VoiceCloak: Adversarial Example Enabled Voice De-Identification with Balanced Privacy and Utility

MENG CHEN, Zhejiang University, China and ZJU-Hangzhou Global Scientific and Technological Innovation Center, China
LI LU*, Zhejiang University, China
JUNHAO WANG, Zhejiang University, China
JIADI YU, Shanghai Jiao Tong University, China
YINGYING CHEN, Rutgers University, USA
ZHIBO WANG, Zhejiang University, China
ZHONGJIE BA, Zhejiang University, China
FENG LIN, Zhejiang University, China
KUI REN, Zhejiang University, China

Faced with the threat of identity leakage during voice data publishing, users are engaged in a privacy-utility dilemma when enjoying the utility of voice services. Existing machine-centric studies employ direct modification or text-based re-synthesis to de-identify users' voices but cause inconsistent audibility for human participants in emerging online communication scenarios, such as virtual meetings. In this paper, we propose a human-centric voice de-identification system, *VoiceCloak*, which uses adversarial examples to balance the privacy and utility of voice services. Instead of typical additive examples inducing perceivable distortions, we design a novel convolutional adversarial example that modulates perturbations into real-world room impulse responses. Benefiting from this, *VoiceCloak* could preserve user identity from exposure by Automatic Speaker Identification (ASI), while remaining the voice perceptual quality for non-intrusive de-identification. Moreover, *VoiceCloak* learns a compact speaker distribution through a conditional variational auto-encoder to synthesize diverse targets on demand. Guided by these pseudo targets, *VoiceCloak* constructs adversarial examples in an input-specific manner, enabling any-to-any identity transformation for robust de-identification. Experimental results show that *VoiceCloak* could achieve over 92% and 84% successful de-identification on mainstream ASIs and commercial systems with excellent voiceprint consistency, speech integrity, and audio quality.

*Corresponding Author

Authors' addresses: Meng Chen, meng.chen@zju.edu.cn, Zhejiang University, School of Cyber Science and Technology, Key Laboratory of Blockchain and Cyberspace Governance of Zhejiang Province, Hangzhou, China and ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China; Li Lu, li.lu@zju.edu.cn, Zhejiang University, School of Cyber Science and Technology, Key Laboratory of Blockchain and Cyberspace Governance of Zhejiang Province, Hangzhou, China; Junhao Wang, wangjunhao@zju.edu.cn, Zhejiang University, School of Cyber Science and Technology, Hangzhou, China; Jiadi Yu, jiadiyu@sjtu.edu.cn, Shanghai Jiao Tong University, Department of Computer Science and Engineering, Shanghai, China; Yingying Chen, yingche@scarletmail.rutgers.edu, Rutgers University, WINLAB, Department of Electrical and Computer Engineering, Piscataway, NJ, USA; Zhibo Wang, zhibowang@zju.edu.cn, Zhejiang University, School of Cyber Science and Technology, Hangzhou, China; Zhongjie Ba, zhongjieba@zju.edu.cn, Zhejiang University, School of Cyber Science and Technology, Hangzhou, China; Feng Lin, flin@zju.edu.cn, Zhejiang University, School of Cyber Science and Technology, Hangzhou, China; Kui Ren, kuiren@zju.edu.cn, Zhejiang University, School of Cyber Science and Technology, Hangzhou, China.

2474-9567/2023/6-ART48 $15.00
https://doi.org/10.1145/3596266

CCS Concepts: • **Security and privacy → Privacy protections**; *Usability in security and privacy*.

Additional Key Words and Phrases: voice de-identification, adversarial examples, voice privacy preservation

ACM Reference Format:
Meng Chen, Li Lu, Junhao Wang, Jiadi Yu, Yingying Chen, Zhibo Wang, Zhongjie Ba, Feng Lin, and Kui Ren. 2023. *VoiceCloak*: Adversarial Example Enabled Voice De-Identification with Balanced Privacy and Utility. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 48 (June 2023), 21 pages. https://doi.org/10.1145/3596266

1 INTRODUCTION

Recent decades have witnessed voice input becoming one of the most prevalent methods widely deployed in various services. Richer functional utility, including automatic speech transcription, efficient voice searching, and live language translation, thus gradually enables humans to enjoy a natural but much more intelligent experience. Especially under the shock of COVID-19 pandemic, people have to shift to online platforms (e.g., Zoom [66], Microsoft Teams [40], Google Meet [24]) for remote work and virtual meeting, where voice services, including real-time speech note and automated speaker annotation, greatly facilitate the communication and collaboration efficiency. However, behind the powerful utility of voice services, the privacy risks of voice data publishing raise extensive public concerns. Many leading tech giants are collecting and storing users' voices in practice [25, 39] or even eavesdropping on users' conversations without any consent [22, 57]. This exposes users to the risk of identity leakage by specialized Automatic Speaker Identification (ASI) tools [29, 41], which can extract voiceprints after listening to only 8~10 words [5]. Such voiceprints may be used to disclose Personal Identifiable Information (PII) for targeted advertisement [45] by user profiling or even malicious impersonation.

Caught in such a dilemma between the high utility of voice services and personal identity privacy, voice de-identification is proposed to eliminate individual traits while maintaining the linguistic content for other downstream tasks (e.g., Automatic Speech Recognition, ASR). Related studies focus on the voice conversion [2, 20, 27, 30, 38, 42, 46, 47, 54, 58, 59, 64] or text-to-speech re-synthesis [3, 31, 47] to transform or exclude individual features in voices. However, these methods are designed for machine-centric tasks, i.e., protecting user identity against ASI while remaining correct speech transcripts from ASR, ignoring human-centric experiences, i.e., the perceptual quality of de-identified voices significantly declines due to the inconsistent voiceprint and severe distortion. This significantly downgrades the user experience of human listeners and cause misunderstandings during interpersonal interactions including virtual meeting, social media publishing, and instant messaging, thus neglecting the speech utility.

Toward this end, we take a different viewpoint to balance the speech utility and identity privacy of voice services. Inspired by the strong threat to learning-based automatic systems and the excellent imperceptibility to humans, we introduce adversarial examples [7, 8, 23, 36, 65] to conceal speaker identity while remaining speech integrity and perceptual consistency, which serves as a more ideal de-identification tool. However, applying adversarial examples to voice de-identification is also a challenging task. On the one hand, existing methods [1, 9–12, 35, 48, 50, 60, 63] generate *additive adversarial perturbations* with amplitude normalization or psychoacoustic masking to constrain the perturbation audibility, which either induces perceivable high-frequency artifacts [50] or is easily corrupted by well-designed filters [18, 28], making them inapplicable for voice de-identification. On the other hand, it's still challenging to resist re-identification attacks with partial or full knowledge about the de-identification [55].

To address the perceivable artifacts caused by additive perturbations, we revise the channel distortion of sound propagation, and propose a novel *convolutional adversarial perturbation*. Theoretically, apart from the ambient additive noises, the channel interference of airborne sound also includes its own delayed reflections due to the multi-path effect. These reflections can be quantified by a Room Impulse Response (RIR) that convolves with the dry voice and behaves as a natural reverberation. Hence, it is more difficult for human to discern them

as an abnormal signal. Inspired by this observation, we wonder whether the RIR can be carefully crafted as a convolutional adversarial perturbation to conceal user identity while being transparent for human audibility. In addition, unlike existing methods that transforms users' identity to a pool of collected real speakers, we also propose to synthesize pseudo targets using generative models to guide the adversarial example generation. This could improve the identity diversity and unlinkability of de-identified voices under re-identification and also reduce the storage overhead for practical deployment.

To achieve these, we propose *VoiceCloak*, a non-intrusive voice de-identification system. First, different from the additive adversarial perturbations, *VoiceCloak* injects convolutional adversarial perturbations to de-identify voices while avoiding perceivable artifacts. By reshaping the convolutional adversarial perturbations into real-world RIRs, *VoiceCloak* approximates the perturbation injection to a natural reverberation effect, remaining the voiceprint consistency, speech integrity and audio quality for human participants. Second, to provide diverse targets for de-identification with limited resources, *VoiceCloak* pre-trains a lightweight conditional variational auto-encoder at the embedding level. With this generative model, users can synthesize any desired target embeddings on demand, which improves the diversity of de-identified voices. Finally, *VoiceCloak* adopts a triplet loss architecture for iterative perturbation construction, whose input-specific manner empowers *VoiceCloak* any-to-any identity transformation, enabling any source user to conceal his/her identity among a large group of different target speakers. Experimental results show that *VoiceCloak* could achieve effective voice de-identification on mainstream ASIs and commercial systems with satisfactory speech recognition and perceptual quality. The de-identified voice examples are provided at the demo page [1].

Our contributions are highlighted as follows:

- To the best of our knowledge, *VoiceCloak* is the first work to employ convolutional adversarial examples to realize voice de-identification, which achieves a good balance between the privacy and utility of voice services.
- We propose a novel convolutional adversarial example method to modulate adversarial perturbations into real-world RIRs, which improves the perceptual consistency in terms of the voiceprint, speech content, and audio quality, realizing a non-intrusive voice de-identification.
- We design a triplet loss architecture for input-specific perturbation construction and develop an embedding-level conditional variational auto-encoder to sample diverse target embeddings on demand, enabling any-to-any identity transformation for robust voice de-identification.
- We conduct extensive experiments against four mainstream and commercial ASIs on two voice datasets. The results show that *VoiceCloak* achieves over 92% and 84% successful de-identification on mainstream and commercial ASI systems with a word accuracy drop of less than 10%, reaching a Mel cepstral distortion of 5.13dB and a short-time objective intelligence over 0.81. The subjective evaluation also shows excellent perceptual quality with a mean opinion score over 4.25.

2 PRELIMINARY

In this section, we illustrate the voice de-identification system and threat model and identify the key design goals of privacy-utility balanced de-identification. To fulfill the design goals, we propose *VoiceCloak* and introduce its basic idea with a high-level overview.

2.1 System and Threat Models

Fig. 1 shows the system and threat models. In a voice service, a user's raw voice is first captured by devices, such as having a virtual meeting on online applications, sharing videos on social media platforms, or interacting with personal voice assistants. The captured voices are transmitted not only to human participants (e.g., conference audience, social friends) but also to a cloud server for specific services (e.g., ASR). We assume the adversary can

[1] https://zju-muslab.github.io/projects/voicecloak

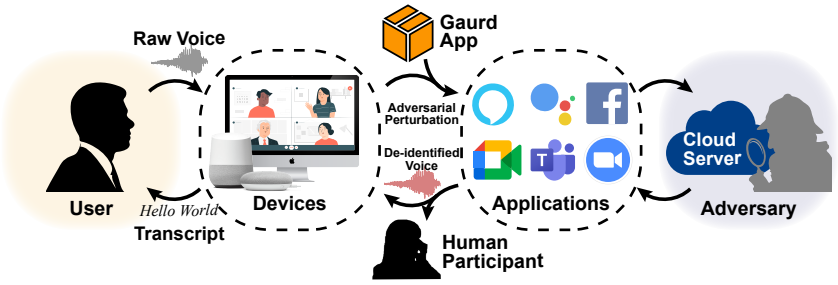


Fig. 1. System and threat model.

access the user's voice data on the cloud server, such as an external hacker, an internal curious data analyst, or even the service provider itself. With the collected voice data, the adversary can exploit ASI tools to extract users' voiceprints and explore speaker identities. Specifically, the adversary may build a pool of speaker models via ASI with available voice recordings from other sources, where the adversary may also retrieve some background or even identifiable information about these victim candidates. And then the adversary can identify a specific person by matching his/her utterance with the stored speaker models. For example, malicious service providers or third-party data consumers can link authenticated (with PII) and unauthenticated (without PII) user accounts by comparing the speaker models of collected voices from different applications. Telecom companies driven by ulterior motives can easily infer the identity of a public utterance by matching its voiceprint to those of stored recordings connected with a phone number. And curious individuals can also crawl massive voice data attached to informative user profiles on the Internet without much effort to examine the identity of target persons. Once the user's voiceprint is exposed, it may be exploited to make user profiles for precise advertising, be shared with third parties for commercial purposes, or even be cloned to craft DeepFake voices.

To protect the user identity privacy while maintaining voice service utility, this work aims to design a guard APP at the user side for voice de-identification. Considering the great threat to neural networks and excellent imperceptibility of adversarial examples, we intend to turn the powerful well-crafted examples into a new privacy-preserving tool to protect speaker identity from leakage to ASIs when using convenient voice services. We assume the guard APP is installed in advance and runs in the background on the user's device. Before the raw voice is uploaded to the cloud, the guard APP is invoked and imposes a subtle perturbation on the raw voice for de-identification. The APP can be set to automation activated by microphone use events, be plugged in the voice input editor, or integrated into the operating system. We discuss the different deployment modes in Section 5. Then the de-identified voices are expected to not only conceal the user identity from ASI and yield correct transcript after ASR, but also maintain perceptually consistent in speaker voiceprint, speech text and audio quality for human participants. This non-intrusive de-identification scheme enables users to balance identity privacy and service utility. In online voice publishing scenarios such as virtual meetings and instant messaging, users can de-identify their voices while keeping voiceprint consistency for normal interpersonal communication. In offline voice post-processing scenarios such as taking Vlogs and sharing singing, users can conceal their identity without impairing the perceptual quality of local audio before uploading them to public platforms.

Blocked by our voice de-identification scheme, the adversary may further take measures to uncover speaker identities from the de-identified voices, i.e., *Speaker re-identification*. In this case, the adversary has full knowledge about the de-identification strategy and implementation details, with which the adversary performs the same processing on collected clean voices to re-identify the de-identified voices [55]. By considering these advanced attacks with strong capability, our de-identification aims to support more robust protection for users.

2.2 Design Goals

Traditional voice de-identification schemes are only designed for affecting intelligent systems, i.e., preserving users' identity from disclosure by ASI, and maintaining correct transcription after ASR. Unlike such machine-centric solutions, our work turns to focus on a user-centric voice de-identification to realize a privacy-utility trade-off, which should not only maintain the correct transcription after ASR, but also need to fulfill the following key design goals to improve user experience:

- **Voiceprint consistency**. During the interpersonal interaction between users and other human participants, it is necessary to have the users' voiceprints perceptually consistent before and after the de-identification, for avoiding misunderstandings of conversation contexts.
- **Speech integrity**. To minimize the interference in normal communication between users and human participants, the de-identified voices are required to maintain integral linguistic content recognized by humans.
- **Audio quality**. For a non-intrusive user experience, the de-identification should produce high-quality audio without perceivable distortion.

2.3 Motivation and Challenges

To fulfill the design goals achieving the privacy-utility balance of voice services, we propose a novel adversarial example-based voice de-identification scheme. Instead of directly manipulating the voiceprint, we are inspired by the adversarial examples' the strong threat to deep neural networks and excellent imperceptibility to human ears. Based on the observation, its basic idea is to inject subtle perturbations on raw voices to generate audio adversarial examples as a de-identification tool for tackling the privacy-utility dilemma, which transforms the digital identity for deceiving ASIs while remains the acoustic characteristics for human perception.

However, applying adversarial examples for voice de-identification is not a straightforward way, due to the following challenges: (1) *Perceivable and fragile additive perturbation.* Most existing methods follow the paradigm of additive adversarial perturbation, i.e., generating and overlaying adversarial perturbations on original voices to derive adversarial examples. To suppress the audibility of such additive noises, amplitude normalization [9, 10, 12, 35, 63] is widely applied to confine their loudness, but still induce perceivable distortions, especially high-frequency artifacts [50]. Following studies [1, 11, 48, 50, 60] turn to apply psychoacoustic masking to construct inaudible perturbations below the hearing threshold of the human auditory system. But unfortunately, a recent work [18] has demonstrated that such perturbations can be easily filtered out by a simple mask operation based on the same psychoacoustic principle. Hence, these additive adversarial examples are either easily noticed by human beings or corrupted by well-designed filters, making them inapplicable for voice de-identification. (2) *Adaptive speaker re-identification.* As mentioned in the threat model, the transparency and availability of a voice de-identification scheme inform the adversary of its internal mechanism and detailed implementation. This may be further exploited by the adversary to re-identify the original identity of users [55]. Therefore, a new non-intrusive, and robust solution to voice de-identification is highly desired for balancing the privacy and utility of voice services.

3 VOICECLOAK DESIGN

3.1 System Overview

To solve the aforementioned challenges, we propose *VoiceCloak*, a non-intrusive and robust voice de-identification system. Fig. 2 shows the system overview of *VoiceCloak*, which consists of three components.

To reduce the perceivable artifacts resulting from additive perturbations, the **Convolutional Perturbation Injector** reshapes a novel convolutional adversarial perturbation to a real-world room impulse response to produce adversarial examples. This modulates the perturbations into the natural reverberation effect of airborne sound propagation, thus maintaining the perceptual characterastics in voice. Moreover, instead of storing a

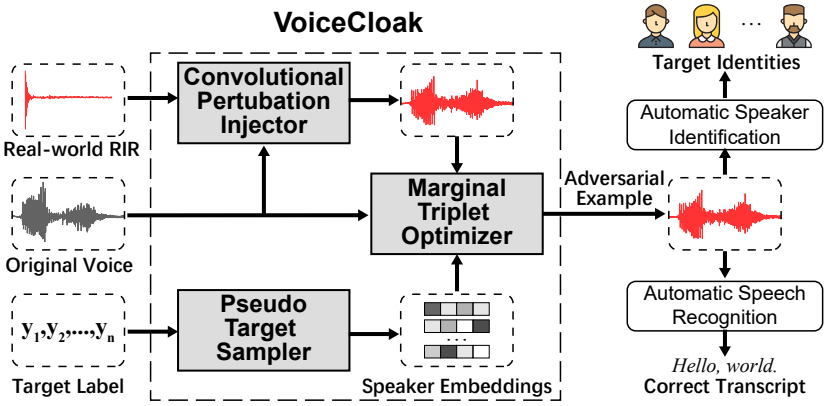


Fig. 2. System overview of *VoiceCloak*.

pool of real target voices as previous studies, the **Pseudo Target Sampler** pre-trains a lightweight conditional variational auto-encoder at the embedding level to sample diverse targets on demand. By explicitly modulating prior identity knowledge to the generative model, the sampler could learn a compact speaker identity distribution in the latent space, and sample the embeddings of any target even a pseudo speaker as instances for perturbation optimization, which improves the diversity and unlinkability of de-identified voices. With the original voice and pseudo speaker embedding, the **Marginal Triplet Optimizer** adopt a triplet loss architecture to optimize adversarial examples in an input-specific manner. This enables any source user to disguise a large group of target speakers in different utterances, thus further increasing the difficulty of speaker re-identification. Consequently, the adversarial examples would be identified as different target speakers by ASI while preserving the correct transcription after ASR.

3.2 Convolutional Perturbation Injector

Compared with voice conversion and speech re-synthesis that directly modify voiceprint or generate audio with significant distortions, VoiceCloak imposes subtle adversarial perturbations on raw voices, which remains a perceptually consistent voiceprint, content, and quality for human participants while effectively deceiving ASIs. Unlike typical additive adversarial perturbations that either induce perceivable artifacts or are easily filtered out, we propose a novel imperceptible perturbation construction approach to minimize intrusiveness to human participants.

Theoretically, the over-the-air propagation of sound waves involves two kinds of channel interference, i.e., additive noise and convolutional reverberation. Different from the typical adversarial example methods that impose perceivable additive noises on original voices, we turn to explore the covolutional reverberation. The reverberation is caused by the multi-path effect during the sound propagation in physical space. As shown in Fig. 3, in an enclosed room, the sound waves coming from the transmitter propagate omnidirectionally so that the received signal at the receiver mainly includes: (1) the waves travel through the direct path; (2) the echoes reflected from surrounding walls with different delays; (3) the ambient noise. Such a roughly linear time-invariant process can be quantified as a convolution on the original voice with a Room Impulse Response (RIR) as shown in Fig. 4. As a result, the RIR-convolved speech exhibits a reverberation effect and is difficult to be distinguished as an anomalous signal by humans. Inspired by this observation, we propose to conduct RIR-like convolutional adversarial perturbations for realizing imperceptibility.

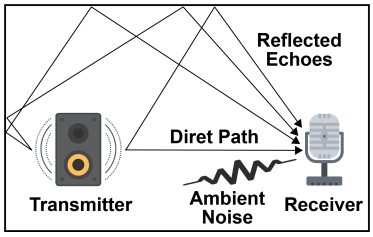


Fig. 3. Illustration of multi-path effect.

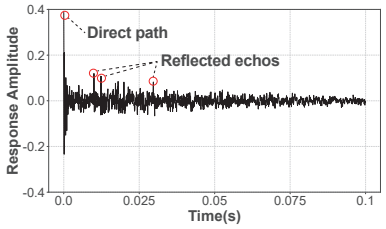


Fig. 4. Example of real-world RIR.

Specifically, unlike the additive perturbation scheme, i.e, $\mathbf{x}'_\mathbf{s} = \mathbf{x}_\mathbf{s} + \delta$, we convolve the perturbation δ and original voice $\mathbf{x}_\mathbf{s}$ to produce adversarial examples, i.e., $\mathbf{x}'_\mathbf{s} = \mathbf{x}_\mathbf{s} * \delta$. For each sampling point, we have:

$$\mathbf{x}'_\mathbf{s}(n) = \sum_{t=0}^{T} \mathbf{x}_\mathbf{s}(t) \cdot \delta(n-t). \tag{1}$$

This indicates that each sampling point of our adversarial example is derived from the weighted combination of the past sampling points of the original voice, thus leading to less artifact and distortion. Moreover, according to the property of the convolution operation, we have:

$$\mathrm{FFT}(\mathbf{x}_\mathbf{s} * \delta) = \mathrm{FFT}(\mathbf{x}_\mathbf{s}) \times \mathrm{FFT}(\delta), \tag{2}$$

which means the convolution in the time domain is equivalent to multiplication in the frequency domain. In other words, our convolutional perturbation is essentially a filter, which determines the significance of different frequency components in the original voice. This filter can serve as a promising approach to manipulating acoustic features (i.e., pitch and harmonics) in voices for de-identification. Then we introduce the real-world RIR h as a template to guide the construction of convolutional adversarial perturbations. Specifically, we penalize the perturbation change in the following optimization objective:

$$\mathcal{L}_{\mathrm{perturb}}(\delta) = \|\delta - h\|_p. \tag{3}$$

In our implementation, L_2 normalization is adopted according to empirical studies. By incorporating this objective into optimization, we reshape the convolutional adversarial perturbations as real-world RIRs to approximate a natural reverberation effect.

Benefiting from such RIR-like convolutional perturbations, we can achieve imperceptible voice de-identification, providing perceptually consistent voiceprint, integral speech, and audio quality for human participants.

3.3 Pseudo Target Sampler

To realize a sufficiently diverse de-identification, previous work collects and stores a pool of voices from real target speakers, whose speaker embeddings are also extracted and updated frequently. To reduce the storage and calculation overhead while providing diverse targets, we propose a lightweight pseudo target sampler to synthesize targets for guiding the perturbation construction.

In order to generate samples with desired speaker identity as targets, we employ a Conditional Variational Auto-Encoder (CVAE), for its strong ability to model continuous distribution and learn semantic representation. Moreover, to avoid redundant voice-embedding transformation, we design an embedding-level β-CVAE to serve as a pseudo target sampler in *VoiceCloak*. As shown in Figure 5, we first extract speaker embeddings from a open-source corpus with an extractor. Following the framework of mainstream ASIs [16, 52], the extractor converts the voice in the corpus to Mel Frequency Cepstral Coefficients (MFCC) [21] as acoustic features. The acoustic features are fed to a Time Delay Neural Network (TDNN) to build the frame-level temporal context,

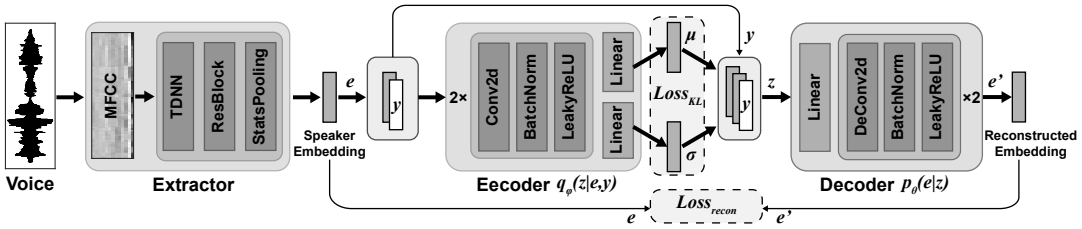


Fig. 5. Network architecture of the embedding-level β-CVAE.

and then input to residual blocks to model the global channel interdependences. Finally, a statistic pooling is performed to aggregate utterance-level features as the output speaker embedding. This embedding extraction serves as a data preparation process, where we employ a state-of-the-art ASI (Ecapa-TDNN [16]) as the extractor to generate a large number of speaker embeddings to construct an embedding dataset for the following β-CVAE training.

Taking the extracted speaker embedding $\mathbf{e}$ as input, the β-CVAE generates a new embedding $\mathbf{e}'$ through an encoder-decoder architecture, where the encoder q_ϕ squeezes the embedding $\mathbf{x}$ into a latent variable $\mathbf{z}$, and the decoder p_θ reconstructs a new embedding $\mathbf{e}'$ through deep neural networks. To learn the underlying identity semantics, the speaker embedding $\mathbf{e}$ is concatenated with the one-hot embedding of the corresponding identity label $\mathbf{y_t}$ and then fed to the encoder. The encoder stacks multiple down-sample blocks consisting of convolution layers and batch normalization with LeakyReLU activation. According to the variational interference, the encoder output $\mathbf{z}$ is assumed to follow a Gaussian distribution: $q_\phi(\mathbf{z}|\mathbf{e},\mathbf{y}) \sim \mathcal{N}(\mu, \sigma^2\mathbf{I})$, where the mean vector μ and the diagonal covariance $\sigma^2\mathbf{I}$ are derived from the following two parallel linear layers. After that, the latent variable $\mathbf{z}$ can be sampled from the distribution, which is reparameterized using the reparameterization tricks [32] in practice: $\mathbf{z} = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Similar to the encoder input, we also explicitly modulate the prior $\mathbf{y_t}$ into the latent variable $\mathbf{z}$ as the input of the decoder. The decoder reforms the latent vector by a linear layer, and then reconstructs a new embedding $\mathbf{e}'$ with multiple up-sample blocks, each of which consists of deconvolution layers and batch normalization with LeakyReLU activation. Finally, the entire generative model can be trained with:

$$\mathcal{L}_{\beta-CVAE} = \mathbb{E}[\|\mathbf{e} - \mathbf{e}'\|_2^2] + \beta KL(\mathcal{N}(\mathbf{z}|\mu, \sigma^2\mathbf{I})\|\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})), \tag{4}$$

where the first term is the reconstruction error, restricting the reconstruction bias of the output embedding, and the second term is the KL divergence regularization, forcing the latent space to approximate a continuous zero-mean unit-variance Gaussian. And the weight β is used to balance the reconstruction quality and the latent space continuity.

With the well-trained β-CVAE, *VoiceCloak* could generate diverse target speaker embeddings for adversarial perturbation construction. Specifically, given the desired identity label, the target embedding can be derived from: (1) *reconstruction*: the generator encodes the speaker embedding and then decodes the latent variable to a new embedding with the corresponding identity. (2) *sampling*: the generator directly samples a stochastic variable in the latent space and synthesizes a new embedding with only the decoder. (3) *interpolation*: the generator performs semantic interpolation between latent variables of different speakers to synthesize new embeddings of unreal speakers. As a result, *VoiceCloak* could sample diverse speaker embeddings with the desired identity from a learned Gaussian distribution without any voice input, or even create unreal embeddings as "pseudo speaker" for improving the identity diversity. Moreover, only the pre-trained decoder module is needed for actual deployment, thus significantly reducing the demand for computing and storage resources.

3.4 Marginal Triplet Optimizer

With the sampled pseudo target, we optimize the convolutional adversarial perturbations to impose on the user's raw voice for de-identification.

Typically, there are two different manners of adversarial perturbation optimization, i.e., the non-targeted and targeted manners. The non-targeted manner tends to yield highly-similar adversarial examples that may be re-identified by the adversary, while the targeted adversarial examples may still be linked to the source user identity due to the unknown enrollment set of adversary's ASI. Therefore, we introduce a triplet loss architecture [51] to combine the non-targeted and targeted manners for effective voice de-identification.

The core idea of triplet loss is to quantify the distance among samples in the latent space, and project the anchor and positive samples in the nearby region, whereas the anchor and negative samples are far away from each other. Specifically, we define the speaker embeddings of the source voice and the sampled target embedding (i.e., $f(\mathbf{x_s})$ and $\mathbf{e_t}$) as the negative and positive samples respectively, and regard the speaker embedding of adversarial example (i.e., $f(\mathbf{x_s} * \delta)$) as the anchor sample, where $f(\cdot)$ denotes the extractor to calculate speaker embedding. Then, we build a triplet and derive the marginal triplet loss:

$$\mathcal{L}_{\text{triplet}}(\delta) = \max\{D(f(\mathbf{x_s} * \delta), \mathbf{e_t}) - \kappa_1, 0\} + \max\{\kappa_2 - D(f(\mathbf{x_s} * \delta), f(\mathbf{x_s})), 0\}, \tag{5}$$

where $D(\cdot)$ refers to the distance metric of speaker embeddings (e.g., PLDA [15], cosine distance), κ_1 and κ_2 are distance margins. With this objective, we iteratively optimize the adversarial example in a gradient descent way under the explicit guidance of the triplet, enabling the user's identity to be close to another target speaker while far away from the source user to the greatest extent. Combining with the perturbation penalty loss in Equation (3), we derive the complete optimization objective:

$$\mathcal{L}(\delta) = \mathcal{L}_{\text{triplet}}(\delta) + \alpha \mathcal{L}_{\text{perturb}}(\delta), \tag{6}$$

where α is used to control the penalty weight. Finally, we summarize our adversarial perturbation construction process for voice de-identification in Algorithm 1 in Appendix A.2.

Since the adversarial example is constructed from a specific source and target as needed, such an input-specific manner empowers *VoiceCloak* any-to-any voice de-identification. On the one hand, this manner allows a source user to disguise a group of different speakers among different voices by sampling different pseudo targets to build triplets, further increasing the difficulty of identity linkage by ASIs. On the other hand, any source users could directly deploy and apply *VoiceCloak* without any additional enrollment, enabling a user-friendly experience.

4 EVALUATION

4.1 Experimental Setup

4.1.1 Speech Datasets. As shown in Table 1, we first train the pseudo target sampler of *VoiceCloak* on LibriSpeech(train) [43] with 251 target speakers. Besides, we enroll another 40 speakers from LibriSpeech(test) and 107 speakers from VCTK(v0.80) [14] as users of *VoiceCloak*, covering a wide range of various accents, professions, and ages. Hence, a total of 7,579 voices from the 147 users (66 males and 81 females) ranging 0.17s~34.96s are used to test ASI systems and construct adversarial examples. Note that the 147 source users and the 251 target speakers are disjoint, i.e., the actual users are completely unseen for our de-identification. In addition, we use real-world RIRs from REVERB2014 [33] for convolutional perturbation initialization.

4.1.2 ASI Systems. As shown in Table 2, we adopt multiple State-Of-The-Art (SOTA) speaker identification models as target ASI systems. Among them, the pre-trained DeepSpeaker [34] is derived from the unofficial implementation[1] due to the lack of original open-source code, while the X-Vector [52] and Ecapa-TDNN [16] are

[1] https://github.com/philipperemy/deep-speaker

pre-trained by SpeechBrain [49]. Moreover, we also evaluate *VoiceCloak* on a commercial ASI system iFLYTEK [29] through the provided HTTP API, whose implementation details are totally unknown to us.

4.1.3 Implementation Details. Table 4 in Appendix A.1 presents the network structure of β-CVAE. We extract speaker embeddings from LibriSpeech(train) for training the β-CVAE, where we set β=2 to optimize Equation (4) with an Adam optimizer (learning rate=0.001) for 30 epochs to derive a compact distribution. After the training, we only remain the well-trained decoder to sample target embeddings. Then we randomly select real-world RIRs to construct convolutional adversarial perturbations for each test voice according to Algorithm 1. To alleviate the reverberation, the length of RIR-like perturbation is set to 0.2s according to empirical study. We use cosine distance as $D(\cdot)$, and optimize Equation (6) with α=5000, κ_1=0.2, κ_2=0.8, η=0.001 for 200 steps until early stopping with a patient of 10. The well-crafted perturbations are normalized and convolved with the user's voices to produce de-identified samples.

4.1.4 Baselines. We compare *VoiceCloak* to SOTA works in voice de-identification, including the frequency warping-based VoiceMask [46], disentangled representation-based voice conversion (DisVC) [13] as well as a typical additive adversarial example method PGD [36]. Specifically, the warping factor $|\alpha|$ is sampled from [0.08,0.10] as recommended in VoiceMask, random targets from LibriSpeech(test) are selected for DisVC, and the perturbation scale is constrained below 0.01 for PGD and optimized in a non-targeted manner.

4.1.5 Metrics. We adopt multiple objective and subjective metrics to evaluate *VoiceCloak* in terms of privacy and utility:

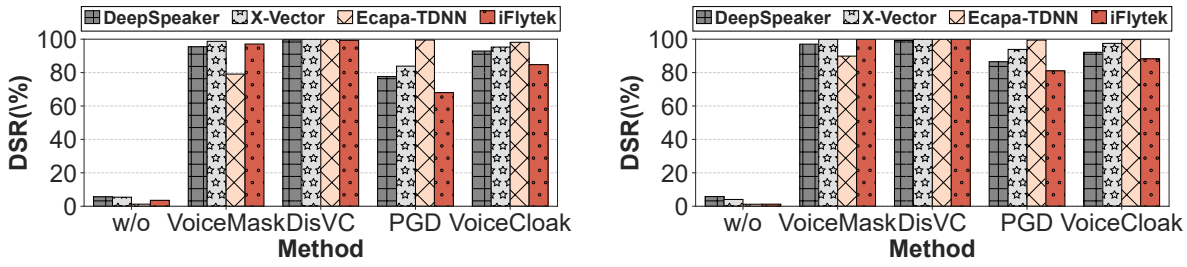
- *De-identification Success Rate (DSR)*: DSR = $\frac{X}{Y}$, where X and Y are the numbers of test voices and successful de-identification respectively.
- *Word Accuracy Drop (WAD)*: WAD = $\frac{C_a - I_a}{N} - \frac{C_b - I_b}{N}$, where N is the sentence length, C_b, I_b and C_a, I_a refer to the number of correct words and extra inserted words in the transcript before and after de-identification respectively. We employ an end-to-end ASR pre-trained by SpeechBrain to transcribe speech and calculate the WAD.
- *Mel Spectral Distortion (MCD)*: an objective audio distortion measurement that quantifies the distance between the MFCCs of the reference and target voices (mc_r, mc_t): $\frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{24}\|mc_r(i) - mc_t(i)\|_2}$. Typically, an MCD below 8dB is acceptable for ASR systems while between 4.5dB~6.0dB is needed for voice conversion systems [26].
- *Short-Time Objective Intelligibility (STOI)*: a widely used metric that is highly correlated with the intelligibility of the degraded version of the speech. A higher STOI indicates better speech intelligibility.
- *Mean Opinion Score (MOS)*: a numerical measure of the human-judged quality of speech with 5 levels: excellent(4~5), good(3~4), fair(2~3), poor(1~2) and bad(0~1).
- *Real-Time Ratio (RTR)*: RTR = $\frac{T_c}{T_d}$, where T_c and T_d are the time cost and voice duration respectively.

Table 1. Voice dataset statistics.

Dataset	#Speaker	#Utterance	Duration(s)
LibriSpeech(train)	251	27,952	3.00~24.53
LibriSpeech(test)	40	2,229	3.00~34.96
VCTK(v0.80)	107	5,350	0.17~19.28
REVERB2014	36 real-world RIRs		1s

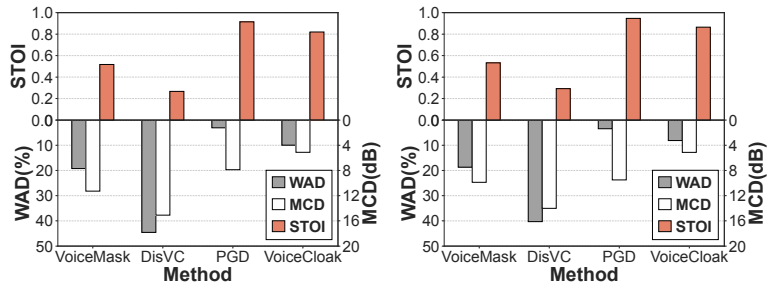
Table 2. SOTA ASI systems.

System	Architecture	Source
DeepSpeaker	ResCNN	Reproduction
X-Vector	TDNN	SpeechBrain
Ecapa-TDNN	SE-ResNet	SpeechBrain
iFlytek	unknown	iFlytek



(a) Librispeech. (b) VCTK.

Fig. 6. Comparison of voice de-identification between *VoiceCloak* and SOTA methods.



(a) Librispeech. (b) VCTK.

Fig. 7. Comparison of speech utility between *VoiceCloak* and SOTA methods.

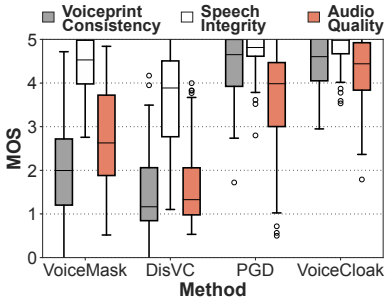


Fig. 8. MOS of voices de-identified by *VoiceCloak* and SOTA methods.

4.2 Overall Effectiveness

We first evaluate the overall performance of *VoiceCloak* on voice de-identification and speech utility. In the experiment, we perform de-identification on the 7,579 voices from the 147 users using *VoiceCloak* and other baselines and then feed both the original and de-identified voices to the four ASIs and the ASR for speaker identification and speech recognition respectively. Then we calculate the WAD, MCD, and STOI for each pair of original and de-identified voices.

As shown in Fig. 6 and Fig. 7, all the DSRs against the four ASIs on both LibriSpeech and VCTK are about 1%∼5%, exhibiting their powerful identification ability on the original voices. After voice de-identification, the DSRs of VoiceMask and DisVC against the four ASIs approach 100% except Ecapa-TDNN, due to their significant voiceprint manipulation. However, such an intrusive manipulation also lead to extremely high WADs of 18.70%∼44.56% and MCDs of 9.88dB∼15.09dB as well as low STOIs of 0.26∼0.53. This result indicates severe distortions and poor intelligibility, which would significantly impair the normal interaction between users and human participants. This result also suggests that existing de-identification methods overemphasize identity privacy while ignoring speech utility. By contrast, adversarial example-based PGD and *VoiceCloak* achieve slight lower but satisfactory DSRs while remaining WADs below 9.96 and MCDs below 9.50 as well as high STOIs over 0.81, inducing much less distortion and better intelligibility. Moreover, compared to PGD, *VoiceMask* achieves a better DSR and MCD but with a worse WAD and STOI. This also indicates that convolutional perturbations tend to have better effectiveness

against different ASIs while causing less audio signal distortions, and the additive perturbations tend to preserve the linguistic-related structure for better integrity and intelligibility.

4.3 Perturbation Non-Intrusiveness

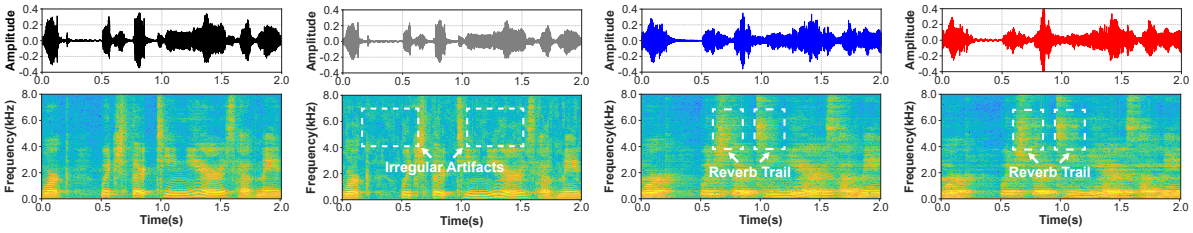
Since the objective metrics (e.g., WAD, MCD, STOI) can not fully reflect the real perception of human beings, we further conduct subjective experiments to evaluate the non-intrusiveness of *VoiceCloak* to human perception. We recruit 50 volunteers (28 males and 22 females) aged 18∼53, who have no hearing disease and are unaware of the specific de-identification techniques. Note that all the subjective experiments on volunteers are validated by the Institutional Review Board (IRB) in our university. We have these volunteers participate in our MOS test that includes a comparing trial and a distinguishing trial.

4.3.1 Comparing Trial. In this trial, the volunteers are asked to listen to 10 pairs of original voices and the corresponding de-identified voices from VoiceMask, DisVC, PGD, and *VoiceCloak* respectively, and then report their intuitive sense of voiceprint consistency, speech integrity, and audio quality respectively. The volunteers' opinions are recorded as a 5-level MOS and shown in Fig. 8. In terms of voiceprint consistency, We can see that the MOSs of VoiceMask and DisVC are distributed between 1∼3, which are much lower than those of PGD and *VoiceCloak*, indicating that adversarial example-based de-identification preserves perceptually consistent voiceprint. As for the speech integrity, the MOSs of PGD and *VoiceMask* approach 5, which also outperforms VoiceMask and DisVC. Moreover, *VoiceCloak* achieves the highest MOS in terms of audio quality, further validating the superior non-intrusiveness of convolutional perturbations.

4.3.2 Distinguishing Trial. In this trial, we first play an original voice for volunteers to refresh their impression and then play 10 original voices and 10 de-identified voices from VoiceMask, Disentangle-VC, PGD, and *VoiceCloak* in random order. For each voice, the volunteers need to determine whether it is original or not. If they regard the voice as not original, they are further asked to choose a reason for this from several options, i.e., *unnatural voiceprint*, *illegible text*, *distorted quality*, *obvious reverb* or provide any *other reasons* supporting their judgment. These options and reasons are described in Table 5 in Appendix A.3. We calculate the distribution of test voices for each reason, as shown in Table 3. We can see that over 45.23% de-identified voices from *VoiceCloak* while only 4.87%, 0.00% and 34.14% of VoiceMask, DisVC and PGD are regarded as original. Among the voices that are regarded as not original, 56.43% from VoiceMask and 36.17% from DisVC are considered to have unnatural voiceprint, degrading the voiceprint consistency for user experience. PGD performs well in voiceprint and text preservation but over 32.41% of voices are considered distorted in quality, and volunteers also reported that they could hear distinct electrical noises in 20.48% of voices from PGD in the *other reasons* option. Instead, *voiceCloak* well balances the voiceprint, text, and quality with only a natural reverberation that does not impair the user experience too much.

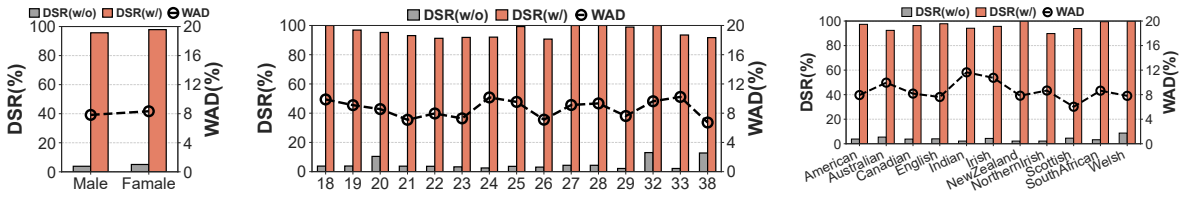
Table 3. Distribution(%) of distinguishing test on voices de-identified by *VoiceCloak* and SOTA methods.

Method	Regard as Original	Unnatural Voiceprint	Illegible Text	Distorted Quality	Obvious Reverb	Other
VoiceMask	4.87	**56.43**	4.83	20.95	12.89	0.00
DisVC	0.00	**36.17**	27.82	24.49	11.13	0.36
PGD	34.14	3.04	0.00	**32.41**	9.91	**20.48**
VoiceCloak	**45.23**	8.45	2.74	13.94	29.61	0.00



(a) Original. (b) De-identified by PGD. (c) Over-the-air. (d) De-identified by *VoiceCloak*.

Fig. 9. Example of original voice, original voice over the air, and de-identified voice by *VoiceCloak*.



(a) Gender. (b) Age. (c) Accent.

Fig. 10. DSR and WAD of *VoiceCloak* with different human voice characteristics.

4.3.3 Spectrum Visualization. To explicitly illustrate the non-intrusive de-identification of *VoiceCloak*, we present a visual comparison between original and de-identified voices. As shown in Fig. 9, compared with the original voice, the de-identified voice by PGD results in many irregular artifacts in the high-frequency components, which would sound like harsh electrical noises. For *VoiceCloak*, although it exhibits a distinct waveform and reverb trails in the spectrum, we can observe similar results in the over-the-air voice, which means our convolutional perturbations successfully approximate RIRs to imitate the natural distortion during over-the-air sound propagation. This would be perceived as reverberation and does not affect the perceptual quality very much. Hence, the RIR-like convolutional perturbations could maintain natural audibility and provide a better experience for human participants.

4.4 Impact Factors

Considering the various usage scenarios of *VoiceCloak* with different conditions, we further study the impact of human voice characteristics and external factors. For simplicity, we use X-Vector as the default ASI in this experiment.

4.4.1 Human Voice Characteristics. To verify the adaptability of *VoiceCloak* on various persons with different voice characteristics, we further study the impact of gender, age, and accent. Specifically, we count the DSR and WAD of 5,350 pairs of original and de-identified voices from 107 users (46 males and 61 females) in VCTK, covering 15 ages and 11 accents. As shown in Fig. 10(a), the WAD of both genders are similar while the DSR of females is slightly higher than males. As for the age shown in Fig. 10(b), *VoiceCloak* realizes high DSRs without significant difference among ages while only sacrificing about 6.7%~10.2% of speech recognition performance. Besides, from Fig. 10(c), we can see that *VoiceCloak* also provides effective de-identifi-cation across different accents with a low WAD below 12%. These results demonstrate the adaptability of *VoiceCloak* to a variety of human voice characteristics.

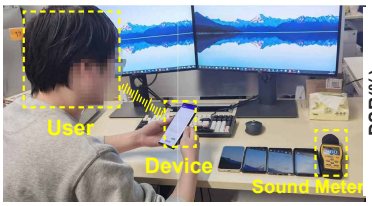
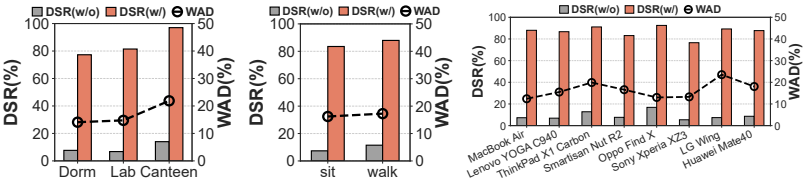
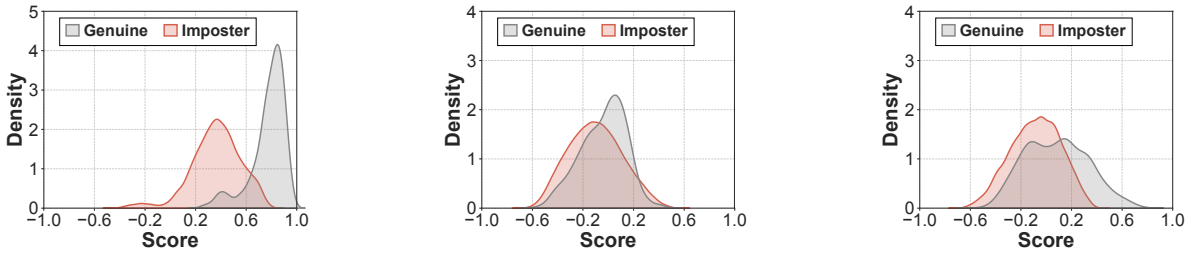


Fig. 11. Experimental setup of human study (Lab).



(a) Environment. (b) State. (c) Device.

Fig. 12. DSR and WAD of *VoiceCloak* with different external factors.



(a) Original. (b) De-identified. (c) Re-identified.

Fig. 13. Score distribution of genuine and imposter trials.

4.4.2 External Factors. Apart from the analysis on the open-source corpus, we also conduct a human study to explore the impact of different environments, devices and motions. We recruit 10 volunteers (6 males, 4 females) as our users, and each of them is asked to speak 5 English sentences to 3 laptops (MacBook Air, Lenovo YOGA C940, and ThinkPad X1 Carbon) and 5 smart phones (Smartisan Nut R1, Oppo Find X, Sony Xperia XZ3, LG Wing, and Huawei Mate40). As shown in Fig. 11, we conduct the experiments in three environments including a dorm, a meeting room and a canteen with ambient noise levels of 44.9dBSPL, 50.8dBSPL and 71.4dBSPL respectively (measured by a sound level meter). The users hold the device at a comfortable distance of 0.2~0.3m while sitting still or walking slowly, whose voice is recorded by the device and then sent to *VoiceCloak* for de-identification. We collect 2,400 voices in total and calculate the DSR with and without de-identification as well as the WAD. As shown in Fig. 12(a), the original voices can be well identified with DSRs of 7.67% and 6.79% in the dorm and lab respectively, while there is a DSR degradation of about 6.5% in the noisier canteen. After de-identification, the DSR of the dorm and lab is around 80% with a similar WAD of about 14%, while the DSR approaches 97% in the canteen with a higher WAD over 20%, indicating that the ambient noise contributes to voice de-identification but also distorts speech. As shown in Fig. 12(b) and Fig. 12(c), although there are some differences in DSR and WAD of different user states and device models, *VoiceCloak* still greatly increases the DSR with an acceptable WAD.

4.5 Unlinkability Analysis

To validate the unlinkability of speaker identity, we further investigate the score distribution and analize the embedding visualization of de-identified voices.

4.5.1 Score Distribution. We present the score distribution on X-Vector of *Genuine* and *Imposter* trials on the original and de-identified voices, i.e., the log-likelihood ratios between same-speaker and different-speaker

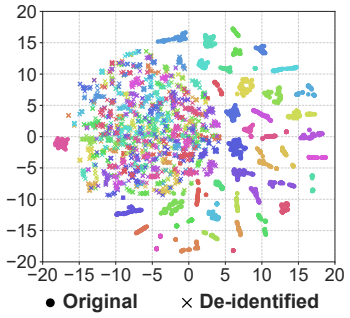


Fig. 14. T-SNE visualization of speaker embedding.

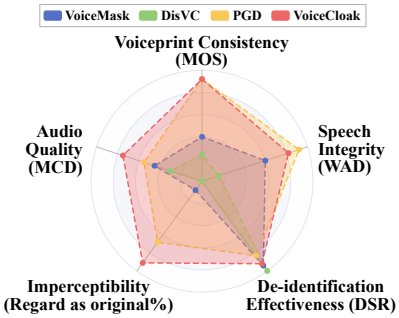


Fig. 15. Comparison between *VoiceCloak* and SOTA methods.

hypotheses [56]. Here we further consider the re-identification attack [55] that aims to link the de-identified voice to the source user. Specifically, the adversary applies the same voice de-identification operations as *VoiceCloak* on the enrolled raw voices and then tries to find out the real identity of the incoming de-identified voices through analyzing their similarity with the enrolled ones. As shown in Fig. 13, the genuine and imposter score distributions of the original voices are separated from each other, so that the ASI can distinguish them. But after our de-identification with diverse pseudo identities, the score distribution of genuine and imposter trials overlap even under re-identification, thus hard to be distinguished and linked by advanced adversaries.

4.5.2 Embedding Visualization. To understand the contribution of identity diversity to de-identification, we also apply t-SNE to visualize the speaker embeddings of the original and de-identified voices from the 40 LibriSpeech users. As shown in Fig. 14, the de-identified voices are far away from the original voices and mapped to diverse targets instead of clustering together, further increasing the difficulty for the adversary to link the de-identified voice to the original user. Hence, *VoiceCloak* can realize diverse identity transformation and robust de-identification even under the re-identification attack.

4.6 Resource Cost

4.6.1 Computation Overhead. We implement *VoiceCloak* on the 8 devices to evaluate the computation and storage overhead. Specifically, we perform de-identification on 1,000 voices with varying length of 3s∼30s, and record the speech duration and the corresponding CPU time to derive the average RTR. We find that the RTR on MacBook Air is 0.93 while those on less powerful Lenovo YOGA C940 and ThinkPad X1 Carbon are 1.34 and 1.65. Due to the limited computing resource and power consumption, the RTRs on the 5 phones are even higher 5.71∼11.38). Through further analysis, we find that the RTR is roughly proportional to the voice duration. Considering the excellent speech and speaker consistency of convolutional perturbations, we slice each voice into multiple 3s segments and then perform a *streaming* de-identification. Based on this, we finally reduce the RTRs into 0.54, 0.95, 1.17 on the laptops and 3.35∼6.87 on the phones without loss of effectiveness. Besides, we also find that the target sampler takes up little CPU time, while the triplet optimizer is the most expensive one (over 95%). This observation verifies the efficiency of our target sampling design, and encourages us to optimize a more efficient perturbation construction in the future, i.e., generative construction [62] rather than iterative optimization, which would further improve the real-time performance of *VoiceCloak*.

4.6.2 Storage Occupancy. As for the storage cost, instead of preparing hundreds of megabytes of target voices, *VoiceCloak*'s target sampler only occupies 13.03M while providing 251 real identities and numerous pseudo identities. Note that this scale can be further expanded by means of the powerful representation learning of CVAE.

Thanks to such a lightweight decoder, *VoiceClock* significantly reduces the storage requirement and enables practical deployment on different devices.

4.7 Privacy-utility Balance

To better understand the privacy-utility balance of *VoiceCloak*, we compare it with SOTA methods in terms of de-identification effectiveness and voice utility using both objective and subjective metrics. Instead of calculating a simple weighted average on multiple metrics [4], we present the overall performance comparison in a radar figure. As summarized in Fig. 15, VoiceMask and DisVC achieve excellent de-identification to support reliable identity privacy protection. However, due to the intrusive voiceprint manipulation, both of them result in poor voice consistency, speech integrity, and audio quality and are obviously perceptible to human participants, thus largely degrading the voice utility. Benefiting from the strength of adversarial examples, PGD and *VoiceCloak* realize comparative de-identification performance and a significant improvement in voice utility. Moreover, additive adversarial perturbation-based PGD presents a slight advantage in preserving speech integrity, but still induces unsatisfactory audio quality and can be easily perceived by human listeners. Especially in the distinguishing test, over 20% volunteers report that they can hear obvious electronic noises from the PGD de-identified samples. This is due to the high-frequency artifacts produced by additive perturbations, which also impair the voice utility. By contrast, *VoiceCloak* utilizes a convolutional adversarial perturbation to approximate natural reverberation, which not only de-identifies voices effectively but also remains excellent voiceprint consistency, speech integrity, and audio quality. More importantly, human listeners can not easily distinguish the de-identified voices by *VoiceCloak* from original samples, further demonstrating its superior non-intrusiveness. To sum up, *VoiceCloak* does not simply pursue the ultimate de-identification performance but pays more attention to balancing the privacy and utility for improving user experience in practical usage.

5 DISCUSSION

Deployment mode in practice. *VoiceCloak* acts as a voice filter at the user side and can be deployed in two modes. The first mode is to integrate *VoiceCloak* into the Operating System (OS) as a basic facility for the full control of all voice input, which allows trusted APPs to access raw voices while feeding untrusted ones the de-identified voices only. This ensures a system-level security guarantee of users' voiceprints but requires additional OS modification. In the second mode, *VoiceCloak* is installed as a voice input plugin on the users' input editor, which performs de-identification every time the voice input interface is invoked. This requires users to install *VoiceCloak* in advance and authorize corresponding permissions.

Possible extension to physical layer. In this work, we focus on speaker de-identification at the digital layer of voice transmission without any hardware equipment. This enables a user-friendly experience during online meetings, social interaction, instant messaging, etc. But it is also possible to extend *VoiceCloak* to the physical layer, where users' voiceprints suffer from exposure by stealthy eavesdropping. In this case, except for taking additional hardware devices for emitting adversarial perturbations, *VoiceCloak* also needs to cope with several key issues: (1) Input-agnostic de-identification. Physical-layer protection needs to process live-streaming voices instead of recorded utterances, so we cannot observe the entire input during de-identification, requiring universal adversarial perturbations that can generalize on voices with different linguistic texts. (2) Live-streaming processing. To prevent live-streaming voices from eavesdropping in real-time, all the de-identification operations need to be done within an extremely short time window, thus raising a higher demand for system efficiency. (3) Channel interference. The physical injection also requires robust adversarial perturbations that are resistant to channel interference. We leave the physical-layer extension a research topic and explore potential solutions to these issues in the future.

Compatibility with voiceprint authentication. Despite voiceprint-irrelevant voice services (e.g., speech recognition), users may also rely on voiceprint-based systems, such as WeChat voiceprint lock [61] and Siri personalized activation [6]. In these cases, these users have to offer their clean voiceprints for accurate authentication, but *VoiceCloak* may hinder such normal usages. Hence, a compatibility mechanism between de-identification and authentication is needed. Considering the fixed wake texts for authentication services usually, a straightforward solution is to shut down *VoiceCloak* temporarily when detecting these specific texts, and automatically restore the protection as soon as the authentication is finished. Another solution may ensure more comprehensive voiceprint protection, where we can transform users' voiceprints to the same target speaker for enrollment and later activation or unlocking. This could enable de-identification and authentication simultaneously.

6 RELATED WORK

Signal processing-based voice de-identification. Early studies [2, 30, 38, 59] on voice de-identification mainly exploit signal processing techniques (e.g., frequency warping, amplitude scaling, duration warping) to modify the spectral and prosodic features for manipulating the in-depth voiceprint. However, these methods require parallel source and target voices to train the voice transformation, i.e., two voices with the same texts and timestamps, limiting its utility in practice. To address this issue, the following works [37, 44] propose to pre-calculate a set of voice transformations between multiple pairs of source and target speakers for online de-identification. But the voices produced by this method are filled with obvious artifacts due to the synthetic speakers. Hence, a new paradigm based on Vocal Tract Length Normalization (VTLN) [17] is proposed to realize voice transformation without parallel corpus. Specifically, VLTN-based works [46, 47, 64] stretch or compress the voice spectrum frame by frame according to a warp function, and synthesize de-identified voices through a waveform vocoder. However, the invertible warping function used in these approaches is probably employed to recover the original voice by informed adversaries [55]. This design indicates the intrinsic vulnerability for voice de-identification.

Deep learning-based voice de-identification. With the advances in artificial intelligence, numerous studies turn to explore voice de-identification based on learning methods. One branch of them [3, 31, 47] integrates speech-to-text and text-to-speech techniques to transform voices into texts and then re-synthesize de-identified voices, which eliminates identify information but introduces unacceptable overhead. Another branch [20, 27, 42, 54, 58] proposes X-Vector-based voice conversion schemes, which replaces the source embedding with the target one in the latent space of X-Vector. Using a vocoder or a neural source filter, these works could synthesize anonymized voices with satisfactory quality. However, a pool of speaker voices needs to be pre-collected, and complex target selection strategies are mandatory to ensure the de-identification performance. These lead to difficulties in practical deployment on users' resource-limited devices. More recent studies [19, 53] propose to learn de-identified speech representation using adversarial training at the user side, but require a redesign of existing service architecture.

Unlike these voiceprint modification or speech re-synthesizing methods, our work aims to design a non-intrusive and robust voice de-identification system using novel convolutional adversarial examples.

7 CONCLUSION

This paper presents a voice de-identification system, *VoiceCloak*, which turns adversarial examples as a defense tool against automatic speaker identification to balance the privacy and utility of voice services. By modulating convolutional adversarial perturbations into the natural reverberation, we realize non-intrusive de-identification with consistent voiceprint, integral speech, and excellent quality. We also design a lightweight conditional variational auto-encoder to generate diverse targets and construct input-specific perturbations through a triplet loss architecture, enabling any-to-any identity conversion for robust de-identification. Experimental results show *VoiceCloak* could effectively de-identify users on mainstream and commercial speaker identification systems, achieving a privacy-utility balance.

ACKNOWLEDGMENTS

This research is sponsored by National Key R&D Program of China (2020AAA0107700), National Natural Science Foundation of China (62102354, 62032021, 62122066, 62172359, 61972348, 62172277), Fundamental Research Funds for the Central Universities (2021FZZX001-27).

REFERENCES

[1] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin R. B. Butler, and Joseph Wilson. 2019. Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems. In *Proceedings of NDSS*. San Diego, California, USA.

[2] Mohamed Abou-Zleikha, Zheng-Hua Tan, Mads Græsbøll Christensen, and Søren Holdt Jensen. 2015. A discriminative approach for speaker selection in speaker de-identification systems. In *Proceedings of IEEE EUSIPCO*. Nice, France, 2102–2106.

[3] Shimaa Ahmed, Amrita Roy Chowdhury, Kassem Fawaz, and Parmesh Ramanathan. 2020. Preech: A System for Privacy-Preserving Speech Transcription. In *Proceedings of USENIX Security*. Virtual Event, 2703–2720.

[4] Rawan Alharbi, Mariam Tolba, Lucia C. Petito, Josiah D. Hester, and Nabil Alshurafa. 2019. To Mask or Not to Mask?: Balancing Privacy with Visual Confirmation Utility in Activity-Oriented Wearable Cameras. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3 (2019), 72:1–72:29.

[5] Alibaba Cloud. 2017. Voiceprint Recognition System — Not Just a Powerful Authentication Tool. https://alibaba-cloud.medium.com/voiceprint-recognition-system-not-just-a-powerful-authentication-tool-6b3702b5c5a.

[6] Apple. 2022. Apple Siri. https://machinelearning.apple.com/research/personalized-hey-siri.

[7] Sourav Bhattacharya, Dionysis Manousakas, Alberto Gil C. P. Ramos, Stylianos I. Venieris, Nicholas D. Lane, and Cecilia Mascolo. 2020. Countering Acoustic Adversarial Attacks in Microphone-equipped Smart Home Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2 (2020), 73:1–73:24.

[8] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *Proceedings of IEEE S&P*. San Jose, CA, USA, 39–57.

[9] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. 2021. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems. In *Proceedings of IEEE S&P*. San Francisco, CA, USA, 694–711.

[10] Meng Chen, Li Lu, Zhongjie Ba, and Kui Ren. 2022. PhoneyTalker: An Out-of-the-Box Toolkit for Adversarial Example Attack on Speaker Recognition. In *Proceedings of IEEE INFOCOM*. London, United Kingdom, 1419–1428.

[11] Meng Chen, Li Lu, Jiadi Yu, Yingying Chen, Zhongjie Ba, Feng Lin, and Kui Ren. 2022. A non-intrusive and adaptive speaker de-identification scheme using adversarial examples. In *Proceedings of ACM MobiCom*. Sydney, NSW, Australia, 853–855.

[12] Qianniu Chen, Meng Chen, Li Lu, Jiadi Yu, Yingying Chen, Zhibo Wang, Zhongjie Ba, Feng Lin, and Kui Ren. 2022. Push the Limit of Adversarial Example Attack on Speaker Recognition in Physical Domain. In *Proceedings of ACM SenSys*. Boston, Massachusetts, 710–724.

[13] Ju-Chieh Chou and Hung-yi Lee. 2019. One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. In *Proceedings of ISCA Interspeech*. Graz, Austria, 664–668.

[14] Veaux Christophe, Yamagishi Junichi, and MacDonald Kirsten. 2016. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. https://datashare.ed.ac.uk/handle/10283/2119.

[15] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4 (2010), 788–798.

[16] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proceedings of ISCA Interspeech*. Shanghai, China, 3830–3834.

[17] Ellen Eide and Herbert Gish. 1996. A parametric approach to vocal tract length normalization. In *Proceedings of IEEE ICASSP*. Atlanta, Georgia, USA, 346–348.

[18] Thorsten Eisenhofer, Lea Schönherr, Joel Frank, Lars Speckemeier, Dorothea Kolossa, and Thorsten Holz. 2021. Dompteur: Taming Audio Adversarial Examples. In *Proceedings of USENIX Security*. 2309–2326.

[19] Fernando M. Espinoza-Cuadros, Juan M. Perero-Codosero, Javier Antón-Martín, and Luis A. Hernández Gómez. 2020. Speaker De-identification System using Autoencoders and Adversarial Training. *CoRR* abs/2011.04696 (2020). arXiv:2011.04696

[20] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas W. D. Evans, and Jean-François Bonastre. 2019. Speaker Anonymization Using X-vector and Neural Waveform Models. *CoRR* abs/1905.13561 (2019). arXiv:1905.13561

[21] Haytham M. Fayek. 2016. Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between. https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.

[22] Forbes. 2021. Apple Just Gave 1.5 Billion iPad, iPhone Users A Reason To Leave. https://www.forbes.com/sites/gordonkelly/2022/02/12/apple-iphone-ipad-siri-audio-recordings-iphone-privacy/?sh=68fc85bd4193.

[23] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *Proceedings of ICLR*. San Diego, CA, USA.
[24] Google. 2022. Google Meet. https://www.bluestacks.com/apps/communication/google-meet-on-pc.html.
[25] Google Privacy & Terms. 2022. How Google Voice works. https://policies.google.com/technologies/voice?hl=en-US.
[26] CMU Speech Group. 2012. Statistical parametirc sythesis and voice conversion techniques. http://festvox.org/11752/slides/lecture11a.pdf.
[27] Yaowei Han, Sheng Li, Yang Cao, Qiang Ma, and Masatoshi Yoshikawa. 2020. Voice-Indistinguishability: Protecting Voiceprint In Privacy-Preserving Speech Data Release. In *Proceedings of IEEE ICME*. London, UK, 1–6.
[28] Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian J. McAuley, and Farinaz Koushanfar. 2021. WaveGuard: Understanding and Mitigating Audio Adversarial Examples. In *Proceedings of USENIX Security*. 2273–2290.
[29] iFLYTEK Open Platform. 2022. Voiceprint Recognition. https://www.xfyun.cn/service/isv.
[30] Qin Jin, Arthur R. Toth, Tanja Schultz, and Alan W. Black. 2009. Voice convergin: Speaker de-identification by voice transformation. In *Proceedings of IEEE ICASSP*. Taipei, Taiwan, 3909–3912.
[31] Tadej Justin, Vitomir Struc, Simon Dobrisek, Bostjan Vesnicer, Ivo Ipsic, and France Mihelic. 2015. Speaker de-identification using diphone recognition and speech synthesis. In *Proceedings of IEEE FG*. Ljubljana, Slovenia, 1–7.
[32] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proceedings of ICLR*. Banff, AB, Canada.
[33] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka. 2016. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP J. Adv. Signal Process.* 2016 (2016), 7.
[34] Lantian Li, Yixiang Chen, Ying Shi, Zhiyuan Tang, and Dong Wang. 2017. Deep Speaker Feature Learning for Text-Independent Speaker Verification. In *Proceedings of ISCA Interspeech*. Stockholm, Sweden, 1542–1546.
[35] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. 2020. AdvPulse: Universal, Synchronization-free, and Targeted Audio Adversarial Attacks via Subsecond Perturbations. In *Proceedings of ACM CCS*. Virtual Event, USA, 1121–1134.
[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of ICLR*. Vancouver, BC, Canada.
[37] Carmen Magariños, Paula Lopez-Otero, Laura Docío Fernández, Eduardo Rodríguez Banga, Daniel Erro, and Carmen García-Mateo. 2017. Reversible speaker de-identification using pre-trained transformation functions. *Comput. Speech Lang.* 46 (2017), 36–52.
[38] Carmen Magariños, Paula Lopez-Otero, Laura Docío Fernández, Eduardo R. Banga, Carmen García-Mateo, and Daniel Erro. 2016. Piecewise linear definition of transformation functions for speaker de-identification. In *Proceedings of IEEE SPLINE*. Aalborg, Denmark, 1–5.
[39] Microsoft. 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.
[40] Microsoft. 2022. Microsoft Teams. https://www.microsoft.com/en-us/microsoft-teams/group-chat-software.
[41] Microsoft Azure Congnitive Service. 2022. Speaker recognition. https://azure.microsoft.com/en-us/services/cognitive-services/speaker-recognition/.
[42] Seyed Hamidreza Mohammadi and Alexander Kain. 2017. An overview of voice conversion systems. *Speech Commun.* 88 (2017), 65–82.
[43] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of IEEE ICASSP*. South Brisbane, Queensland, Australia, 5206–5210.
[44] Miran Pobar and Ivo Ipsic. 2014. Online speaker de-identification using voice transformation. In *Proceedings of IEEE MIPRO*. Opatija, Croatia, 1264–1267.
[45] Popular Mechanics. 2018. Hundreds of Apps Can Eavesdrop Through Phone Microphones to Target Ads. https://www.popularmechanics.com/technology/security/a14533262/alphonso-audio-ad-targeting/.
[46] 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 ACM SenSys*. Shenzhen, China, 82–94.
[47] Jianwei Qian, Feng Han, Jiahui Hou, Chunhong Zhang, Yu Wang, and Xiang-Yang Li. 2018. Towards Privacy-Preserving Speech Data Publishing. In *Proceedings of IEEE INFOCOM*. Honolulu, HI, USA, 1079–1087.
[48] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. In *Proceedings of ICML*, Vol. 97. Long Beach, California, 5231–5240.
[49] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit. *CoRR* abs/2106.04624 (2021). arXiv:2106.04624
[50] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. 2019. Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. In *Proceedings of NDSS*. San Diego, California, USA.

[51] Matthew Schultz and Thorsten Joachims. 2003. Learning a Distance Metric from Relative Comparisons. In *Proceedings of NIPS*. Vancouver and Whistler, British Columbia, Canada, 41–48.
[52] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *Proceedings of IEEE ICASSP*. Calgary, AB, Canada, 5329–5333.
[53] Brij Mohan Lal Srivastava, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. 2019. Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?. In *Proceedings of ISCA Interspeech*. Graz, Austria, 3700–3704.
[54] Brij Mohan Lal Srivastava, Natalia A. Tomashenko, Xin Wang, Emmanuel Vincent, Junichi Yamagishi, Mohamed Maouche, Aurélien Bellet, and Marc Tommasi. 2020. Design Choices for X-Vector Based Speaker Anonymization. In *Proceedings of ISCA Interspeech*. Virtual Event, Shanghai, China, 1713–1717.
[55] Brij Mohan Lal Srivastava, Nathalie Vauquier, Md. Sahidullah, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. 2020. Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers. In *Proceedings of IEEE ICASSP*. Barcelona, Spain, 2802–2806.
[56] Brij Mohan Lal Srivastava, Nathalie Vauquier, Md. Sahidullah, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. 2020. Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers. In *Proceedings of IEEE ICASSP*. Barcelona, Spain, 2802–2806.
[57] The New York Times. 2019. Amazon's Alexa Never Stops Listening to You. Should You Worry? https://www.nytimes.com/wirecutter/blog/amazons-alexa-never-stops-listening-to-you/.
[58] Henry Turner, Giulio Lovisotto, and Ivan Martinovic. 2022. Generating identities with mixture models for speaker anonymization. *Comput. Speech Lang.* 72 (2022), 101318.
[59] Tavish Vaidya and Micah Sherr. 2019. You Talk Too Much: Limiting Privacy Exposure Via Voice Input. In *Proceedings of IEEE S&P Workshops*. San Francisco, CA, USA, 84–91.
[60] Qing Wang, Pengcheng Guo, and Lei Xie. 2020. Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition. In *Proceedings of ISCA Interspeech*. Virtual Event, 4228–4232.
[61] Wechat Official. 2015. Voiceprint: The New WeChat Password. https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password.
[62] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. 2021. Enabling Fast and Universal Audio Adversarial Attack Using Generative Model. In *Proceedings of AAAI*. Virtual Event, 14129–14137.
[63] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter. 2018. CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition. In *Proceedings of USENIX Security*. Baltimore, MD, USA, 49–64.
[64] Guanglin Zhang, Sifan Ni, and Ping Zhao. 2020. Enhancing Privacy Preservation in Speech Data Publishing. *IEEE Internet Things J.* 7, 8 (2020), 7357–7367.
[65] Yuxuan Zhou, Huangxun Chen, Chenyu Huang, and Qian Zhang. 2022. WiAdv: Practical and Robust Adversarial Attack against WiFi-based Gesture Recognition System. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2 (2022), 92:1–92:25.
[66] Zoom. 2022. One platform to connect. https://zoom.us/.

A APPENDIX

A.1 Network Structure of β-CVAE

As shown in Table 4, the input D-dimensional speaker embedding is concatenated with a D-dimensional one-hot embedding of identity label. Such a combined embedding is squeezed to D/4 by the encoder through two down-sample blocks, each of which includes a 2D convolution layer. After a following full-connected layer, the encoder learns the D/2-dimensional mean and covariance vector through two parallel linear layers. Through the reparameterization trick, we derive the D/2-dimensional latent vector and concatenate it with the one-hot embedding. After that, the decoder reforms the latent vector to 16D-dimensional with a full-connected layer, and then expand it through two up-sample blocks.

A.2 Perturbation Construction Algorithm

We summarize the whole convolutional adversarial perturbation construction process in Algorithm 1.

A.3 MOS Test Descriptions

In the distinguishing trial, if the volunteers regard the voice as not original, we further require them to select a reason from several options or provide their own evidence. Table 5 shows the detailed description.

Table 4. Network structure of β-CVAE.

Module	Block	Input→Output	Layer Specification
Encoder	Down-sample Block	(2,D)→(32,D/2)	Conv2d(channel=32, kernel=(2,6), stride=(1,2))
	Down-sample Block	(32,D/2)→(64,D/4)	Conv2d(channel=64, kernel=(1,4), stride=(1,2))
	Full-connected Layer	(1,16D)→(1,D)	Linear(in_features=16D,out_features=D)
	Mean Vector	(1,D)→(1,D/2)	Linear(in_features=D,out_features=D/2)
	Covariance Vector	(1,D)→(1,D/2)	Linear(in_features=D,out_features=D/2)
Decoder	Latent Vector	(1,3D/2)→(1,D)	Linear(in_features=3D/2,out_features=D)
	Full-connected Layer	(1,D)→(1,16D)	Linear(in_features=D,out_features=16D)
	Up-sample Block	(64,D/4)→(32,D/2)	DeConv2d(channel=32, kernel=(1,4), stride=(1,2))
	Up-sample Block	(32,D/2)→(1,D)	DeConv2d(channel=1, kernel=(1,6), stride=(1,2))

Algorithm 1 Adversarial Perturbation Construction

Input: Extractor $f(\cdot)$ with distance metric $D(\cdot)$, voice of source speaker $\mathbf{x_s}$, label of target speaker $\mathbf{y_t}$, real-world RIR h, pre-trained β-CVAE decoder $d(\cdot)$, penalty weight α, distance margins κ_1, κ_2, learning rate η
Output: Well-crafted convolutional perturbation δ

Normalize the RIR: $h \leftarrow \frac{h}{\|h\|}$
Initialize the perturbation: $\delta \leftarrow h$
Extract the source embedding: $f(\mathbf{x_s})$
Sample a target embedding: $\mathbf{e_t} \leftarrow d(\mathbf{y_t})$
for each step **do**
  Extract the adversarial embedding: $f(\mathbf{x_s} * \delta)$
  Calculate the marginal triplet loss: $\mathcal{L}_{\text{triplet}}(\delta)$
  Calculate the perturbation penalty loss: $\mathcal{L}_{\text{perturb}}(\delta)$
  $\mathcal{L}(\delta) \leftarrow \mathcal{L}_{\text{triplet}}(\delta) + \alpha\mathcal{L}_{\text{perturb}}(\delta)$
  $\delta \leftarrow \delta - \eta\nabla_\delta\mathcal{L}(\delta)$
end for

Table 5. Description of options and reasons in the distinguishing trial.

Option	Description
Yes	The test audio is the original voice.
No	The test audio is not the original voice.
Reason	**Description**
Unnatural Voiceprint	The voiceprint of the test audio sounds synthetic but not real human voice.
Illegible Text	The text of the test audio is corrupted and hard to recognize.
Distorted Quality	The quality of the test audio is distorted.
Obvious Reverb	There are reverb echos in the test audio.
Other Reason	Any other reasons from volunteers.