V-Cloak: Intelligibility-, Naturalness- & Timbre-Preserving Real-Time Voice Anonymization

Jiangyi Deng, Fei Teng, and Yanjiao Chen, *Zhejiang University;* Xiaofu Chen and Zhaohui Wang, *Wuhan University;* Wenyuan Xu, *Zhejiang University*

https://www.usenix.org/conference/usenixsecurity23/presentation/deng-jiangyi-v-cloak

V-Cloak: Intelligibility-, Naturalness- & Timbre-Preserving Real-Time Voice Anonymization

Jiangyi Deng
Zhejiang University

Fei Teng
Zhejiang University

Yanjiao Chen
Zhejiang University

Xiaofu Chen
Wuhan University

Zhaohui Wang
Wuhan University

Wenyuan Xu
Zhejiang University

Abstract

Voice data generated on instant messaging or social media applications contains unique user voiceprints that may be abused by malicious adversaries for identity inference or identity theft. Existing voice anonymization techniques, e.g., signal processing and voice conversion/synthesis, suffer from degradation of perceptual quality. In this paper, we develop a voice anonymization system, named V-Cloak, which attains real-time voice anonymization while preserving the intelligibility, naturalness and timbre of the audio. Our designed anonymizer features a one-shot generative model that modulates the features of the original audio at different frequency levels. We train the anonymizer with a carefully-designed loss function. Apart from the anonymity loss, we further incorporate the intelligibility loss and the psychoacoustics-based naturalness loss. The anonymizer can realize untargeted and targeted anonymization to achieve the anonymity goals of unidentifiability and unlinkability.

We have conducted extensive experiments on four datasets, i.e., LibriSpeech (English), AISHELL (Chinese), CommonVoice (French) and CommonVoice (Italian), five Automatic Speaker Verification (ASV) systems (including two DNN-based, two statistical and one commercial ASV), and eleven Automatic Speech Recognition (ASR) systems (for different languages). Experiment results confirm that V-Cloak outperforms five baselines in terms of anonymity performance. We also demonstrate that V-Cloak trained only on the VoxCeleb1 dataset against ECAPA-TDNN ASV and DeepSpeech2 ASR has transferable anonymity against other ASVs and cross-language intelligibility for other ASRs. Furthermore, we verify the robustness of V-Cloak against various de-noising techniques and adaptive attacks. Hopefully, V-Cloak may provide a cloak for us in a *prism* world.

1 Introduction

Voiceprint is a critical biometric that can uniquely identify a person. As massive personal data is collected and processed by online services, there are rising concerns for privacy leakage. In 2018, the European Union enforced the General Data Protection Regulation (GDPR) [1] for personal data protection, especially for biometric data. However, an avalanche of voice data is generated daily on social media (e.g. Facebook/Meta, WeChat, TikTok) and in communication applications (e.g. Zoom, Slack, Microsoft Teams, Ding Talk), and automated processing methods, e.g., ASV, can easily extract voiceprint for ill use. For example, as shown in Figure 1, an adversary may infer the speaker identity of a private conversation from voice messages uploaded to the cloud with an ASV [19, 22, 28]. Therefore, there is an urgent demand for voice anonymization to help users protect voiceprint while enjoying voice-related services (e.g., speech recognition by ASR) and interpersonal communication (e.g., human listeners can identify the speaker).

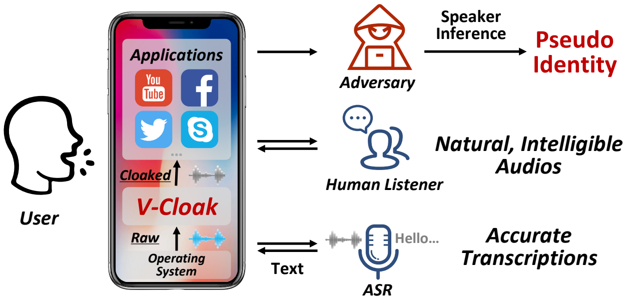

Figure 1: Voiceprint in voice data may be leveraged by malicious adversaries for identity inference or identity theft. The raw audio is cloaked with V-Cloak before being passed to applications, thus malicious service providers or third parties can only obtain a pseudo identity/voiceprint.

Existing voice anonymization methods are mainly based on voice signal processing (SP), voice conversion (VC) and voice synthesis (VS). SP [34, 53] methods directly apply signal processing techniques to modify speaker-related features in audios to obscure voiceprints. Nonetheless, SP-based voice anonymization usually induces large quality degradation as intelligibility and naturalness are not considered.

Table 1: V-Cloak versus existing works.

Method	**Type***	**Intelligibility#**	**Naturalness#**	**Timbre-preserving**	**Real-Time Coef.↓**	**User-agnostic†**
VoiceMask [38, 39]	VC	✗	✗	✗	0.041	✓
Yoo [56]	VC	✗	✓	✗	*N.K.*	✗
NSF [14, 16]	VS	✓	✓	✗	0.110	✗
HFGAN [29]	VS	✓	✓	✗	0.104	✓
Justin [21]	VS	✓	✓	✗	*N.K.*	✓
McAdams [34]	SP	✗	✗	✗	0.030	✓
Vaidya [53]	SP	✗	✗	✗	*N.K.*	✓
V-Cloak **(Ours)**	**Adv**	✓	✓	✓	**0.011**	✓

(i) *: Voice Conversion (VC). Voice Synthesis (VS). Signal Processing (SP). Adversarial examples (Adv). (ii) #: whether the method has explicit constraints on intelligibility or naturalness. (iii) ↓: Real-time coefficient (RTC), the ratio between the processing time and the duration of the audio. **The lower the RTC, the more efficient the method.** We measure the five methods under the same computing resource conditions. *N.K.*, not known, the authors did not evaluate the efficiency of their methods or make their codes available. (iv) †: whether the method needs to be trained for a new user.

VC [38, 39, 48, 56] and VS [14, 16, 21, 29] methods convert the original audio into another audio that sounds completely different from the original speaker. Although VC and VS may achieve anonymity, they are not suitable for scenarios where the user wants to hide their identity from ASVs but hopes to preserve their personal timbre to human audiences, e.g., posts of celebrities on social media, voice messages with acquaintances.

In this paper, we make the first attempt to design a real-time voice anonymization system, named V-Cloak, which achieves anonymity while preserving intelligibility, naturalness, and timbre of the audios. A comparison of V-Cloak with existing works is shown in Table 1. Nonetheless, to realize these design goals with a practical real-time system is challenging in three aspects.

- *How to achieve real-time voice anonymity against adaptive attacks?*

Different from traditional signal processing and voice conversion & synthesis, we are inspired by the adversarial examples that can trick ASV into misidentifying the speaker but induce imperceptible differences to the human auditory system. Nonetheless, directly applying adversarial examples to voice anonymization has two major issues. First, most of the existing ASV adversarial examples [7, 11, 26, 27, 58] are constructed via iterative updates, which cannot achieve real-time voice anonymization. As far as we are concerned, there is only one ASV adversarial attack named FAPG that creates adversarial examples using a one-shot generative model [55]. Unfortunately, FAPG needs to train a feature map for each potential target speaker and the original paper only evaluates for an ASV with 10 speakers. Furthermore, the adversary may be informed of the anonymization method and the model (anonymizer), and then launches an adaptive attack to de-anonymize the anonymized audio.

To tackle these problems, we adapt a lightweight generative model Wave-U-Net [49] for V-Cloak. We equip Wave-U-Net with two novel components, i.e., *VP-Modulation* and *Throttle*. *VP-Modulation* modulates the feature elements of the original audio at each frequency level according to the voiceprint of a target speaker. *Throttle* adjusts the weights of features of the original audio at different frequency levels to conform to the constraint on the anonymization perturbations. The trained anonymizer can produce anonymized audios targeting any speaker/voiceprint under any anonymization perturbation constraint without re-training. Furthermore, we conduct theoretical analysis and experiments to verify the anonymity of V-Cloak in the case of adaptive attacks.

- *How to maintain objective and subjective intelligibility of anonymized audios?*

It is desirable for the anonymized audios to be intelligible to ASRs (objective intelligibility) such that the users can still enjoy speech-to-text services; and to humans (subjective intelligibility) such that voice messages can be understood. However, SP- and VC-based anonymization, as well as voice adversarial examples, do not consider intelligibility constraint and may introduce noises that greatly degrade intelligibility.

To address this issue, we impose an intelligibility loss when training the anonymizer. The intelligibility loss is based on the decoding error rate of the ASR. Instead of the commonly-used Connectionist Temporal Classification (CTC) loss of ASR, we acquire the graphemic posteriorgram (GPG) loss, which preserves the full alignment of the transcription and the grapheme of each frame. The subjective intelligibility is achieved by constraining the anonymization perturbations by our proposed *Throttle* module and better masking the anonymization perturbations based on psychoacoustics.

- *How to preserve naturalness and timbre of anonymized audios?*

Naturalness and timbre preservation are important to human audiences or listeners of anonymized audios. Signal processing and existing ASV adversarial examples did not consider naturalness such that the processed audios may sound mechanical. In addition, signal processing, voice conversion and voice synthesis all distort the timbre of the original speaker such that the anonymized audio sounds unlike being spoken by the original speaker (e.g., a friend or a celebrity).

To cope with this problem, we introduce a naturalness & timbre loss when training the anonymizer based on the psychoacoustic theory of masking effects. Our user study verifies that the anonymized audios of V-CLOAK receive high naturalness and timbre scores.

We implement a fully-functional prototype of V-CLOAK, evaluated with extensive experiments on five ASVs (anonymity) and eleven ASRs (intelligibility) with datasets of four languages (English, Chinese, French, Italian). The comparison with five baselines demonstrates that V-CLOAK achieves the best anonymization performance with the second-best intelligibility performance. Cross-language experiments show that the anonymizer of V-CLOAK trained on one ASV and one ASR can be transferred to other ASVs and ASRs (with different languages). A user study with 102 volunteers confirms the intelligibility-, naturalness- and timbre-preserving properties of V-CLOAK.

We summarize our main contributions as follows.

- We propose V-CLOAK, an intelligibility-, naturalness- and timbre-preserving voice anonymization system. V-CLOAK is proved and evaluated to fulfil the anonymization goals of unidentifiability and unlinkability against naive and adaptive adversaries.

- We develop a real-time anonymizer that transforms the original audio into targeted or untargeted anonymized audios. The anonymizer is trained with anonymity, intelligibility, naturalness and timbre loss, generalizing to any new original speaker or new target speaker without the need for re-training.

- We conduct extensive experiments to verify the effectiveness and efficiency of V-CLOAK under various testing conditions and a user study to confirm the practicality and applicability of V-CLOAK.

2 Background

2.1 Voice Data

In the digital world, voice data of a user is massively generated and distributed for various purposes, e.g., communications via voice messages or video posts on social media. These wildly exposed voice data may be easily collected by service providers or third parties. For instance, Facebook is collecting audio data from voice messages on its social network platform, and even attempts to transcribe the content of these private messages [22]. TikTok revised its privacy policy to legitimize faceprints and voiceprints collection from the videos uploaded by users, and even claimed the possibility of data sharing for business purposes [28].

Voice data contains two kinds of information, i.e., speech contents and phonetic features.

- Speech contents. Speech contents refer to the linguistic information contained in the voice data, i.e., "what are the words spoken." Speech contents determine the intelligibility of the voice data.

- Phonetic features. Phonetic features refer to the way the speech contents are conveyed in the voice data, i.e., "how are the words spoken." Phonetic features affect the timbre of the voice data.

Voiceprint is a phonetic feature that can uniquely identify a speaker. However, voiceprint contained in voice data may be abused for identity inference or identity theft. On the one hand, voiceprint may be used to infer the identity of speakers of a private conversation by automatic speaker verification (ASV) systems. On the other hand, the voiceprint of a speaker may be extracted from audios to synthesize audios to pass voiceprint-based authentication systems. For example, WeChat, a popular messaging app in China, allows users to login via voiceprint [3]. In face of these potential privacy leakages, it is essential for users to anonymize voice data before sending voice messages or publishing videos on social media.

2.2 Psychoacoustics

Psychoacoustics is the study of the relationship between subjective psychological perceptions (e.g., perceived volume, pitch) and objective physical parameters (e.g., sound pressure level, frequency) [57]. The masking effect is one of the most common psychoacoustic phenomena [15]. There are two forms of masking: temporal and spectral. Temporal masking refers to the situation where a sound cannot be perceived if a sudden louder sound appears immediately preceding or following the first one. The louder sound is called the *masker*. Spectral masking refers to the imperceptibility of a sound component due to other frequency components played simultaneously. The *perception threshold* of this component varies due to both sound signals (e.g., frequency) and the listener. We leverage the spectral masking effect to make anonymization perturbations more imperceptible to human users.

2.3 Automatic Speaker Verification & Automatic Speech Recognition

An Automatic Speaker Verification (ASV) system aims to deduce the speakers of audios based on their voiceprints.

Speaker inference via an ASV system includes the enrollment phase and the inference phase. In the enrollment phase, clean audio samples of the speaker to be recognized are fed into the ASV such that the voiceprint can be extracted and stored in the ASV. In the inference phase, the ASV takes an audio sample as input and outputs whether the input audio belongs to the enrolled speaker. There are two mainstream methods of extracting and matching voiceprints, i.e., statistical models and Deep Neural Network (DNN)-based models. Gaussian mixture model (GMM) is a traditional statistical model to extract ivector voiceprints. ivector-PLDA is a popular ASV implementation that matches ivector voiceprints via probabilistic linear discriminant analysis (PLDA). X-vector is a DNN-based voiceprint extractor, which outperforms GMM as DNNs are more effective in extracting feature representations from large-scale voice datasets. ECAPA-TDNN [10] is the state-of-the-art ASV implementation using end-to-end training, i.e., training the front-end and the back-end jointly as an integrated network [54].

An Automatic Speech Recognition (ASR) system aims to transcribe the speech contents from audio samples (without the need to know the speaker). In the training process, audios are first transformed into a sequence of spectral frames. Each frame is then transformed into a feature vector. Commonly used features include Filter Bank (FBank) [43], Mel-Frequency Cepstral Coefficients (MFCC) [30], Spectral Subband Centroid (SSC) [51] and Perceptual Linear Predictive (PLP) [17]. Then, the posterior probability of the lingueme (e.g., phoneme, grapheme, or word) contained in each frame is estimated. The linguemes are usually represented as tokens. For example, 29 tokens are used for the English language, i.e., letters a∼z, space, apostrophe and the special blank token ϕ. Next, the Connectionist Temporal Classification (CTC) module sums the probability of all possible alignments that reduce to the ground-truth sequence. For example, a three-frame sequence of $[a\ b\ \phi]$, $[a\ \phi\ b]$ and $[\phi\ a\ b]$ will all be reduced to the ground-truth sequence of $[a\ b]$. Finally, the model is updated to increase the probability of producing the ground-truth sequence. In the inference phase, a language model may be used to provide a prior probability to find the lingueme sequence of the highest probability.

2.4 Voice Anonymization

Voice anonymization refers to the practice of removing voiceprint from voice data. A voice anonymization system needs to satisfy various requirements to fulfil different purposes. Regarding the digital voice data privacy, we aim to achieve the following performance goals.

Anonymity. An anonymized audio should not reveal the identity of the speaker. More specifically, we consider concealing speaker identities in the digital domain from ASVs.

Intelligibility. An anonymized audio should be intelligible to both humans and ASRs. More specifically, the speech contents of the anonymized audio can be correctly understood by humans and transcribed by ASRs.

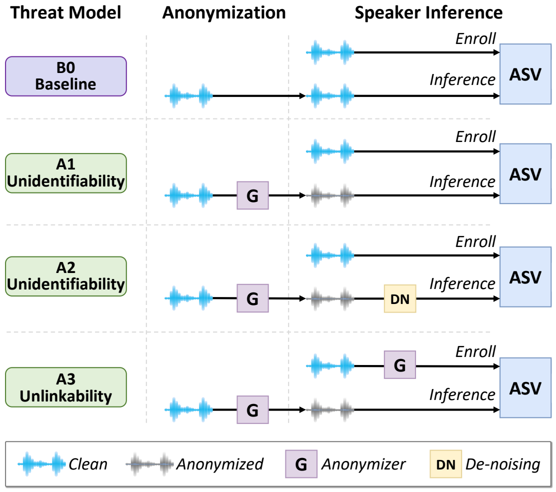

Figure 2: Threat model. **A1:** *ignorant* adversary who enrolls clean audios into the ASV and feeds anonymized audio into the ASV to infer the speaker. **A2:** *semi-informed* adversary who enrolls clean audios into the ASV and feeds de-noised anonymized audio into the ASV to infer the speaker. **A3:** *informed* adversary who enrolls anonymizer-processed audios into the ASV and feeds the anonymized audio into the ASV to infer the speaker.

Naturalness & Timbre. An anonymized audio should sound natural and like the timbre of the original speaker to humans. Studies show that most people find highly mechanical audios irritating and discomforting to listen to, thus natural-sounding anonymized audios are more user-friendly [50]. For voice messages and video posts on social media, it is ideal to make the anonymized audios sound authentic as the original speaker to audiences, especially for communications between acquaintances and publicity of celebrities.

Voice anonymization can be realized in various ways, as summarized in Table 1.

Voice signal processing. Signal processing techniques attempt to contort the voiceprint by directly modifying the voice signals in terms of formant positions, pitch, tempo, or pause [34, 53]. Though simple and fast, signal processing may degrade the intelligibility and naturalness of audios.

Voice conversion & synthesis. Voice conversion & synthesis techniques aim to replace the voiceprint of the original speaker in an audio with the voiceprint of another speaker [14, 16, 21, 29, 38, 39, 48, 56]. Voice conversion & synthesis preserve the intelligibility and naturalness, but alter the timbre so that the audio sounds unlike the original speaker. This may reduce the authenticity of voice messages to acquaintances and video posts of celebrities.

Voice adversarial examples. Adversarial example attacks against ASVs add imperceptible noises to audios such that the ASV cannot recognize the speaker [7, 11, 26, 27, 55, 58]. Adversarial perturbations can be generated in two ways.

- *Iterative optimization.* Optimization-based methods formulate the problem of adversarial perturbation generation as a constrained optimization problem [7, 11, 26, 27, 58]. As the formulated optimization problems are usually NP-hard, the solutions can only be approximated through iterative updates, which is quite time-consuming. Therefore, iterative optimization methods cannot be applied to real-time services.

- *One-shot generative model.* Generative models can be trained to produce adversarial perturbations in one shot. Commonly used generative models include Generative Adversarial Networks (GAN) and autoencoders [35, 36, 47]. As far as we know, there is only one study on generative model-based adversarial examples against ASV, named FAPG [55]. However, FAPG mainly focuses on deceiving ASVs but not preserving intelligibility and naturalness of audios.

2.5 Threat Model

We define the threat model in terms of the adversary's knowledge and capability, then we elaborate the performance goals of voice anonymization under the defined threat model. As shown in Figure 2, we consider three kinds of adversaries, i.e., *ignorant* (A1), *semi-informed* (A2), and *informed* (A3), with different knowledge and capabilities.

Knowledge. The adversary has an anonymized audio whose speaker is unknown. The adversary has collected a few clean samples of a pool of potential speakers to help with identity inference. Adversary A1 does not know that the audio is anonymized. Adversary A2 knows that the audio is anonymized but does not know the specific anonymizer. Adversary A3 has full knowledge of the anonymizer.

Capability. Adversary A1, A2, and A3 can use any ASVs to infer the speaker of the anonymized audio. As shown in Figure 2, A1 and A2 enroll potential speakers in the ASV using clean audios, and A3 enrolls potential speakers in the ASV using audio samples processed by the anonymizer. In the inference phase, A1 directly feeds the anonymized audio into the ASV; A2 applies de-noising methods to the anonymized audio and feeds the de-noised audio into the ASV; A3 also directly feeds the anonymized audio into the ASV.

In the face of the knowledge and the capability of adversaries, we further elaborate the goal of achieving anonymity regarding different types of adversaries. More specifically, in the case of *ignorant* and *semi-informed* adversaries, the speaker of the anonymized audio should be unidentifiable, and in the case of *informed* adversaries, the speaker and the anonymized audio should be unlinkable.

Unidentifiability. For A1 and A2 who enroll clean voiceprints into the ASV, the speaker of an anonymized audio should not be identified during the inference phase.

Unlinkability. For adversary A3 who enrolls anonymizer-processed voiceprints into the ASV, the speaker of an anonymized audio should be undistinguishable from other speakers.

3 Problem Formulation

Before delving into the design details of V-Cloak, in this section, we formally formulate the voice anonymization as a constrained optimization problem.

Given an audio sample $x = [x_1, \cdots, x_D] \in \mathbb{R}^{1\times D}$, where $\mathbb{R}^{1\times D}$ is a D-dimensional real number field, and D is the length of the audio. Without loss of generality, we assume $x_i \in [-1, 1]$. We aim to obtain an anonymized audio $\tilde{x}$ such that the ASV cannot match the voiceprint of $\tilde{x}$ with that of x. Let $\mathcal{V} : \mathbb{R}^{1\times \cdot} \to \mathbb{R}^{1\times N}$, denote the voiceprint extraction function that outputs a voiceprint of a fixed length N, and $\mathcal{G} : \mathbb{R}^{1\times D} \to \mathbb{R}^{1\times D_o}$ denote the anonymizer function.

<u>**Basic Formulation:**</u>

$$\min_{\mathcal{G}} \; L_{\text{ASV}}$$
$$\text{s.t. } \; \|\tilde{x} - x\|_\infty \le \varepsilon \text{ and } x, \tilde{x} \in [-1, 1],$$

where

$$L_{\text{ASV}} = \begin{cases} \mathcal{S}\big(\mathcal{V}(\tilde{x}), \mathcal{V}(x)\big), & \text{untargeted anonymization,} \\ -\mathcal{S}\big(\mathcal{V}(\tilde{x}), v\big), & \text{targeted anonymization.} \end{cases}$$
$$\tilde{x} = \begin{cases} \mathcal{G}(x), & \text{untargeted anonymization,} \\ \mathcal{G}(x, v), & \text{targeted anonymization.} \end{cases} \tag{1}$$

where ε constrains the l_∞ norm difference between x and $\tilde{x}$, $\mathcal{S}(\cdot,\cdot)$ is the scoring function measuring the similarity between the voiceprints of x and $\tilde{x}$, and v is the voiceprint of a speaker other than x. With untargeted anonymization, the voiceprint of the anonymized audio is diverted from that of the original audio as much as possible, which guarantees *unidentifiability*, i.e., the voiceprint of the anonymized audio will not match the voiceprint of the original audio. With targeted anonymization, the voiceprints of two anonymized audios with different original speakers but the same target speaker v will both be matched with v (thus be matched together), which guarantees both *unidentifiability* and *unlinkability*. We theoretically analyze the *unidentifiability* and the *unlinkability* of targeted and untargeted anonymizations in Appendix A and perform corresponding evaluations in §5.

The anonymized audio obtained by Equation (1) satisfies the basic goal of anonymity but may suffer from quality degradation in terms of intelligibility, naturalness and timbre. To tackle this problem, we equip the basic optimization problem with loss terms that address the performance goals of intelligibility, naturalness and timbre preservation. More specifically, we introduce an ASR-related loss term, which maintains the

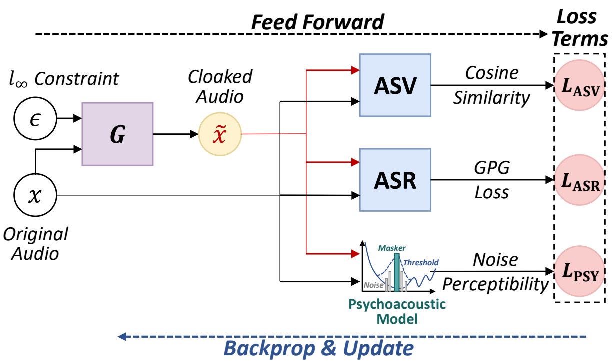

Figure 3: Architecture of V-CLOAK. The anonymizer G produces the anonymized audio $\tilde{x}$ given the original audio x and the threshold ϵ. G is trained to minimize the loss function related to the performance goals of anonymity, intelligibility, naturalness and timbre.

intelligibility of the anonymized audio for ASR. We also add a psychoacoustic-related loss term and an l_2-norm loss term to improve naturalness of the anonymized audio. Overall, the refined optimization problem is

V-CLOAK **Formulation:**

$$\begin{aligned} &\min_{\mathcal{G}} \; L_{\text{ASV}} + \alpha \cdot L_{\text{ASR}}(x,\tilde{x}) + \beta \cdot L_{\text{PSY}}(x,\tilde{x}) + \gamma \cdot \|\tilde{x}-x\|_2, \\ &\text{s.t. } \|\tilde{x}-x\|_\infty \le \epsilon \text{ and } x,\tilde{x} \in [-1,1], \end{aligned} \tag{2}$$

where parameters α, β, γ balance the trade-off among the performance goals.

4 V-CLOAK: Design Details

The optimization problem in Equation (2) is difficult to solve directly. Therefore, we propose a framework named V-CLOAK to derive the solution of Equation (2) based on a generative model. As shown in Figure 3, the main component of V-CLOAK is the anonymizer $\mathcal{G}$. $\mathcal{G}$ takes the original audio x and the threshold ϵ as inputs, and creates the anonymized audio in one shot. We first introduce the model architecture of $\mathcal{G}$, and then elaborate the training process of $\mathcal{G}$.

4.1 Anonymizer Design

To realize real-time voice anonymization, we create anonymized audios through one-shot generation instead of iterative updates. We develop a generative model-based anonymizer based on Wave-U-Net [49], as shown in Figure 4.

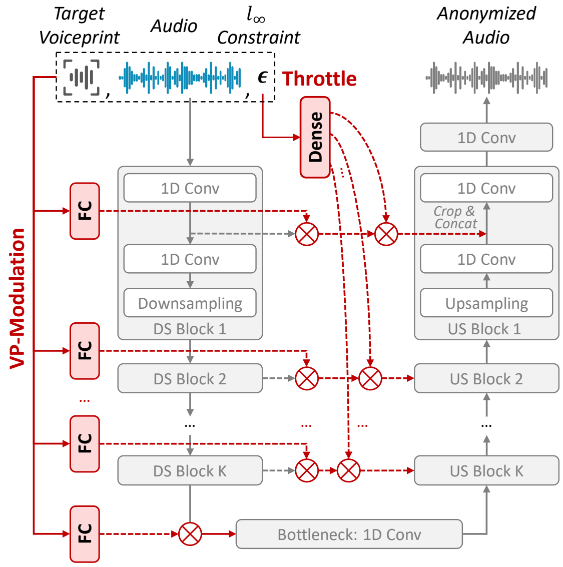

Figure 4: The design of the anonymizer $\mathcal{G}$. *VP-Modulation* modulates the elements in the feature vector extracted by each downsampling (DS) block (the same frequency level) based on the target voiceprint. *Throttle* adjusts features at different frequency levels according to the constraint ϵ.

Wave-U-Net is originally used for audio source separation [49]. The vanilla Wave-U-Net is U-shaped with the downsampling (DS) and the upsampling (US) sub-networks, as illustrated by the grey blocks in Figure 4. The input audio passes through a sequence of DS blocks, where a deeper DS block extracts a longer feature vector at a lower frequency level. There is a shortcut that transports the output of the first convolutional layer in each DS block to the final convolutional layer in each US block to combine the features at different frequency levels.

The anonymizer of V-CLOAK innovates Wave-U-Net in two ways with the *VP-Modulation* and the *Throttle* modules as shown in Figure 4.

4.1.1 VP-Modulation

Wave-U-Net in the previous work [55] creates targeted ASV adversarial examples by converting the audio of the original speaker towards a target speaker with a *feature map* of the target speaker inserted in the *bottleneck* layer at the bottom of the Wave-U-Net. Unfortunately, this suffers from two limitations. First, the feature map of every potential target speaker needs to be trained from scratch with relatively high overhead. The feature map needs to be replaced if another speaker is targeted, thus targeting an untrained speaker is not possible. Second, the feature map resides on the bottom of the Wave-U-Net, which means that the feature map represents the feature at the lowest frequency level. This lowest-frequency feature is too coarse-grained to capture the distinctive traits of different speakers, limiting the ability of the model to target a larger pool of speakers.

To address this limitation, we design *VP-Modulation* to guide the audio conversion process by the target voiceprint. As shown in Figure 4, at each frequency level, the target voiceprint v is transformed by a fully-connected layer (FC) into a modulation vector with a dimension consistent with the output at each shortcut and the output of the last DS block. The modulation vector rescales the shortcut features extracted by each DS block at each frequency level. In this way, the trained bottleneck layer needs no modification for a new target. In addition, the voiceprint of any target speaker can be fed into the network to realize a targeted anonymization without the need for re-training.

4.1.2 Throttle

Wave-U-Net in the previous work [55] imposes a fixed constraint on the converted audio during training to ensure that the difference between the converted audio and the original audio is beneath a threshold ε. However, during the voice anonymization phase, the threshold ε cannot be flexibly changed according to different requirements.

To cope with this problem, we design *Throttle* to learn to adapt anonymization perturbations under different constraints during training. In particular, *Throttle* takes constraint ε as input, and computes a K-dimensional adjustment vector, where K is the number of DS/US blocks. The adjustment vector controls the magnitude of each shortcut feature during combination. The output perturbation will be clipped to conform to the l_∞-norm constraint of ε, and back-propagate the loss to the *Throttle* to alter the adjustment vector. In this way, the *Throttle* learns the optimal adjustment vector under different l_∞-norm constraints. Note that the modulation vector produced by *VP-Modulation* weights the elements in the feature vector at a specific frequency level, and the adjustment vector produced by *Throttle* weights the feature vectors at different frequency levels.

4.2 Anonymizer Training

To train the anonymizer $\mathcal{G}$ to fulfil the performance goals, we materialize the loss function in Equation (2) as follows.

4.2.1 Anonymity Loss

In L_{ASV}, we utilize cosine similarity to measure the resemblance of two voiceprints $\mathcal{S}(\cdot,\cdot)$. In general, the voiceprint extraction function $\mathcal{V}$ can follow any ASV. In our experiments, we utilize the state-of-the-art DNN-based ASV, ECAPA-TDNN [10], to encode varied-length audios into fixed-length voiceprints, $\mathcal{V}: \mathbb{R}^{1\times\cdot} \to \mathbb{R}^{1\times 192}$. We demonstrate that the anonymized audio is transferable to different ASVs in our experiments.

For untargeted anonymization, minimizing L_{ASV} pushes the voiceprint of $\tilde{x}$ away from that of x, such that the voiceprint of the anonymized audio cannot be matched with that of the original audio, i.e., unidentifiability is guaranteed. For targeted anonymization, minimizing L_{ASV} drives the voiceprint of $\tilde{x}$ towards the target voiceprint v, such that the anonymized audio cannot be matched with that of the original audio when a proper v is selected, i.e., unidentifiability is guaranteed. Furthermore, for targeted anonymization, voiceprints of anonymized audios of two different speakers will be matched together (both matched with the same voiceprint v), i.e., unlinkability is guaranteed. A theoretical proof of unidentifiability and unlinkability is provided in Appendix A.

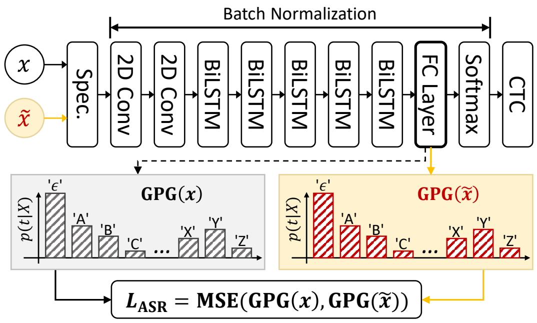

Figure 5: Intelligibility Loss. Above is the structure of DeepSpeech2. Graphemic posteriorgram (GPG) output by the FC layer is used to constrain the linguistic distortion, and GPG loss is measured with the MSE between the original audio and the cloaked audio.

4.2.2 Intelligibility Loss

For a similar reason, L_{ASR} can be instantiated with any ASR. In our experiments, we adopt DeepSpeech2 [4] with two 2D convolutional layers, five bidirectional LSTM layers, a fully-connected layer and a softmax layer, as shown in Figure 5. DeepSpeech2 uses graphemes, i.e., the smallest functional unit in a writing system in linguistics, as tokens. For the English language, the last softmax layer outputs the posterior probability of 29 tokens, i.e., a-z, space, apostrophe and the special ϕ token. The ϕ token indicates 'blank' or no label.

Connectionist temporal classification (CTC) loss is commonly used in ASR to train a sequence-to-sequence model when the alignment between the input spectral-frame sequence and the output token sequence is unknown. However, the CTC loss does not preserve the exact grapheme sequence. For instance, when the input is a three-frame audio, the final fully-connected layer in DeepSpeech2 outputs $[a\ \phi\ b]$ and $[a\ b\ \phi]$, where a, b and ϕ are the tokens of each frame. The CTC loss will ignore the blank token and reduce both sequences to $[a\ b]$.

To better improve intelligibility, we utilize another loss term instead, i.e., the graphemic posteriorgram (GPG) loss. As shown in Figure 5, GPG loss is based on the posterior probability output by the softmax layer (we use the output of the preceding fully connected layer in practice). Therefore, we have $L_{\text{ASR}}(x, \tilde{x}) = \text{MSE}(\text{GPG}(x), \text{GPG}(\tilde{x}))$, where $\text{MSE}(\cdot,\cdot)$ is the mean squared error. We evaluate the effects of the CTC loss and the GPG loss in our ablation study in §5.7.

4.2.3 Naturalness & Timbre Loss

We ensure naturalness and timbre preservation of the anonymized audio based on the psychoacoustic theory of masking effect. In particular, we leverage the spectral masking effect as the original audio and the anonymization perturbations are played at the same time. We treat the original audio as the *masker*, which masks the presence of anonymization perturbations.

To materialize L_{PSY}, we first compute the *masking threshold* [40] of the original audio, $\mathcal{M}(x)$, an F-dimensional vector, in which each element represents the maximum tolerable perturbation at a certain frequency level (a total of F frequency components). Then L_{PSY} is computed as the sum of excesses of the perturbations in the anonymized audio.

$$L_{\text{PSY}}(\tilde{x},x) = \min\{\psi,\ \frac{1}{F}\max\{0,\ \text{PSD}(\tilde{x}-x) - \mathcal{M}(x)\}\}, \tag{3}$$

where $\text{PSD}(\cdot)$ computes the log-magnitude power spectral density, and $\max\{0,\cdot\}$ preserves the positive but not negative parts of a function. We constrain the magnitude of the loss by ψ, since we find in our experiments that the L_{PSY} term is unstable and of large variance, especially in the early stage of training. Note that we further add an l_2-norm $\|\tilde{x}-x\|_2$ in the loss function in order to limit the energy of the anonymization perturbations.

5 Evaluation

5.1 Experiment Setup

Prototype. We have implemented a prototype of V-Cloak on the PyTorch [33] platform and trained the model according to Equation (2) using two NVIDIA 3090 GPUs. We set the default configuration as $D = 41,641$, $D_o = 32,089$, $N = 192$, $F = 1,025$, $K = 5$, and a batchsize of 64. The l_∞-norm constraint, ε, is sampled from a normal distribution $\mathcal{N}(\mu,\sigma)$ with the mean $\mu = 0.05$ and the variance $\sigma = 0.05$. Note that we randomize ε, the constraint value of l_∞ to train the *Throttle* module in V-Cloak to learn to adjust the magnitude of each feature under different constraints. In the training phase, we use an Adam [23] optimizer to update the parameters of the anonymizer $\mathcal{G}$ for 50 epochs, with a learning rate of 4e-4. The default adversary is A1.

Dataset. We train V-Cloak on VoxCeleb1 [31], an English dataset with 352-hour audios from 1,251 speakers. Four widely-used datasets are adopted to evaluate the effectiveness of V-Cloak, i.e., LibriSpeech (English) [32], AISHELL (Chinese) [6], CommonVoice (French) [5], and CommonVoice (Italian) [5]. We use the test sets of the four datasets for evaluation. Note that CommonVoice datasets are used for ASR with no credible speaker identity information, so we only use them to test whether the anonymized audios maintain intelligibility. More details about the datasets are listed in Table 2.

Table 2: Dataset statistics.

Dataset	Subset	#Speaker	#Utterance	Duration (s)
LibriSpeech (English)	*test-clean*	40	2,620	1.3 ∼ 35.0
AISHELL (Chinese)	*test*	20	7,176	1.9 ∼ 14.7
CommonVoice (French)	*test*	-*	5,000†	1.6 ∼ 11.5
CommonVoice (Italian)	*test*	-*	5,000†	3.2 ∼ 11.4

(i) *: CommonVoice datasets have no credible speaker identity information.
(ii) †: We use the first 5,000 utterances in CommonVoice for evaluation.

Table 3: ASVs used for evaluation.

Model	Alias	Category	Source	EER†(%)
ECAPA-TDNN	**EP**	DNN-based	SpeechBrain	0.70
X-vector	**XV**	DNN-based	SpeechBrain	6.53
GMM-UBM	**GMM**	Statistical	Kaldi	11.39
ivector-PLDA	**IV**	Statistical	Kaldi	6.03
iFlytek	**IF**	Commercial	iFlytek	9.44

†: We test the EERs of the five ASVs on *test-clean* of LibriSpeech (B0).

ASV. As shown in Table 3, we use five widely-used ASVs to test the effectiveness and transferability of V-Cloak, i.e., ECAPA-TDNN, X-vector [45], ivector-PLDA [8], GMM-UBM [42] and iFlytek ASV [2]. ECAPA-TDNN and X-vector are DNN-based ASVs implemented using SpeechBrain [41] and trained on VoxCeleb1&2. GMM-UBM and ivector-PLDA are traditional statistical model-based ASVs implemented using Kaldi toolkit [37] and trained on VoxCeleb1&2. iFlytek ASV is a commercial ASV API provided by iFlytek Open Platform [2], with private training data that is unknown. We download pretrained models of the five ASVs. The baseline performance of the five ASVs presented in Table 3 is tested on the *test-clean* set of LibriSpeech.

ASR. We evaluate the decoding error of anonymized audios on eleven ASRs, including five English ones, two Chinese ones, two French ones and two Italian ones. These ASRs are trained on different datasets or have different architectures. More details about the ASRs are listed in Table 6 in the extended version [9]. The baseline performance of the ASRs presented in Table 6 is tested on the corresponding datasets in Table 2.

Evaluation metrics. Six metrics are used to evaluate the performance of voice anonymization.

- *Miss-Match Rate (MMR)*, the probability that the voiceprint of the anonymized audio cannot be matched with that of the original speaker by the ASV, which is like the False Rejection Rate (FRR) of the ASV.

- *Wrong-Match Rate (WMR)*, the probability that the voiceprint of the anonymized audio is matched with that of a wrong speaker, which is like the False Acceptance Rate (FAR) of the ASV.

- *Equal Error Rate (EER)*, the rate at which MMR equals WMR (FRR equals FAR), which measures the overall

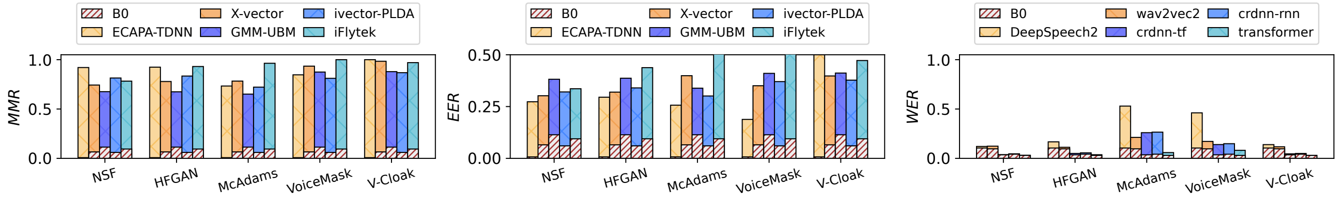

Figure 6: Comparison with existing works. (a) MMR. (b) EER. (c) WER. V-Cloak yields the highest average MMR of 94.02% and the highest average EER of 46.10%. V-Cloak obtains a low average WER of 7.65% second only to the NSF (7.19%).

anonymization power.

- *Word Error Rate (WER)/Character Error Rate (CER)*, metrics that measure the differences between the transcription given by the ASR and the ground-truth. WER (resp. CER) is calculated as,

$$\text{WER (resp. CER)} = \frac{N_{sub} + N_{del} + N_{ins}}{N_{ref}},$$

where N_{sub}, N_{del} and N_{ins} are the numbers of substitution, deletion, and insertion errors of words (resp. characters), respectively. N_{ref} is the ground-truth number of words (resp. characters).

- *Signal-to-noise Ratio (SNR)*, computed as $\text{SNR(dB)} = 10\log_{10}(P_x/P_\delta)$, where $\delta = \tilde{x} - x$, P_x and P_δ are the average power of the original audio and the anonymization perturbation, respectively. SNR is used to evaluate the objective naturalness of the anonymized audios. We also evaluate the subjective naturalness and timbre of the anonymized audios with a user study.

- *Real-time coefficient (RTC)*, computed as $\text{RTC} = T_{cvt}/T_{audio}$, where T_{audio} is the duration of the original audio, and T_{cvt} is the time to anonymize the audio. RTC is used to evaluate the efficiency of the voice anonymization system.

MMR (i.e., false negative/rejection rate) and WMR (i.e., false positive/acceptance rate) are commonly-used metrics for speaker verification systems. EER and WER (CER) are widely used by existing works, e.g., NSF, HFGAN, and McAdams. SNR is a commonly-used metric to measure the imperceptibility of adversarial perturbations. RTC is often used to gauge the efficiency of voice processing methods [38]. To achieve desirable anonymization performance, MMR, WMR, and EER are better to be higher. To maintain intelligibility, WER is better to be lower, and the SNR is better to be higher. To realize real-time voice anonymization, RTC is better to be lower.

Baselines. We compare V-Cloak with the baselines. The performance of ASVs and ASRs with clean/unprocessed audios is referred to as B0. Moreover, we reproduce four state-of-the-art voice anonymization methods, i.e., NSF [14], HFGAN [29], McAdams [34], and VoiceMask [38, 39]. Details of the baselines are in the Appendix B (extended version [9]).

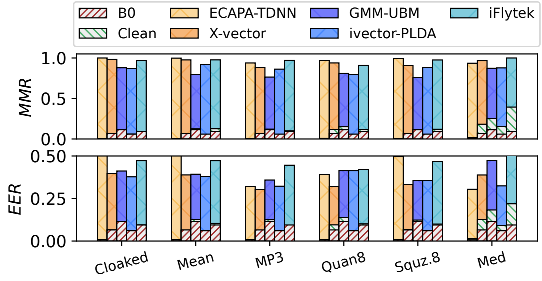

Figure 7: Unidentifiability under adversary A2. The most effective de-noising method, MP3 compression, only causes a decrease in the MMR of 3.27% and the EER of 9.26%.

Hyperparameter tuning. Hyperparameters affect the convergence, performance and generalization of the anonymizer. When balancing different optimization goals of anonymity, intelligibility and naturalness, we need to tune the hyperparameters α, β and γ. We adopt a stepwise hyperparameter tuning, which increases the weights of other loss terms as the anonymity loss decreases. When $L_{\text{ASV}} \geq 0.3$, $\alpha = 0.5, \beta =$1e-6, $\gamma = 0.1$. When $0.15 \leq L_{\text{ASV}} < 0.3$, $\alpha = 0.8, \beta =$2e-6, $\gamma = 0.1$. When $L_{\text{ASV}} < 0.15$, $\alpha = 1, \beta =$3e-6, $\gamma = 0.1$.

5.2 Comparison With Existing Works

As shown in Figure 6 (Table 7 in the extended version [9]), we compare V-Cloak with four existing anonymization methods on the *test-clean* set of LibriSpeech against adversary A1. V-Cloak ($\varepsilon = 0.1$) achieves the highest average MMR of 94.02% and the highest average EER of 46.10%. In the worst case where the attacker adopts the best ASV, the lowest EER of V-Cloak is still 8.23% higher than the lowest EER of all other baselines. V-Cloak gives a low average WER of 7.65% second only to NSF (7.19%). Although NSF achieves the lowest WER, it only gains an MMR of 78.74% and an EER of 32.27%. Although VoiceMask has high unidentifiable performance, it causes severe linguistic distortions, with an average WER of 20.05%. Overall, V-Cloak provides the best anonymity while maintaining a low ASR decoding error (high intelligibility).

Discussion. As far as we are concerned, existing works on voice anonymization have not standardized the threshold of EER for *enough* anonymization. We expect that in the future, standards on data anonymization, especially on biometric data, will be implemented, e.g., by NIST.

5.3 Cross-Language Performance

Apart from English, we also test V-Cloak on Chinese, French and Italian datasets. Note that V-Cloak is trained only on the English ASV and ASR.

As shown in Figure 10 and Table 8 in the extended version [9], V-Cloak can effectively anonymize Chinese audios, with an average MMR of 98.19% and an EER of 51.44%. The cross-language transferability of V-Cloak may be attributed to the language-agnostic ECAPA-TDNN used in the training process. Moreover, we test V-Cloak with two Chinese ASRs, and the CER results are presented in Figure 10. V-Cloak induces a CER increase of only 2.68% with $\varepsilon = 0.1$. Due to a lack of identity information provided by CommonVoice datasets, we only use the French and the Italian datasets for ASR intelligibility test. As shown in Figure 12, Table 9 & 10 in the extended version [9], V-Cloak leads to a WER increase of 6.59% with French audios (from 16.33% to 22.91%) and 6.55% with Italian audios (from 15.11% to 21.66%).

It demonstrates that even if only an English dataset is used for training, the intelligibility-preserving property of V-Cloak can generalize to significantly different languages, i.e., Chinese, French, and Italian in our case.

5.4 Unidentifiability Under Adaptive Attacker A2

We compare the unidentifiability of V-Cloak and baselines against adversary A2, who applies de-noising techniques to try to remove the anonymization perturbation. We consider six methods that are commonly used to remove adversarial perturbations, i.e., smoothing (mean filter and median filter with a kernel of 3, band-pass filter with a passband of 50∼7,500Hz), quantization (from 32-bits to 8-bits), audio squeezing (0.8× the sampling rate), and MP3 compression. The results are shown in Figure 7 (Table 11-16 in the extended version [9]). Most of the de-noising techniques have little influence on the effectiveness of V-Cloak. Among them, MP3 compression has the most obvious influence, i.e., an average MMR decrease of 3.27% (to 90.75%) and an EER decrease of 9.26% (to 36.84%). Compared with the four baselines, V-Cloak achieves the highest EERs in the worst-case scenario against all six de-noising methods. It indicates that the anonymized audios of V-Cloak are robust in terms of unidentifiability under de-noising techniques.

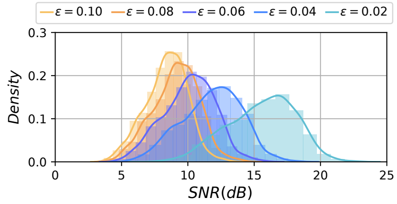

Figure 8: The SNRs at different anonymization levels.

5.5 Unlinkability Under Adaptive Attacker A3

We compare the performance of V-Cloak and baselines against adaptive attacker A3, who has the anonymizer $\mathcal{G}$ and attempts to link the voiceprint of an anonymized audio to that of an anonymizer-processed audio. As shown in Table 4, with the untargeted anonymization, V-Cloak achieves the second highest anonymization performance (only slighly worse than HFGAN). With the targeted anonymization, V-Cloak achieves the highest EER among all methods. For adversary A3, without the voiceprint key, the anonymizer-processed audio (enrollment audio) still cannot be matched with the anonymized audio (test trial), achieving a highest EER of 42.57% in the average case and 37.56% in the worst case. In the most challenging scenario where the adversary has access to the anonymizer and the voiceprint key v, audio samples anonymized with v from any speakers (not only the original speaker of the anonymized audio, but also other speakers) will be matched to the same speaker v, thus producing a highest EER of 36.47%. Overall, our experiment results show that unlinkability is achieved in the face of adversary A3. To verify the targeted anonymization effectiveness of V-Cloak, we convert audios of *test-clean* to 10 speakers in *dev-clean*, and test the success rate. As shown in Table 17 in the extended version [9], 99.99% of the converted audios are successfully matched with the target speakers.

5.6 Anonymization Levels

As shown in Figure 11 and Table 19 in the extended version [9], we test the performance of V-Cloak at different anonymization levels, i.e., vary the ε from 0.02 ∼ 0.10, on the *test-clean* set of LibriSpeech. We can observe that the MMRs and EERs of V-Cloak are much higher than those of clean audios (B0). As ε increases, i.e., larger anonymization perturbations, both MMRs and EERs increase. Note that although the anonymizer of V-Cloak is trained using the ECAPA-TDNN ASV, the anonymization is also effective for the other four ASVs including a commercial ASV, iFlytek, whose architecture and training set are unknown. The transferability is gained probably because we optimize at the intermediate voiceprint layer rather than the last classification layer.

As for intelligibility, we can see that only a slight increase of WER is induced by the anonymization perturbation, i.e., increased by 1.46% from 6.19% to 7.65% ($\varepsilon = 0.1$). Although

Table 4: The performance under adaptive attacker A3.

Model		B0 (%)	NSF (%)	HFGAN (%)	McAdams (%)	VoiceMask (%)	V-CLOAK (%)		
							Untargeted	**Targeted w/o key**†	**Targeted w key**†
ASV	**EP**	0.70	24.50	24.79	9.01	8.80	22.03	46.93	32.98
	XV	6.53	27.56	27.11	9.13	14.75	18.20	44.50	29.67
	GMM	11.39	30.36	31.40	23.33	32.58	39.50	43.44	51.15
	IV	6.03	11.17	26.51	8.76	18.25	32.90	40.42	37.22
	IF	9.44	28.24	27.85	13.31	18.95	19.93	37.56	31.34
	AVG	6.82	24.37	27.53	12.71	18.67	26.51	42.57	36.47
	WCS	-	11.17	24.79	8.76	8.80	18.20	37.56	29.67

(i) †: w/ or w/o key means that the voiceprint of the target speaker is known or unknown to the adversary. (ii) **AVG**: the average-case scenario, **WCS**: the worst-case scenario. **EP**: ECAPA-TDNN, **XV**: X-vector, **GMM**: GMM-UBM, **IV**: ivector-PLDA, **IF**: iFlytek.

only the DeepSpeech2 ASR is used in the training process, the intelligibility of V-CLOAK generalizes to other ASRs of different architectures and training sets.

As shown in Figure 8, the average SNRs of anonymized audios range from 8.4dB ($\varepsilon = 0.1$) to 15.5dB ($\varepsilon = 0.02$). The lower the ε, the larger the average SNR and the variance, which means that for those *hard-to-anonymize* audios, V-CLOAK adaptively generates larger anonymization perturbations, thus V-CLOAK can maintain high MMR (87.86%) even when $\varepsilon = 0.02$.

5.7 Ablation Study

Intelligibility loss. As shown in Table 20 in the extended version [9], without the intelligibility loss, the anonymizer can achieve a higher MMR of 97.16% and a higher EER of 53.53%, since the constraint on the anonymization process is looser. However, the intelligibility is greatly reduced (an average WER of 107.66%) without the intelligibility loss term. In addition, we evaluate the effectiveness of the CTC loss and the GPG loss. We find that GPG is more effective in decreasing the ASR decoding error and stabilizing the convergence of ASV loss.

Throttle. We remove *Throttle* from the anonymizer and present the results in Figure 13 and Table 18 in the extended version [9]. We can see that removing *Throttle* induces an obvious performance degradation on GMM-UBM, ivector-PLDA, and iFlytek, which demonstrates the necessity and effectiveness of *Throttle*.

5.8 Data Separation

In the default experiment setup, the training datasets of V-CLOAK and baselines are partially overlapped with the training datasets of ASVs (note that the test datasets are not overlapped with any training datasets). In this section, we examine the performance of V-CLOAK and baselines when their training datasets do not overlap with those of the ASVs. More specifically, we train V-CLOAK on VoxCeleb1&2 and LibriSpeech (train-clean-100, train-other-500). NSF and HFGAN are trained on VoxCeleb1&2, LibriSpeech (train-clean-100, train-other-500), and LibriTTS (train-clean-100). McAdams and VoiceMask are model-free methods and do not need to be trained. For ASVs, we train ECAPA-TDNN, X-vector, and DeepSpeaker [24] all on LibriSpeech (train-clean-360). The test set is LibriSpeech (test-clean). As shown in Table 5, V-CLOAK achieves the highest average and worst-case MMR and EER compared with four baselines. This verifies the transferability of V-CLOAK, i.e., V-CLOAK is effective in fooling ASVs trained on totally different datasets.

5.9 User Study

To demonstrate the intelligibility-, naturalness-, and timbre-preserving properties of V-CLOAK, we conduct a user study, which is approved by the Institutional Review Board (IRB) of our institutes.

Setup. We have recruited 102 participants to answer two sets of questions. The participants are aged 18∼28 with 72 males and 30 females. Before answering each question, users are asked to listen to one or two audios and give a score according to the quality of the audio(s). Users can listen to the audios for multiple times if they are unsure of the answer. The audios are played via a JBL GO2 loudspeaker in a quiet environment. Each audio sample is clipped to 10s. Note that we anonymize an audio piece by piece so that the anonymization will not degrade over long pieces of audio. The user study is carried out in a controlled laboratory environment, where we guarantee that no bots, scripts or automated answering tools are used in the process. In addition, we randomly insert attention-check questions to ensure that the participants pay attention to the questions.

Timbre. The first set of questions evaluates the timbre of the anonymized audios. We select 5 audios and anonymize each audio with V-CLOAK and four baselines, resulting in a total of 25 pairs of audios. Each participant is asked to listen to the original audio and the anonymized audio from the same speaker and to rate the similarity between the two with a score from 1∼10 points (1 for completely different and 10 for completely the same). The user is also asked to assume that the speaker of the original audio is a celebrity or their acquaintances and rate whether they accept the anonymized audio as from the same speaker. There is a total of 25 similarity ratings and 25 acceptability ratings given by each participant.

Intelligibility & Naturalness. The second set of questions evaluates the intelligibility and the naturalness of the

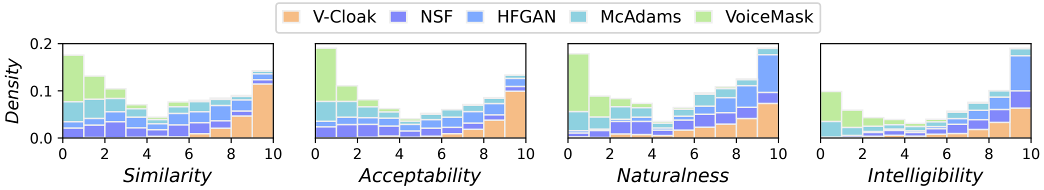

Figure 9: The results of user study.

Table 5: Comparison with existing works under data separation.

Model		B0 (%)	NSF (%)			HFGAN (%)			McAdams (%)			VoiceMask (%)			V-Cloak (%)		
		EER	MMR	WMR	EER	MMR	WMR	EER	MMR	WMR	EER	MMR	WMR	EER	MMR	WMR	EER
ASV	EP	3.72	88.89	3.89	38.09	87.33	3.89	42.21	46.53	3.89	20.69	70.15	3.89	23.40	97.90	3.89	42.21
	XV	5.74	87.33	4.73	34.47	88.89	4.73	39.05	84.28	4.73	40.19	95.73	4.73	37.79	100.0	4.73	44.73
	DP	3.72	93.97	4.05	39.70	89.39	4.05	33.13	80.15	4.05	35.00	99.24	4.05	41.37	99.77	4.05	49.47
	AVG	4.39	90.06	4.22	37.42	88.54	4.22	38.13	70.32	4.22	31.96	88.37	4.22	34.19	99.22	4.22	45.47
	WCS	-	87.33	3.89	34.47	87.33	3.89	33.13	46.53	3.89	20.69	70.15	3.89	23.40	97.90	3.89	42.21

AVG: average, **WCS**: worst-case scenario. **EP**: ECAPA-TDNN, **XV**: X-vector, **DP**: DeepSpeaker.

anonymized audios. We select 5 audios and anonymize each audio with V-Cloak and the four baselines, resulting in a total of 25 audios. Each participant is asked to listen to one audio sample and to rate the intelligibility of the audio with a score from 1∼10 points (1 for completely unintelligible and 10 for completely intelligible). The user is also asked to rate the naturalness of the audio with a score from 1∼10 points (1 for completely unnatural and 10 for completely natural). There is a total of 25 naturalness ratings and 25 intelligibility ratings given by each participant.

Results. As shown in Figure 9, V-Cloak gains the highest scores in similarity and acceptability, with most scores distributed between 7∼10, which means that V-Cloak preserves the timbre of audios for users to trust that the audio is from the genuine speaker. For naturalness and intelligibility, V-Cloak obtains scores as high as HFGAN. User studies verify that V-Cloak preserves intelligibility, naturalness and timbre of the audios.

5.10 Efficiency

We test V-Cloak and the baselines [14, 29, 34, 38, 39] under the same computing resources on the *test-clean* set of LibriSpeech. The results are shown in Table 1, which indicates that V-Cloak has the highest efficiency.

6 Related Work

6.1 Voice Signal Processing

Conventional signal processing techniques are used to alter the voice signals. Patino et al. [34] proposed to utilize the McAdams coefficient to shift the formant positions in an utterance for speaker anonymization. Vaidya et al. [53] modified the pitch, tempo, pause, and MFCCs of an audio to alter the voice characteristics. Voice signal processing methods do not consider preserving naturalness of audios, thus they induce large distortions. In comparison, V-Cloak limits the distortion with a psychoacoustics-based loss in the training phase to mask the introduced anonymization perturbations.

6.2 Voice Conversion & Synthesis

Voice Conversion (VC)/ Synthesis (VS) methods aim to convert the speaker features of an original audio into those of the target speaker while preserving naturalness.

VTLN-based VC. VTLN is a traditional voice conversion method that utilizes a frequency warping function to rescale the frequency axis of the voice spectrogram [13]. Qian et al. [38, 39] utilized a bilinear warping function with randomly-chosen parameters to conceal the original voiceprint. However, they do not consider intelligibility and naturalness of the sanitized audios. According to our user study, the method has low scores of intelligibility and naturalness. Srivastava et al. [48] investigated the performance of different target speaker selection strategies in two VTLN-based VC and one DNN-based VC. Their results show that VTLN-based VC methods suffer from a 6.5%∼10.4% WER increase on ASR. In contrast, we consider intelligibility and naturalness in training the anonymizer to achieve lower WER/CER and high subjective scores in user studies.

DNN-based VC/VS. Yoo et al. [56] proposed to anonymize audios by VC technique based on CycleVAE-GAN, which modifies the speaker identity vectors of the VAE input. Fang et al. [14] proposed to disentangle linguistic and speaker identity features from an utterance, replace the latter with a pseudo identity, and re-synthesize an anonymized audio. Han et al. [16] designed a voiceprint privacy metric according to differential privacy [12] and adapted Fang [14] to a voice data release mechanism that satisfies the privacy metric. Miao et al. [29] followed the basic framework of Fang [14] but proposed to use a HuBERT-based content encoder, an ECAPA-TDNN speaker encoder, and a HiFi GAN to re-synthesize the speech. Justin et al. [21] proposed to transform only the linguistic content of an audio into an audio of another speaker with a speech synthesis system. However, the features of pitch, rhythm, tempo, and pause in the audio are all lost. DNN-based VC/VS methods convert the original speaker features into those of another speaker, which may not be suitable for instant messaging and social media applications. In comparison, we preserve the timbre of the original speaker while hiding the voiceprint of the speaker from the ASV.

6.3 Voice Adversarial Examples

As far as we know, there is only one work on generative model-based audio adversarial attacks, called FAPG. FAPG [55] trains an audio adversarial example generator and a series of feature maps. Each feature map is trained for each target speaker. A feature map can be concatenated with the anonymizer to produce adversarial examples of a specific target speaker. V-CLOAK differs from FAPG in four aspects. Firstly, FAPG attacks a speaker classification model with fixed and known classes and requires a re-training of feature maps for unseen classes/speakers. Secondly, FAPG utilizes the last softmax layer of the classification model, which is training-set-specific, resulting in low transferability [18]. In comparison, V-CLOAK can realize targeted anonymization with the input of any target speaker without the need to retrain the anonymizer. Thirdly, the size of the FAPG model increases with the number of feature maps, while the V-CLOAK model has a constant size. Finally, the design of FAPG only considers one fixed constraint on the added noise. In contrast, V-CLOAK introduces a learnable structure to allow diversified constraints on the anonymization perturbation.

7 Ethics Discussion

V-CLOAK is designed to protect voiceprint, a sensitive biometric, as the General Data Protection Regulation (GDPR) [1] enacted by the European Union grants natural persons “the right to the personal data protection". The processing of biometric data for the purpose of uniquely identifying a natural person is prohibited except for certain cases [1], the prominent ones of which include

- Criminal convictions and offences.
- Social security.
- Scientific or historical research.
- Public health.
- Consent by the data owner.

We shall take proper measures to prevent the abuse of V-CLOAK in these legitimate cases, including but not limited to 1) providing necessary details of V-CLOAK to bodies that can legally conduct voice analysis, 2) withholding the release of the code of V-CLOAK for 90 days after notification.

8 Conclusion & Future Work

In this work, we present the design, implementation and evaluation of V-CLOAK, a real-time voice anonymization system, which preserves intelligibility, naturalness and timbre of the audios. Extensive experiments on four language datasets with various ASVs and ASRs confirm the effectiveness and transferability of V-CLOAK. The user study demonstrates the high perceptual quality of the anonymized audios generated by V-CLOAK. In the future, V-CLOAK can be further improved in several aspects.

Psychoacoustics. Apart from spectral masking, there are other psychoacoustic effects that may be leveraged to further improve the performance of V-CLOAK. For instance, we may find out the non-silent segments of audios utilizing Voice Activity Detection (VAD) [46], and only anonymize the non-silent parts, which may further improve naturalness.

Analog voice data. We mainly consider anonymization for digital voice data in applications such as voice messaging and social media. Anonymization for analog voice data or over-the-air digital voice data may be necessary in the case where the adversary physically records public speeches or private conversations. To realize over-the-air or analog voice data anonymization, a possible way is to incorporate room impulse responses (RIRs) [20] in the optimization problem of voice anonymization. This is our future direction.

Other attacks. We assume that the adversary knows that voice anonymization is performed and only focuses on de-anonymization. However, it is possible that the adversary does not know whether voice anonymization is conducted and tries to detect its presence. Anonymization detection can be performed by A1 ∼A3 before de-anonymization. The attacker may also train the ASV with (denoised) anonymized samples and feed denoised samples into ASV during inference. However, a possible loophole is that the performance of ASV on clean samples may degrade. We consider these attacks as our future direction.

Extension to other applications. The approach we design for V-CLOAK may be extended to other tasks. For example, we may train an adversarial model with a similar architecture against an audio Deepfake model, preventing the generation of Deepfake audio, similar to Fawkes for facial images [44]. We consider the extensions of V-CLOAK as our future work.

Acknowledgments

We sincerely thank our Shepherd and all the anonymous reviewers for their valuable comments. This work is supported by China NSFC Grant 61925109. Yanjiao Chen is the corresponding author.

References

[1] General Data Protection Regulation. https://gdpr-info.eu/.

[2] iFLYTEK Open Platform. https://global.xfyun.cn.

[3] WeChat: Free messaging and calling app. https://www.wechat.com.

[4] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *International Conference on Machine Learning*. PMLR, 2016.

[5] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *Language Resources and Evaluation Conference*. European Language Resources Association, 2020.

[6] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *IEEE Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment*, 2017.

[7] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. Who Is Real Bob? Adversarial attacks on speaker recognition systems. In *IEEE Symposium on Security and Privacy*, 2021.

[8] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[9] Jiangyi Deng, Fei Teng, Yanjiao Chen, Xiaofu Chen, Zhaohui Wang, and Wenyuan Xu. V-Cloak: Intelligibility-, naturalness- & timbre-preserving real-time voice anonymization extended version. https://person.zju.edu.cn/person/attachments/2022-10/01-1664784704-858031.pdf.

[10] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Conference of the International Speech Communication Association*. ISCA, 2020.

[11] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. SirenAttack: Generating adversarial audio for end-to-end acoustic systems. In *ACM Asia Conference on Computer and Communications Security*, 2020.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference*. Springer, 2006.

[13] Ellen Eide and Herbert Gish. A parametric approach to vocal tract length normalization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996.

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

[15] Stanley A Gelfand. *Hearing: An introduction to psychological and physiological acoustics*. CRC Press, 2017.

[16] Yaowei Han, Sheng Li, Yang Cao, Qiang Ma, and Masatoshi Yoshikawa. Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release. In *IEEE International Conference on Multimedia and Expo*, 2020.

[17] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[18] Qian Huang, Isay Katsman, Zeqi Gu, Horace He, Serge J. Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[19] Mara Hvistendahl. How a Chinese AI giant made chatting—and surveillance—easy. https://www.wired.com/story/iflytek-china-ai-giant-voice-chatting-surveillance.

[20] Marco Jeub, Magnus Schäfer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *IEEE International Conference on Digital Signal Processing*, 2009.

[21] Tadej Justin, Vitomir Struc, Simon Dobrisek, Bostjan Vesnicer, Ivo Ipsic, and France Mihelic. Speaker de-identification using diphone recognition and speech synthesis. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015.

[22] Kayla Kibbe. Facebook has been collecting audio data from voice messages. https://www.insidehook.com/daily_brief/tech/facebook-has-been-collecting-audio-data-from-voice-messages.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*. OpenReview.net, 2015.

[24] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

[25] Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

[26] Zhuohang Li, Cong Shi, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. Practical adversarial attacks against speaker recognition systems. In *ACM International Workshop on Mobile Computing Systems and Applications*, 2020.

[27] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In *ACM SIGSAC Conference on Computer and Communications Security*, 2020.

[28] Megan McCluskey. TikTok has started collecting your faceprints and voiceprints. Here's what it could do with them. https://time.com/6071773/tiktok-faceprints-voiceprints-privacy.

[29] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia A. Tomashenko. Language-independent speaker anonymization approach using self-supervised pre-trained models. *arXiv preprint, arXiv:2202.13097*, 2022.

[30] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[31] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. In *Conference of the International Speech Communication Association*. ISCA, 2017.

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems*. PMLR, 2019.

[34] Jose Patino, Natalia A. Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans. Speaker anonymisation using the McAdams coefficient. In *Conference of the International Speech Communication Association*. ISCA, 2021.

[35] Huy Phan, Yi Xie, Siyu Liao, Jie Chen, and Bo Yuan. CAG: A real-time low-cost enhanced-robustness high-transferability content-aware adversarial attack generator. In *AAAI Conference on Artificial Intelligence*, 2020.

[36] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge J. Belongie. Generative adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[38] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *ACM Conference on Embedded Networked Sensor Systems*, 2018.

[39] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. Speech Sanitizer: Speech content desensitization and voice anonymization. *IEEE Transactions on Dependable and Secure Computing*, 18(6):2631–2642, 2021.

[40] Yao Qin, Nicholas Carlini, Garrison W. Cottrell, Ian J. Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International Conference on Machine Learning*. PMLR, 2019.

[41] 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. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

[42] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[43] Korin Richmond. *Estimating articulatory parameters from the acoustic speech signal*. PhD thesis, University of Edinburgh, UK, 2002.

[44] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *USENIX Security Symposium*, 2020.

[45] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-Vectors: Robust DNN embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[46] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

[47] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Annual Conference on Neural Information Processing Systems*. PMLR, 2018.

[48] Brij Mohan Lal Srivastava, Nathalie Vauquier, Md. Sahidullah, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Evaluating voice conversion-based privacy protection against informed attackers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[49] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. In *International Society for Music Information Retrieval Conference*, 2018.

[50] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

[51] Norman Poh Hoon Thian, Conrad Sanderson, and Samy Bengio. Spectral subband centroids as complementary features for speaker authentication. In *Biometric Authentication First International Conference*. Springer, 2004.

[52] Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Hubert Nourtel, Pierre Champion, Massimiliano Todisco, Emmanuel Vincent, Nicholas Evans, Junichi Yamagishi, and Jean-François Bonastre. The VoicePrivacy 2022 Challenge evaluation plan. *arXiv preprint arXiv:2203.12468*, 2022.

[53] Tavish Vaidya and Micah Sherr. You Talk Too Much: Limiting privacy exposure via voice input. In *IEEE Security and Privacy Workshops*, 2019.

[54] Dong Wang, Lantian Li, Zhiyuan Tang, and Thomas Fang Zheng. Deep speaker verification: Do we need end to end? In *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017.

[55] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. Enabling fast and universal audio adversarial attack using generative model. In *AAAI Conference on Artificial Intelligence*, 2021.

[56] In-Chul Yoo, Keonnyeong Lee, Seong-Gyun Leem, Hyunwoo Oh, BongGu Ko, and Dongsuk Yook. Speaker anonymization for personal information protection using voice conversion techniques. *IEEE Access*, 8:198637–198645, 2020.

[57] William A Yost. Psychoacoustics: A brief historical overview. *Acoustics Today*, 11(3):46–53, 2015.

[58] Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. Black-box adversarial attacks on commercial speech platforms with minimal information. In *ACM SIGSAC Conference on Computer and Communications Security*, 2021.

A Theoretical Analysis

In this part, we prove the unidentifiability achieved by the untargeted and targeted anonymization, and prove the unlinkability achieved by the targeted anonymization.

We first define two probabilities p_A and p_R related to the distinguishability of ASV,

$$\forall X, \exists t_A,\ \mathbb{P}\Big(\forall x_1, x_2 \in X,\ \mathcal{S}(\mathcal{V}(x_1), \mathcal{V}(x_2)) \geq t_A\Big) = p_A, \tag{4a}$$

$$\forall X, Y, \exists t_R,\ \mathbb{P}\Big(\forall x \in X, y \in Y,\ \mathcal{S}(\mathcal{V}(x), \mathcal{V}(y)) < t_R\Big) = p_R, \tag{4b}$$

where X, Y are the data distributions of two different speakers. $t_A, t_R \in (-1, 1)$ are two thresholds, and we have $t_A \geq t_R$. $\mathcal{S}(\cdot, \cdot)$ is the cosine similarity function. Without loss of generality, we further assume that $\forall x, \|\mathcal{V}(x)\| = 1$. (4a) states that if two utterances are from the same speaker, the ASV outputs the score no less than t_A with a probability of p_A. (4b) states that if two utterances are from two different speakers, the ASV outputs the score less than t_R with a probability of p_R. The ASV guarantees that both p_A and p_R are close to 1.

A.1 Untargeted Anonymization

According to the basic formulation (1), we train an untargeted anonymizer $\mathcal{G}$ that satisfies

$$\mathbb{P}\Big(\forall x,\ \mathcal{S}(\mathcal{V}(\tilde{x}), \mathcal{V}(x)) < k_u\Big) = p_G, \tag{5}$$

where $\tilde{x} = \mathcal{G}(x), k_u \in (-1, 1)$. (5) states that for any audio x, $\mathcal{G}$ outputs $\tilde{x}$ such that the score between x and $\tilde{x}$ is less than k_u with a probability of p_G close to 1.

Theorem 1 *$\forall x_0, x_1 \in X$, $\forall t_R \in (\sqrt{5} - 2, 1)$, $\exists k_u \in (-1, 1)$, such that $\mathcal{S}(\mathcal{V}(\tilde{x}_0), \mathcal{V}(x_1)) < t_R$ with a probability higher than $p_A \cdot p_G$.*

Proof A.1

$$\begin{aligned}
&\mathcal{S}(\mathcal{V}(\tilde{x}_0), \mathcal{V}(x_1)) = \mathcal{V}(\tilde{x}_0) \cdot \mathcal{V}(x_1) \\
&= \mathcal{V}(\tilde{x}_0) \cdot \big[\mathcal{V}(x_1) - \mathcal{V}(x_0)\big] + \mathcal{V}(\tilde{x}_0) \cdot \mathcal{V}(x_0) \\
&\leq \|\mathcal{V}(\tilde{x}_0)\| \cdot \|\mathcal{V}(x_1) - \mathcal{V}(x_0)\| + \mathcal{V}(\tilde{x}_0) \cdot \mathcal{V}(x_0) \\
&= \|\mathcal{V}(\tilde{x}_0)\| \cdot \sqrt{\|\mathcal{V}(x_1)\|^2 + \|\mathcal{V}(x_0)\|^2 - 2\mathcal{V}(x_1) \cdot \mathcal{V}(x_0)} \\
&\qquad + \mathcal{V}(\tilde{x}_0) \cdot \mathcal{V}(x_0) \\
&= \sqrt{2 - 2\mathcal{V}(x_1) \cdot \mathcal{V}(x_0)} + \mathcal{V}(\tilde{x}_0) \cdot \mathcal{V}(x_0)
\end{aligned}$$

From (4a),

$$\mathbb{P}\Big(\sqrt{2 - 2\mathcal{V}(x_1) \cdot \mathcal{V}(x_0)} \leq \sqrt{2 - 2t_A}\Big) = p_A. \tag{6}$$

From (5),

$$\mathbb{P}\Big(\mathcal{V}(\tilde{x}_0) \cdot \mathcal{V}(x_0) < k_u\Big) = p_G. \tag{7}$$

Thus,

$$\begin{aligned}
&\mathbb{P}\Big(\mathcal{S}(\mathcal{V}(\tilde{x}_0), \mathcal{V}(x_1)) \leq \sqrt{2 - 2\mathcal{V}(x_1) \cdot \mathcal{V}(x_0)} \\
&\qquad + \mathcal{V}(\tilde{x}_0) \cdot \mathcal{V}(x_0) < \sqrt{2 - 2t_A} + k_u\Big) \geq p_A \cdot p_G.
\end{aligned}$$

$\forall t_R \in (\sqrt{5} - 2, 1)$, $t_A \geq t_R$,

$$\begin{aligned}
&\mathcal{S}(\mathcal{V}(\tilde{x}_0), \mathcal{V}(x_1)) < \sqrt{2 - 2t_A} + k_u \leq \sqrt{2 - 2t_R} + k_u, \\
&\sqrt{2 - 2t_R} + k_u < t_R \Leftrightarrow k_u < \big(t_R - \sqrt{2 - 2t_R}\big) \in (-1, 1).
\end{aligned}$$

Therefore, $\exists k_u \in (-1, 1)$, such that the probability of $\mathcal{S}(\mathcal{V}(\tilde{x}_0), \mathcal{V}(x_1)) < t_R$ is higher than $p_A \cdot p_G$, which is close to 1.

A.2 Targeted Anonymization

A.2.1 Unidentifiability

According to the basic formulation (1), we train a targeted anonymizer $\mathcal{G}$ that satisfies

$$\mathbb{P}\Big(\forall x,\ \mathcal{S}(\mathcal{V}(\tilde{x}), v) > k_t\Big) = p_G, \tag{8}$$

where $\tilde{x} = \mathcal{G}(x, v), k_t \in (-1, 1)$. (8) states that for any audio x, the $\mathcal{G}$ outputs $\tilde{x}$ such that the similarity score between v and $\tilde{x}$ is larger than k_t with a probability of p_G. We assume that the target speaker has a voiceprint v that is far away from that of the original speaker, i.e.,

$$\mathbb{P}\Big(\forall x \in X,\ \mathcal{S}(\mathcal{V}(x), v) < t_R - \eta\Big) = p_R, 0 \leq \eta < t_R + 1, \tag{9}$$

where $(t - \eta)$ stands for how far the target voiceprint v is from the user's, $\mathcal{V}(x)$, and η is a margin.

Theorem 2 *$\forall x_0, x_1 \in X$, $\forall t_R \in (-\frac{1}{2}, 1)$, $\exists k_t \in (-1, 1), \eta \in (0, t_R + 1)$, such that $\mathcal{S}(\mathcal{V}(\tilde{x}_0), \mathcal{V}(x_1)) < t_R$ with a probability higher than $p_R \cdot p_G$.*

Proof A.2

$$\begin{aligned}
&\mathcal{S}(\mathcal{V}(\tilde{x}_0), \mathcal{V}(x_1)) = \mathcal{V}(\tilde{x}_0) \cdot \mathcal{V}(x_1) \\
&= \mathcal{V}(\tilde{x}_0) \cdot \big[\mathcal{V}(x_1) + v\big] - \mathcal{V}(\tilde{x}_0) \cdot v \\
&\leq \|\mathcal{V}(\tilde{x}_0)\| \cdot \|\mathcal{V}(x_1) + v\| - \mathcal{V}(\tilde{x}_0) \cdot v \\
&= \|\mathcal{V}(\tilde{x}_0)\| \cdot \sqrt{\|\mathcal{V}(x_1)\|^2 + \|v\|^2 + 2\mathcal{V}(x_1) \cdot v} \\
&\qquad - \mathcal{V}(\tilde{x}_0) \cdot v \\
&= \sqrt{2 + 2\mathcal{V}(x_1) \cdot v} - \mathcal{V}(\tilde{x}_0) \cdot v
\end{aligned} \tag{10}$$

From (4b),

$$\mathbb{P}\Big(\sqrt{2 + 2\mathcal{V}(x_1) \cdot v} < \sqrt{2 + 2(t_R - \eta)}\Big) = p_R.$$

From (8),

$$\mathbb{P}\Big(\mathcal{V}(\tilde{x}_0)\cdot v > k_t\Big) = p_G.$$

Thus,

$$\mathbb{P}\Big(\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(x_1)) \le \sqrt{2+2\mathcal{V}(x_1)\cdot v} \\ -\mathcal{V}(\tilde{x}_0)\cdot v < \sqrt{2+2(t_R-\eta)}-k_t\Big) \ge p_R\cdot p_G.$$

Let $\eta = \frac{1}{2}$, $\forall t_R \in (-\frac{1}{2},1)$,

$$\sqrt{2+2(t_R-1/2)}-k_t < t_R \\ \Leftrightarrow k_t > \big(\sqrt{1+2t_R}-t_R\big) \in (\frac{1}{2},1).$$

Therefore, $\exists k_t \in (-1,1)$, *such that the probability of* $\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(x_1)) < t_R$ *is higher than* $p_R\cdot p_G$, *which is close to 1.*

A.2.2 Unlinkability

A3 does not have the key. We first prove that if A3 does not have the voiceprint key v and converts the clean samples of potential speakers with a random key u, then the anonymized audio $\tilde{x}_0$ will not be matched with any enrollment samples with high probability.

Theorem 3 $\forall x_0 \in X$, $\forall y$, $\exists k_t \in (-1,1)$, *such that* $\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(\tilde{y})) < t_R$ *with a high probability, where* $\tilde{x}_0 = \mathcal{G}(x_0,v)$, $\tilde{y} = \mathcal{G}(y,u)$ *and* $u \sim U(\{u \in \mathbb{R}^{1\times N} : \|u\| = 1\})$.

Proof A.3

$$\begin{aligned}&\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(y)) \\ &= \big[\mathcal{V}(\tilde{x}_0)-v\big]\cdot\big[\mathcal{V}(y)-u\big]+u\cdot\mathcal{V}(\tilde{x}_0)+v\cdot\mathcal{V}(y)-u\cdot v \\ &= \big[\mathcal{V}(\tilde{x}_0)-v\big]\cdot\big[\mathcal{V}(y)-u\big]+u\cdot\big[\mathcal{V}(\tilde{x}_0)-v\big] \\ &\qquad +v\cdot\big[\mathcal{V}(y)-u\big]+u\cdot v \\ &\le \|\mathcal{V}(\tilde{x}_0)-v\|\cdot\|\mathcal{V}(y)-u\|+\|u\|\cdot\|\mathcal{V}(\tilde{x}_0)-v\| \\ &\qquad +\|v\|\cdot\|\mathcal{V}(y)-u\|+u\cdot v,\end{aligned} \tag{11}$$

of which,

$$\mathbb{P}\Big(\|\mathcal{V}(\tilde{x}_0)-v\| = \sqrt{\|\mathcal{V}(\tilde{x}_0)\|^2+\|v\|^2-2\mathcal{V}(\tilde{x}_0)\cdot v} \\ \le \sqrt{2-2\mathcal{V}(\tilde{x}_0)\cdot v} < \sqrt{2-2k_t}\Big) = p_G.$$

Similarly,

$$\mathbb{P}\Big(\|\mathcal{V}(y)-u\| < \sqrt{2-2k_t}\Big) = p_G.$$

Lemma 1 *[25] The surface area of an N-sphere in N-dimensional Euclidean space, of radius R, can be given in the closed form:*

$$A_N(R) = \frac{2\pi^{\frac{N}{2}}}{\Gamma(\frac{N}{2})}R^{N-1}, \tag{12}$$

where Γ *is the gamma function. An N-sphere can be cut into two caps, by a hyperplane. We denote the colatitude angle, i.e., the angle between a vector of the sphere and its positive* N^{th}*-axis, as* φ*. We only consider the smaller cap in the following, i.e.,* $0 \le \varphi \le \frac{\pi}{2}$.

Lemma 2 *[25] The surface area of a hyperspherical cap described above can be given in the closed form:*

$$A_N^{cap}(R,\varphi) = \frac{1}{2}A_N(R)I_{sin^2_\varphi}\Big(\frac{N-1}{2},\frac{1}{2}\Big), \tag{13}$$

where $I_{sin^2_\varphi}(\cdot,\cdot)$ *is the regularized incomplete beta function.*

Thus,

$$\mathbb{P}(u\cdot v \ge \cos\varphi) = \frac{A_N^{cap}(R,\varphi)}{A_N(R)} = \frac{1}{2}I_{sin^2_\varphi}\Big(\frac{N-1}{2},\frac{1}{2}\Big) \tag{14}$$

Let $\varphi = \arccos(\alpha\cdot t_R), 0 < \alpha < 1$,

$$\begin{aligned}&\mathbb{P}(u\cdot v < \alpha t_R) = 1-\frac{1}{2}I_{sin^2_\varphi}\Big(\frac{N-1}{2},\frac{1}{2}\Big) \\ &= \frac{1}{2}+\frac{1}{2}I_{1-sin^2_\varphi}\Big(\frac{1}{2},\frac{N-1}{2}\Big) = \frac{1}{2}+\frac{1}{2}I_{cos^2_\varphi}\Big(\frac{1}{2},\frac{N-1}{2}\Big) \\ &= \frac{1}{2}+\frac{1}{2}I_{\alpha^2t_R^2}\Big(\frac{1}{2},\frac{N-1}{2}\Big)\end{aligned} \tag{15}$$

Thus,

$$\begin{aligned}&\mathbb{P}\Big(\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(y)) < 2-2k_t+2\sqrt{2-2k_t}+\alpha\cdot t_R\Big) \\ &\ge p_G^2\cdot\left(\frac{1}{2}+\frac{1}{2}I_{\alpha^2t_R^2}\Big(\frac{1}{2},\frac{N-1}{2}\Big)\right) \\ &\to 1(p_G \to 1, N \to \infty) \\ &\lim_{k_t\to 1}\big(2-2k_t+2\sqrt{2-2k_t}+\alpha\cdot t_R\big) = \alpha\cdot t_R < t_R\end{aligned} \tag{16}$$

Therefore, $\forall t_R \in (-1,1)$, $\exists k_t \in (-1,1)$ *such that* $\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(y)) < t_R$ *with a probability close to 1.*

A3 has the key. Next, we prove that if A3 has the voiceprint key v, and converts the clean samples of potential speakers with the exact v, then the anonymized audio $\tilde{x}_0$ will be matched with any enrollment samples with high probability.

Theorem 4 $\forall x_0 \in X$, $t_A \in (-1,1)$, $\exists k_t \in (-1,1)$, $\forall y$, $\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(\tilde{y})) > t_A$ *with a high probability.*

Proof A.4 *From (8),*

$$\mathbb{P}\Big(\mathcal{V}(\tilde{x}_0)\cdot v > k_t\Big) = p_G, \\ \mathbb{P}\Big(\mathcal{V}(\tilde{y})\cdot v > k_t\Big) = p_G.$$

Thus, $\forall y$,

$$\mathbb{P}\Big(\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(\tilde{y})) = \big[\mathcal{V}(\tilde{x}_0)\cdot v\big]\big[\mathcal{V}(\tilde{x}_1)\cdot v\big] > k_t^2\Big) \ge p_G^2.$$

Therefore, $\forall t_A \in [-1,1]$, *let* $k_t = \sqrt{t_A+\alpha(1-t_A)}, 0 < \alpha < 1$, *such that,*

$$\mathcal{S}(\mathcal{V}(\tilde{x}_0),\mathcal{V}(\tilde{y})) > k_t^2 = t_A+\alpha(1-t_A) > t_A.$$