We list the average value in Table 5.

5.6 Overall Performance

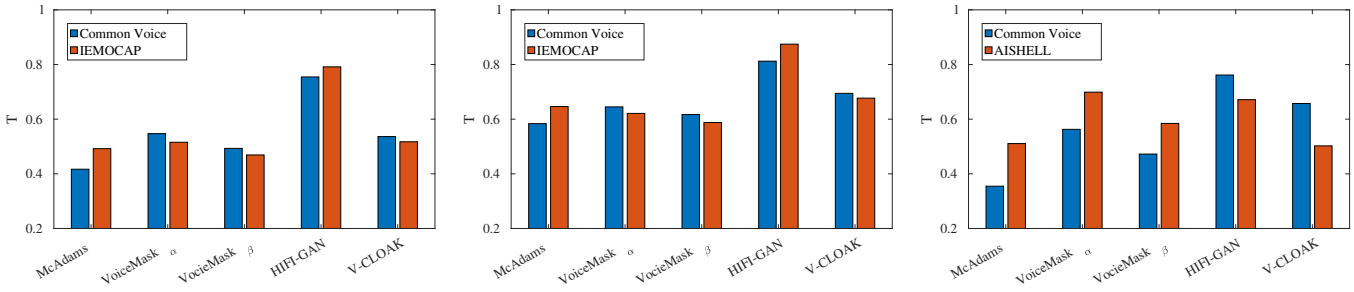
Fig. 8 plots the U, P, and T for all five models (details of other metrics are listed in Table 5). For the signal processing-based approaches, McAdams (U = 0.8606, P = 0.5872, T = 0.4971) slightly performs worse than VoiceMask. VoiceMask$_\alpha$ (U = 0.8431, P = 0.6717, T = 0.5558) has overall better performance in U and P than VoiceMask$_\beta$ (U = 0.8404, P = 0.6432, T = 0.5303). HiFi-GAN performs with the best tradeoff (T = 0.8545) among all five anonymizers, followed by V-CLOAK with a tradeoff of 0.6574. MaskCycleGAN preserves the highest privacy (P = 0.8510), but its utility (U = 0.3261) has the worst performance, resulting in the worst tradeoff (T = 0.2775). The reason is that its WER is as high as 72.12%. Recent work [33] trained and tested Cycle-GAN based voice converter [32] on Librispeech data, which showed that the WER is around 70% while significantly not improving with synthetic speech data augmentation. The CycleGAN networks can convert one speaker's voice to another to successfully impersonate the target. However, it is still challenging to make the converted voice more intelligible and recognizable by ASRs. The WER for McAdams, VoiceMask, HiFi-GAN, and V-CLOAK is less than 30%. *VoicePM* takes advantage of the normalized U to evaluate the utility, which shows U is greater than 0.84. A clean speech dataset would result in a lower WER.

5.7 Runtime Overhead

We ran *VoicePM* in an AMD Opteron (TM) 6128 2 GHz Processor with 32 GB RAM to test the real-time coefficient (RTC), which computes the ratio between the run time to anonymize the audio and the duration of the original audio. As the total CPU time is proportional to the duration of the utterance, we performed a linear function to fit the runtime. Fig.9 shows the CPU time for the six algorithms for the utterance duration from 1 to 8 seconds. The result shows that the runtime for the two VTLN-based approaches (VoiceMask) minimally increases linearly with a coefficient of 0.001, while the runtime for McAdams increases linearly with a coefficient of 0.041. HiFi-GAN takes one to two orders of magnitude longer than the signal processing-based approaches (e.g., VTLN and McAdams) with a linear coefficient of 0.4877. MaskCycleGAN increases linearly with a coefficient of 0.213, while V-CLOAK increases linearly with a coefficient of 0.012. The result demonstrates that signal processing-based approaches and adversarial examples take significantly less runtime.

5.8 Generalizability across Datasets

We use the IEMOCAP test set with gender and emotion state information to evaluate the generalizability of *VoicePM* when trained using the CV dataset. We keep the same ASR system, attribute inference models (i.e., only for gender and emotion), speaker verification model, and anonymization models. Due to MaskCycleGAN's poor performance (i.e., low utility), we drop this model for cross-dataset evaluation. Fig. 10a show that both datasets have a similar ranking from the highest tradeoff to the lowest tradeoff (HiFi-GAN -> VoiceMask$_\alpha$ -> V-CLOAK -> VoiceMask$_\beta$ ->McAdams). Detail listed in Table 6 in Appendix A. We also changed the γ parameter from 0.5 to 0.75 to see if the relative ranking changes. Fig. 10b shows that all the tradeoff points rise, but the overall relative ranking does not change across different γ values. Our design enables users to



(a) (b) (c)

Figure 10: Generalizability of tradeoff across different dataset with (a) γ = 0.5 and (b) γ = 0.75. (c) Transferability of tradeoff metric across speakers of different languages with γ = 0.5.

select a suitable γ value based on their preference. For instance, if the weight of S and J is considered equal, $\gamma = 0.5$ can be used, while a larger γ value indicates a higher weight on S (i.e., hiding speaker recognition).

5.9 Transferability across Different Languages

Next, we use the AISHELL1 Chinese Mandarin dataset to evaluate *VoicePM* performance across speakers of different languages. We keep all settings as §5.8 but only use a trained Chinese Mandarine vocoder [35] for the ASR. As the AISHELL1 dataset only have gender attribute, the Jaccard similarity only considers gender. Our baseline gender accuracy is 91.38% by using the same inference model trained on the CV dataset. Fig. 10c shows the tradeoff plot between Common Voice and AISHELL1 test set. Table 7 in Appendix A lists the detailed measurement. The relative tradeoff ranking in Fig. 10c is different from that of Fig. 10a, and this can be attributed to two factors. Firstly, all models, including attribute inference and anonymization models, were trained and validated using the English language, not Chinese (thus, highlighting the impact of transfer learning). Secondly, the privacy (P) of the Chinese dataset was heavily influenced by the accuracy of gender inference as that was the only attribute available for the AISHELL1 dataset. A Chinese dataset with more labeled attributes (e.g., age, accent, and emotion attributes) could potentially lead to better privacy representation, thus, a higher tradeoff value. For example, V-CLOAK and HiFi-GAN methods had higher gender inference accuracy using the Chinese dataset than the English dataset, resulting in lower privacy (P). As $T = U \times P$, T ends with a significant drop in the Chinese AISHELL1 dataset.

6 DISCUSSION

Our evaluation shows that different anonymization models can hide emotion, age, accent, and gender to different degrees. For example, VoiceMask$_\alpha$ can reduce the gender inference accuracy around or below 50%, which is lower than a random guess, and V-CLOAK can reduce the age inference accuracy from the baseline accuracy of 80.95% to 25.13%. HiFi-GAN can normalize target speakers with a fixed gender, accent, age group, and neutral emotional state. Anonymization models with varying privacy levels can be pre-defined. *VoicePM* can enable vendors to design a privacy configuration (i.e., privacy settings) to allow users to hide their voice attributes at different levels.

Feasibility for Attributes Configuration. We list the tradeoff measurement of all possible 16 combinations from four-voice attributes: emotion, age, accent, and gender in Table 8 (Appendix A). *VoicePM* can provide a tradeoff rank for each combination. For example, if the user wants to hide the {emotion} attribute alone, *VoicePM* would recommend HiFi-GAN, which has the best tradeoff (T = 0.6748), followed by V-CLOAK with a tradeoff of 0.557 (with a lower computational overhead). *VoicePM* can consider many anonymized models. Thus, *VoicePM* can provide a better solution with a higher tradeoff based on the user's privacy configuration and system computation capability.

Limitations. First, we have tried our best to implement the state-of-the-art attribute inference models. However, the accuracy of the emotion and age inference model is relatively low. *VoicePM* could be updated with newer models once more advanced models are available. Second, we use the self-reported binary gender label from CV dataset. We removed the 'other' gender label due to its small portion of data. In addition, the reported gender identity could differ from the biological determinants of sex. These might degrade the performance of the gender inference model if many such audio samples are included in the train or test set. Thirdly, we limited our analysis to audio files of less than 8 seconds due to computation constraints. Lastly, our analysis lacks human perception verification of the altered audio, which we leave as future work.

7 CONCLUSION

In this paper, we build and evaluate voice attribute inference models, including emotion, age, accent, and gender, by adopting the latest speech representation learning networks, such as `wav2vec2` and `ECAPA-TDNN`. We then 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. We extensively evaluate using three datasets and develop tradeoff metrics to study the feasibility of existing anonymization approaches. Our experiments show that *VoicePM* can effectively measure the tradeoff of different anonymization models for a larger set of voice attributes. *VoicePM* has the potential to foster the design of privacy settings for emerging voice assistants.

ACKNOWLEDGEMENTS

We thank our anonymous reviewers and shepherd for their valuable feedback. This material is based upon work supported in parts by a Meta research gift.

REFERENCES

[1] 2017. Google, Amazon Patent Filings Reveal Digital Home Assistant Privacy Problems. http://www.consumerwatchdog.org/sites/default/files/2017-12/Digital%20Assistants%20and%20Privacy.pdf.

[2] 2022. How does Microsoft protect my privacy while improving its speech recognition technology? https://support.microsoft.com/en-us/windows/how-does-microsoft-protect-my-privacy-while-improving-its-speech-recognition-technology-f465d7a7-4a4f-40b7-9441-f0e6e97e24ec

[3] 2022. *Improve Siri and Dictation & Privacy.* https://www.apple.com/legal/privacy/data/en/improve-siri-dictation/

[4] 2022. *Voiceprints and Biometric Litigation.* https://tinyurl.com/2meuz9j5

[5] 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. https://www.usenix.org/conference/usenixsecurity23/presentation/deng

[6] Ranya Aloufi, Hamed Haddadi, and David Boyle. 2019. Emotion filtering at the edge. In *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems*. 1–6.

[7] Ranya Aloufi, Hamed Haddadi, and David Boyle. 2020. Privacy-preserving voice analysis via disentangled representations. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*. 1–14.

[8] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 4218–4222.

[9] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. *Advances in neural information processing systems* 31 (2018).

[10] Levent M Arslan and John HL Hansen. 1996. Language accent classification in American English. *Speech Communication* 18, 4 (1996), 353–367.

[11] 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.

[12] Hamid Behravan, Ville Hautamäki, Sabato Marco Siniscalchi, Tomi Kinnunen, and Chin-Hui Lee. 2015. I-vector modeling of speech attributes for automatic foreign accent recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 1 (2015), 29–41.

[13] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AIShell-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline. In *Oriental COCOSDA 2017*.

[14] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[15] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church. 2021. Speech emotion recognition with multi-task learning. In *Proceedings of Interspeech*, Vol. 2021.

[16] Joon Son Chung, Arsha Nagrani, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman. 2019. VoxSRC 2019: The first VoxCeleb speaker recognition challenge. *arXiv preprint arXiv:1912.02522* (2019).

[17] 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).

[18] David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. 2018. An open-source speaker gender detection framework for monitoring gender equality. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 5214–5218.

[19] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre. 2019. Speaker anonymization using x-vector and neural waveform models. *arXiv preprint arXiv:1905.13561* (2019).

[20] Luciana Ferrer, Diego Castan, Mitchell McLaren, and Aaron Lawson. 2022. A Discriminative Hierarchical PLDA-Based Model for Spoken Language Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2396–2410.

[21] Daniel Garcia-Romero and Carol Y Espy-Wilson. 2011. Analysis of i-vector length normalization in speaker recognition systems. In *Proceedings of the 12th annual conference of the international speech communication association.*

[22] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).

[23] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, and Anne-Maria Laukkanen. 2015. Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication* 72 (2015), 13–31.

[24] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen. 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Interspeech*. 930–934.

[25] Khaled Hechmi, Trung Ngo Trong, Ville Hautamäki, and Tomi Kinnunen. 2021. VoxCeleb Enrichment for Age and Gender Recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 687–693.

[26] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

[27] Huafeng Jin and Shuo Wang. 2018. Voice-based determination of physical and emotional characteristics of users. US Patent 10,096,319.

[28] Lewis Kamb. 2022. Lawsuit claims Amazon using Alexa to target ads at customers. https://www.axios.com/local/seattle/2022/06/16/lawsuit-amazon-alexa-target-ads-customers.

[29] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2019. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6820–6824.

[30] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2020. Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion. *arXiv preprint arXiv:2010.11672* (2020).

[31] 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.

[32] Gokce Keskin, Tyler Lee, Cory Stephenson, and Oguz H Elibol. 2019. Many-to-many voice conversion with out-of-dataset speaker support. *arXiv preprint arXiv:1905.02525* (2019).

[33] Gokce Keskin, Tyler Lee, Cory Stephenson, and Oguz H Elibol. 2019. Measuring the effectiveness of voice conversion on speaker identification and automatic speech recognition systems. *arXiv preprint arXiv:1905.12531* (2019).

[34] Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In *Proc. Interspeech 2017*. 2–6.

[35] 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.

[36] Jacob Leon Kröger, Leon Gellrich, Sebastian Pape, Saba Rebecca Brause, and Stefan Ullrich. 2022. Personal information inference from voice recordings: User awareness and privacy concerns. *Proceedings on Privacy Enhancing Technologies* 2022, 1 (2022), 6–27.

[37] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems* 32 (2019).

[38] Damian Kwasny and Daria Hemmerling. 2021. Gender and age estimation methods based on speech using deep neural networks. *Sensors* 21, 14 (2021), 4785.

[39] Regina Helena Garcia Martins, Tatiana Maria Gonçalvez, Adriana Bueno Benito Pessin, and Anete Branco. 2014. Aging voice: presbyphonia. *Aging clinical and experimental research* 26, 1 (2014), 1–5.

[40] Sutapa Mondal, Mangesh S Gharote, and Sachin P Lodha. 2022. Privacy of Personal Information: Going incog in a goldfish bowl. *Queue* 20, 3 (2022), 41–87.

[41] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6922–6926.

[42] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60 (2020), 101027.

[43] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. *Proc. Interspeech 2017* (2017), 2616–2620.

[44] Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, et al. 2019. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language* 58 (2019), 441–480.

[45] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, Vol. 1. 380–384.

[46] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[47] Anjali Pahwa and Gaurav Aggarwal. 2016. Speech feature extraction for gender recognition. *International Journal of Image, Graphics and Signal Processing* 8, 9 (2016), 17.

[48] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of the 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

[49] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans. 2021. Speaker Anonymisation Using the McAdams Coefficient. In *Interspeech 2021*. ISCA, 1099–1103.
[50] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans. 2021. Speaker Anonymisation Using the McAdams Coefficient. In *Interspeech 2021*. ISCA, 1099–1103.
[51] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. *Proc. Interspeech 2021*, 3400–3404.
[52] Cyril R Pernet and Pascal Belin. 2012. The role of pitch and timbre in voice gender categorization. *Frontiers in psychology* 3 (2012), 23.
[53] 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.
[54] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624* (2021).
[55] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language* 27, 1 (2013), 4–39.
[56] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.
[57] Rita Singh. 2019. *Profiling humans from their voice*. Vol. 41. Springer.
[58] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. Spoken language recognition using x-vectors.. In *Odyssey*. 105–111.
[59] 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.
[60] Daniel Stoller, Sebastian Ewert, and Simon Dixon. 2018. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185* (2018).
[61] David Sundermann and Hermann Ney. 2003. VTLN-based voice conversion. In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795)*. IEEE, 556–559.
[62] Naohiro Tawara, Atsunori Ogawa, Yuki Kitagishi, and Hosana Kamiyama. 2021. Age-vox-celeb: Multi-modal corpus for facial and speech estimation. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6963–6967.
[63] Ingo R Titze and Daniel W Martin. 1998. Principles of voice production.
[64] Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Hubert Nourtel, Pierre Champion, Massimiliano Todisco, Emmanuel Vincent, Nicholas Evans, Junichi Yamagishi, and Jean François Bonastre. 2022. The VoicePrivacy 2022 Challenge Evaluation Plan. *arXiv preprint arXiv:2203.12468* (2022).
[65] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al. 2022. The VoicePrivacy 2020 Challenge: Results and findings. *Computer Speech & Language* 74 (2022), 101362.
[66] Danielle Toth. 2021. Amazon Uses Alexa to Unlawfully Collect, Store, Uses Biometric Data, Class Action Lawsuit Claims. https://topclassactions.com/lawsuit-settlements/privacy/amazon-uses-alexa-to-unlawfully-collect-store-uses-biometric-data-class-action-lawsuit-claims/.
[67] Henry Turner, Giulio Lovisotto, Simon Eberz, and Ivan Martinovic. 2022. I Am Hearing Different Voices: Anonymous Voices to Protect User Privacy. *arXiv preprint arXiv:2202.06278* (2022).
[68] Nik Vaessen and David A Van Leeuwen. 2022. Fine-tuning wav2vec2 for speaker recognition. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7967–7971.
[69] Zhi-Feng Wang, Gang Wei, and Qian-Hua He. 2011. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *Proceedings of the IEEE 2011 International Conference on Machine Learning and Cybernetics (ICMLC)*, Vol. 4. 1708–1713.
[70] Hamid Yeganeh. 2021. Emerging social and business trends associated with the Covid-19 pandemic. *Critical perspectives on international business* (2021).
[71] Cheng Yi, Jianzong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2021. Transfer Ability of Monolingual Wav2vec2. 0 for Low-resource Speech Recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*. 1–6.
[72] Haitong Zhang and Yue Lin. 2022. Improve few-shot voice cloning using multi-modal learning. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8317–8321.
[73] Hao Zhu, Yanyong Zhang, Xing Guo, and Xiang-Yang Li. 2021. Anti Leakage: Protecting Privacy Hidden in Our Speech. In *2021 7th International Conference on Big Data Computing and Communications (BigCom)*. 114–120.
[74] Ellie Zolfagharifard. 2018. Google and Amazon patent creepy SPY systems that use cameras and sensors in your home to know everything from your mood to your medical conditions. http://www.dailymail.co.uk/sciencetech/article-5569121/Google-Amazon-patent-creepy-Big-Brother-style-systems-spy-you.html.

APPENDIX

A PERFORMANCE OF DIFFERENT VOICE ANONYMIZATION MODELS

Table 6: Performance of different voice anonymization models on the IEMOCAP dataset.

Model	Gender (%)	Emotion (%)	Jaccard	Speaker Acc (%)	EER (%)	WER (%)	S	J	U	P	T
Baseline	81.49	77.31	0.7399	97.10	0.10	24.57	0.0000	1.0000	1.0000	0.0000	0.0000
McAdams	63.16	66.42	0.5717	80.00	10.50	39.04	0.9905	0.7727	0.8082	0.6089	0.4921
VocieMask$_\alpha$	38.84	65.70	0.4325	70.90	13.20	44.77	0.9924	0.5846	0.7323	0.7039	0.5155
VocieMask$_\beta$	51.18	64.61	0.4985	67.00	16.00	46.39	0.9938	0.6738	0.7107	0.6600	0.4691
HiFi-GAN	43.19	23.77	0.2577	50.40	48.70	27.62	0.9979	0.3483	0.9595	0.8248	0.7914
V-CLOAK	59.17	69.15	0.5650	52.40	39.80	36.75	0.9975	0.7637	0.8386	0.6169	0.5173

Table 7: Performance of different voice anonymization models on the AISHELL-1 (Mandarin Chinese) dataset.

Model	Gender (%)	Jaccard	Speaker Acc (%)	EER (%)	WER (%)	S	J	U	P	T
Baseline	91.39	0.9139	82.76	0.97	5.04	0.0000	1.0000	1.0000	0.0000	0.0000
McAdams	75.85	0.7585	77.64	19.93	13.51	0.9513	0.8300	0.9108	0.5607	0.5107
VocieMask$_\alpha$	44.38	0.4438	60.32	28.96	10.38	0.9665	0.4857	0.9437	0.7404	0.6988
VocieMask$_\beta$	68.27	0.6827	56.09	41.53	9.74	0.9766	0.7470	0.9505	0.6148	0.5844
HiFi-GAN	50.39	0.5039	47.20	54.11	10.87	0.9821	0.5514	0.9386	0.7153	0.6714
V-CLOAK	66.29	0.6629	50.00	50.70	10.35	0.9809	0.7254	0.9441	0.6277	0.5926

Table 8: Tradeoff with various attributes selection. *VoicePM* would recommend the anonymization model with a higher tradeoff based on user's configuration.

Attributes	McAdams (U=0.8466)		VoiceMask$_\alpha$ (U=0.8274)		VocieMask$_\beta$ (U=0.8245)		HiFi-GAN (U=0.9130)		MaskCycleGAN (U=0.3261)		V-CLOAK (U=0.8911)	
	P	T	P	T	P	T	P	T	P	T	P	T
basic privacy	0.4431	0.3752	0.4488	0.3714	0.4493	0.3704	0.4764	0.4350	0.4715	0.1538	0.4784	0.4263
emotion	0.5066	0.4289	0.5420	0.4484	0.5399	0.4451	0.7391	0.6748	0.7595	0.2477	0.6237	0.5557
age	0.7634	0.6463	0.7562	0.6257	0.7627	0.6288	0.8628	0.7878	0.8296	0.2706	0.8312	0.7406
accent	0.5878	0.4976	0.6798	0.5625	0.6529	0.5383	0.9053	0.8266	0.8583	0.2799	0.6791	0.6052
gender	0.4189	0.3547	0.6803	0.5628	0.5729	0.4723	0.8342	0.7616	0.7356	0.2399	0.5023	0.4476
emotion+accent	0.5900	0.4996	0.6609	0.5468	0.6436	0.5306	0.8638	0.7886	0.8445	0.2754	0.7015	0.6251
emotion+age	0.6936	0.5872	0.7040	0.5825	0.7081	0.5838	0.8438	0.7704	0.8373	0.2731	0.7835	0.6981
emotion+gender	0.4924	0.4169	0.6612	0.5471	0.5981	0.4931	0.8264	0.7545	0.7861	0.2564	0.6018	0.5363
age+accent	0.7265	0.6151	0.7651	0.6331	0.7566	0.6237	0.9105	0.8313	0.8744	0.2852	0.8024	0.7149
gender+accent	0.5428	0.4596	0.7286	0.6029	0.6579	0.5424	0.8985	0.8203	0.8363	0.2727	0.6386	0.5690
gender+age	0.6548	0.5544	0.7643	0.6324	0.7196	0.5933	0.8823	0.8056	0.8264	0.2695	0.7312	0.6515
emotion+age+accent	0.6889	0.5832	0.7275	0.6020	0.7208	0.5943	0.8825	0.8058	0.8624	0.2812	0.7797	0.6948
emotion+accent+gender	0.5577	0.4722	0.7013	0.5803	0.6503	0.5361	0.8731	0.7971	0.8337	0.2719	0.6655	0.5930
emotion+age+gender	0.6347	0.5373	0.7274	0.6019	0.6939	0.5721	0.8620	0.7871	0.8296	0.2706	0.7257	0.6466
gender+age+accent	0.6611	0.5597	0.7679	0.6354	0.7283	0.6005	0.9047	0.8260	0.8569	0.2795	0.7431	0.6622
emotion+age+accent+gender	0.6454	0.5464	0.7393	0.6117	0.7072	0.5830	0.8849	0.8080	0.8510	0.2775	0.7378	0.6574