



Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition

Xingyu Zhang¹ · Xiongwei Zhang¹ · Meng Sun¹ · Xia Zou¹ · Kejiang Chen² · Nenghai Yu²

Received: 29 November 2021 / Accepted: 1 May 2022 / Published online: 17 June 2022
© The Author(s) 2022

Abstract

Automatic speaker recognition is an important biometric authentication approach with emerging applications. However, recent research has shown its vulnerability on adversarial attacks. In this paper, we propose a new type of adversarial examples by generating *imperceptible* adversarial samples for *targeted* attacks on *black-box* systems of automatic speaker recognition. Waveform samples are created directly by solving an optimization problem with waveform inputs and outputs, which is more realistic in real-life scenario. Inspired by *auditory masking*, a regularization term adapting to the energy of speech waveform is proposed for generating imperceptible adversarial perturbations. The optimization problems are subsequently solved by *differential evolution algorithm* in a black-box manner which does not require any knowledge on the inner configuration of the recognition systems. Experiments conducted on commonly used data sets, LibriSpeech and VoxCeleb, show that the proposed methods have successfully performed targeted attacks on state-of-the-art speaker recognition systems while being imperceptible to human listeners. Given the high SNR and PESQ scores of the yielded adversarial samples, the proposed methods deteriorate less on the quality of the original signals than several recently proposed methods, which justifies the imperceptibility of adversarial samples.

Keywords Automatic speaker recognition · Adversarial examples · Imperceptibility · Black-box attack · Differential evolution · Auditory masking

Introduction

Automatic speaker recognition is a technique to identify a person from the characteristics of his/her voice, which has been applied in voice interaction systems, such as smartphones [1], front-end of voice wake-up devices [2], and on-site access control for secured rooms [3], etc. However, recent works have shown that speaker recognition is vulnerable to malicious attacks, e.g., spoofing and adversarial attacks. In the case of spoofing attacks, audio files sound like the target victim and can get access to speaker recognition systems [4, 5]. The typical methods of spoofing include impersonation, replay, speech synthesis, and voice

conversion [6, 7]. On the contrary, audio files in adversarial attacks do not have to sound like the target victim but can also get access to the speaker recognition system, e.g., for white-box attacks in [8–13] and black-box attacks in [8, 9, 14, 15]. Actually, there are plenty of methods proposed for conducting adversarial attacks on image classification, including Box-Constrained L-BFGS [16], Fast Gradient Sign Method (FGSM) [17], Basic Iterative Method (BIM) [18], Jacobian-based Saliency Map Attack (JSMA) [19], One Pixel Attack [20], Carlini and Wagner Attacks (C&W) [21], DeepFool [22], Universal Adversarial Perturbations [23], etc. Unlike the existing extensive studies on image classification, adversarial attacks on speaker recognition are just emerging recently, as presented in a recent work [24]. It is a straightforward solution to deploy fraud detection systems when protecting real-world voice-based biometrics authentication systems [25]. However, in most previous fraud detection works, small perturbations such as adversarial examples have not been considered [26]. In those works, fraud detection systems may be ineffective on detecting adversarial examples given the fact that adversarial examples have only quite

✉ Meng Sun
sunmeng@aeu.edu.cn

¹ Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing, China

² Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China

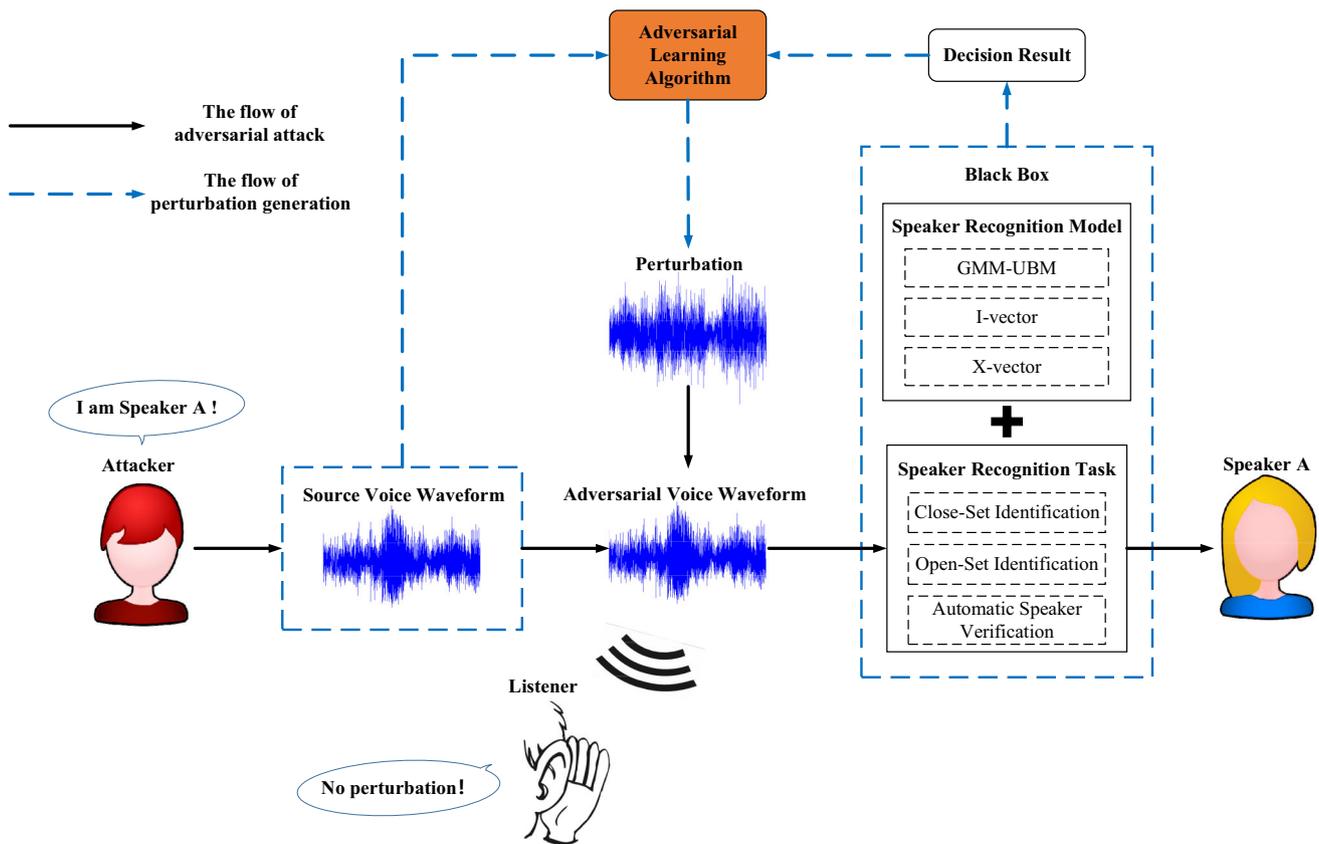


Fig. 1 Attacking scenario of this paper. The blue dotted line represents the generation process of adversarial perturbations, and the black solid line represents the conducting process of an adversarial attack. An attacker first has the source voice. An adversarial learning algorithm subsequently generates a perturbation signal based on the source voice

and adds them together to obtain an adversarial voice. Human listener cannot tell the difference between the source voice and the adversarial one. However, the speaker recognition model will be fooled to recognize the adversarial voice as from speaker A in miscellaneous tasks

small modifications regarding to their original counterparts. In this paper, we consider the attacking scenario depicted in Fig. 1. The adversarial attacking takes the following steps. An attacker first has the source voice. An adversarial learning algorithm subsequently generates a perturbation signal based on the source voice and adds them together to obtain an adversarial voice. Human listeners cannot tell the difference between the source voice and the adversarial one. However, a speaker recognition system can be fooled to recognize the adversarial voice as from speaker A. Given this property, adversarial examples can be utilized to protect the privacy of users of voice interface by preventing the users' voice being identified by some speaker recognition system. Therefore, our work could reduce the chance of malicious usages of one's voice biometrics.

A feasible method for crafting an adversarial example is adding a tiny amount of well-tuned additive perturbation on the source voice to fool speaker recognition. There are generally two kinds of attacks according to how to fool speaker recognition. One is *untargeted* attack, in which speaker

recognition fails to identify the correct identity of the modified voice. The other is *targeted* attack, in which speaker recognition recognizes the identity of an adversarial sample as the specific speaker. Given the experiences on adversarial attacks for image classification [27] and speech recognition [28], targeted attack is much more challenging than the untargeted one. In this paper, we investigate *targeted* attack for automatic speaker recognition.

A key property of a successful adversarial attack is the difference between the adversarial sample and the source one should be *imperceptible* to human perception. Unfortunately, some studies have not paid enough attention on this property, where additive perturbations could be too large to be imperceptible by human listening as illustrated in Fig. 1, e.g., significant background noises were introduced to conduct a successful attack but also deteriorated the quality of the source voice when adding perturbations [14]. A feasible solution would be considering the psychoacoustic property of sounds as studied in [12] and [29]. In this paper, inspired by auditory masking, we will improve the *imperceptibility*

of adversarial samples by constraining both the amounts and the amplitudes of the adversarial perturbations.

The less the prior knowledge required by an attack, the easier the attack conducted in practical usage. Given the assumption that an attacker does not have any knowledge on the inner configuration of the recognition systems, we focus on the *black-box* adversarial attack in this paper, where an attacker can at most access the decision results or the scores of predictions, following the definitions in [30]. However, the difficulty of black-box attack is much greater than that of white-box attack, as reported in [14]. An algorithm assisted by differential evolution was proposed in [20] to perform black-box attack on image classification by only modifying a few of pixels to create an adversarial image. In this paper we make adversarial audio samples by only modifying partial points of an utterance. Given the fact that excessive large amplitudes in audio samples would produce harsh noises, constraints on the amplitudes of adversarial samples will be considered in our methods.

With the vastly usage of time–frequency features as the inputs of speaker recognition, Mel-Spectrum, Mel-frequency cepstral coefficients (MFCCs) and log power magnitude spectrums (LPMSs) were utilized to generate adversarial features in [8] and [9], respectively, where the attacks were performed on feature space rather than on time domain signals. In this paper, we generate adversarial perturbations directly on *waveform-level* (and not on the spectrogram) to yield high-quality samples for attacking, which is more realistic in real-life scenarios as pointed out in [13].

There are three typical tasks for automatic speaker recognition, say open-set identification (OSI) [31], close-set identification (CSI) [32] and automatic speaker verification (ASV) [33]. Existing works have investigated attacks on CSI [10–13], ASV [8, 9] and OSI [14]. In this paper, we comprehensively study targeted adversarial attacks towards all these three tasks within the proposed framework. Comparison and analysis will also be presented with respect to the existing works.

In summary, our method has the following properties, *imperceptibility*, *black-box*, *waveform-level*, and availability to *multiple tasks*. An overview of the working flow of the proposed method is shown in Fig. 1. The rest of the paper is organized as follows. In second section, we first revisit state-of-the-art speaker recognition systems and three typical tasks and then describe the configurations of adversarial attacks. The proposed algorithms are presented in third section. Experimental settings are described in fourth section

and the results and analysis are reported in fifth section. The conclusion is given in sixth section.

Speaker recognition and adversarial attacks

In this section, state-of-the-art speaker recognition systems and three typical tasks involved in speaker recognition are introduced. The configurations of adversarial attacks are also presented.

Speaker recognition: tasks and models

Tasks and their decision functions

We consider three typical tasks in speaker recognition, OSI, CSI and ASV.

An OSI system enrolls multiple speakers in the enrollment stage, say a group of speakers G with IDs $\{1, 2, \dots, n\}$. In the following testing stage, for an arbitrary testing voice \mathbf{x} , the system tries to decide whether \mathbf{x} is from one of the enrolled speakers or none of them, according to the similarity of \mathbf{x} regarding to the enrolled utterances of the speakers in G . A predefined threshold θ is taken to conduct the binary decision, the decision function $D(\mathbf{x})$ in OSI is hereby

$$D(\mathbf{x}) = \begin{cases} \arg \max_{i \in G} f_i(\mathbf{x}) & \text{if } \max_{i \in G} f_i(\mathbf{x}) \geq \theta, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $f_i(\mathbf{x})$ denotes the similarity score of \mathbf{x} on the utterances of the enrolled speaker i . Intuitively, the system classifies the input utterance \mathbf{x} as from speaker i if and only if the score $f_i(\mathbf{x})$ is the highest one and not lower than the threshold θ . Therefore, a successful targeted attack on OSI should satisfy the following two conditions: (1) the score on the target speaker is the highest one among G and (2) the score is larger than θ .

CSI is a task to identify the identity of a speaker in a closed set, i.e., an input utterance will always be classified as from one of the enrolled speakers. The decision function is thus

$$D(\mathbf{x}) = \arg \max_{i \in G} f_i(\mathbf{x}). \quad (2)$$

A successful targeted attack on CSI only needs to make the score of the target speaker the highest one among G .

ASV is a task to verify if an utterance spoken by a *claimed* speaker. ASV has exactly one enrolled speaker and its decision function is thus

$$D(\mathbf{x}) = \begin{cases} 1, & \text{if } f(\mathbf{x}) \geq \theta, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where $f(\mathbf{x})$ is the score of the testing utterance \mathbf{x} on the enrolled speaker. A successful attack on ASV is hence to make the score as large as possible to be higher than the threshold of the system. As seen above, ASV is a special case of OSI, but it will be studied separately in this paper given its importance and vastly usage in biometric authentication.

State-of-the-art models

In this paper, state-of-the-art speaker recognition systems, Gaussian Mixture Models—Universal Background Models (GMM–UBM), *i-vector* and *x-vector*, are considered in our experiments.

GMM–UBM is a traditional statistics-based method in speaker recognition, which has been utilized in both academia and industries [33]. Our study on the vulnerability of GMM–UBM is helpful to understand the weaknesses of speaker recognition *based on statistics*.

An *i-vector* system consists of two basic components: GMM–UBM and a total variability matrix [34]. *i-vector* had been a benchmark method for implementing speaker recognition systems before the emergence of deep learning-based methods. The extraction of *i-vector* is essentially a factor analysis process, which tends to yield high accuracy with inputs of long utterances. Our study on the vulnerability of *i-vector* is helpful to understand the vulnerability of models using *factor analysis*.

In an *x-vector* system, Time Delay Neural Networks (TDNN) is introduced to produce speaker-discriminative embeddings [35]. Network architectures have been extensively studied recently towards better representation of speaker characteristics, among which *x-vector* is an open-source benchmark one. Our study on the adversarial examples of *x-vector* system is helpful to understand the weakness of speaker recognition systems based on *neural networks*.

General formulation of adversarial attacks

Given an audio sample \mathbf{x} , the key step to craft an adversarial sample is to obtain the perturbed signal $e(\mathbf{x})$ by solving the following optimization problem:

$$\begin{aligned} \min_{e(\mathbf{x})} \quad & Q(D(\mathbf{x} + e(\mathbf{x})), y_{\text{tar}}) \\ \text{s.t.} \quad & \|e(\mathbf{x})\|_p < \varepsilon, \end{aligned} \quad (4)$$

whose the goal is to fool the classifier to produce erroneous output for $\hat{\mathbf{x}} = \mathbf{x} + e(\mathbf{x})$ and where *s.t.* means “subject to”, $\|\cdot\|_p$ means p -norm. In other words, if the label of the target

victim speaker is y_{tar} , a successful attack fools the classifier (e.g., GMM–UBM, *i-vector* or *x-vector*) to produce y_{tar} for the perturbed sample $\hat{\mathbf{x}}$. In this paper, two norms l_∞ and l_0 are taken as $\|\cdot\|_p$ to constrain the maximum amplitude and length of perturbation, respectively.

The proposed methods

In this section, the procedures of generating imperceptible, black-box, waveform-level adversarial examples are formulated into an optimization problem. Specifically, objective functions are configured for the tasks of OSI, CSI, and ASV, respectively.

Given our motivation on performing waveform-level adversarial attacks, an input utterance is represented by a vector where each entry is a sample point. Let D be the target speaker recognition model, and let $\mathbf{x} = (x_1, \dots, x_n)$ be the source utterance with x_i denoting the i -th sample point. The adversarial attack is thus to deceive D in Eqs. (1), (2) and (3) by modifying \mathbf{x} with an additive adversarial perturbation vector $e(\mathbf{x}) = (e_1, \dots, e_n)$, i.e., $\hat{\mathbf{x}} = \mathbf{x} + e(\mathbf{x})$. As explained in the Introduction, the difference between $\hat{\mathbf{x}}$ and \mathbf{x} should be imperceptible, and D should classify $\hat{\mathbf{x}}$ as the targeted speaker predefined by the attacker. Therefore, l_0 is introduced to measure the number of non-zero sample points in $e(\mathbf{x})$ with d as its upper bound; l_∞ is introduced to constrain the amplitude of $e(\mathbf{x})$ to be less than a given small tolerance value ζ . Thus, solving the perturbation $e(\mathbf{x})$ boils down to the following constrained optimization problem,

$$\begin{aligned} \min_{e(\mathbf{x})} \quad & Q_0(D(\mathbf{x} + e(\mathbf{x})), y_{\text{tar}}) + \lambda Q_1(\mathbf{x}, e(\mathbf{x})) \\ \text{s.t.} \quad & \|e(\mathbf{x})\|_0 \leq d \text{ and } \|e(\mathbf{x})\|_\infty \leq \zeta, \end{aligned} \quad (5)$$

where Q_0 is a loss function to evaluate if the modification $\mathbf{x} + e(\mathbf{x})$ has successfully been classified as the target speaker y_{tar} . The configuration of Q_0 for OSI, CSI and ASV will be presented in “Attacks on OSI”, “Attacks on CSI” and “Attacks on ASV” sections. Q_1 is a regularization term proposed in this paper to further improve imperceptibility by considering the energy distribution over \mathbf{x} , which is inspired by auditory masking and will be presented in “Regularization to promote imperceptibility” section.

Attacks on OSI

The goal of the targeted attack on OSI is to find an optimized perturbation $e^*(\mathbf{x})$ such that $\hat{\mathbf{x}} = \mathbf{x} + e^*(\mathbf{x})$ can be classified as some target speaker with $\text{tar} \in G$. To conduct a successful attack on OSI, the following two conditions should be satisfied simultaneously: the score $f_{\text{tar}}(\hat{\mathbf{x}})$ of the target speaker

tar should be the highest one among all the enrolled speakers in G , and not lower than the preset threshold θ . Given how OSI makes decisions in (1), the loss function Q_0 to be minimized is hereby

$$Q_0(\mathbf{x}) = \max(\theta, \max_{j \in G, j \neq tar} f(\hat{\mathbf{x}})) - f_{tar}(\hat{\mathbf{x}}), \tag{6}$$

where $f_{tar}(\hat{\mathbf{x}})$ and $\max_{j \in G, j \neq tar} f(\hat{\mathbf{x}})$ refer to the score on target speaker *tar* and the highest score on the remaining speakers, respectively. The objective function tries to increase the gap between $f_{tar}(\hat{\mathbf{x}})$ and $\max(\theta, \max_{j \in G, j \neq tar} f(\hat{\mathbf{x}}))$. In detailed optimization processes, if $\max(\theta, \max_{j \in G, j \neq tar} f(\hat{\mathbf{x}})) = \theta$, it means that all the enrolled speakers are identified outside G , except the target one. Therefore, one only needs to ensure $f_{tar}(\hat{\mathbf{x}})$ to be larger than the threshold θ . If $\max(\theta, \max_{j \in G, j \neq tar} f(\hat{\mathbf{x}})) = \max_{j \in G, j \neq tar} f(\hat{\mathbf{x}})$, one has to increase the gap between the target speaker and the maximum one among the remaining speakers, i.e., to increase the gap between $f_{tar}(\hat{\mathbf{x}})$ and $\max_{j \in G, j \neq tar} f(\hat{\mathbf{x}})$.

Attacks on CSI

Given the closed set involved in CSI in (2), it is expected to find some small perturbation $e(\mathbf{x})$ such that the score on the target speaker is as large as possible, while the score on the second largest one is as small as possible. The loss function to obtain $\hat{\mathbf{x}}$ towards a successful targeted attack on CSI would be

$$Q_0(\mathbf{x}) = \max_{j \in G, j \neq tar} f(\hat{\mathbf{x}}) - f_{tar}(\hat{\mathbf{x}}). \tag{7}$$

Attacks on ASV

As an important special case of OSI, ASV has exactly one enrolled speaker and verifies if the input voice is from the enrolled speaker or not. Given how ASV makes decisions in (3), the loss function for making adversarial samples in ASV turns out to be

$$Q_0(\mathbf{x}) = \theta - f_{tar}(\hat{\mathbf{x}}). \tag{8}$$

The objective is to make the score on the target speaker $f_{tar}(\hat{\mathbf{x}})$ as large as possible to be higher than the threshold θ .

Regularization to promote imperceptibility

Beyond the constraints on $e(\mathbf{x})$ in (5), the imperceptibility of adversarial perturbation can further be enhanced in an adaptive way by considering the energy distribution of \mathbf{x} . This is actually an auditory masking effect. In psychoacoustics,

auditory masking occurs when the auditory perception of a sound (named by reference sound) is affected by the presence of another sound (named by masking sound) [36]. Inspired by auditory masking, a regularization term is proposed to promote the design of $e(\mathbf{x})$, that is

$$Q_1(\mathbf{x}, e(\mathbf{x})) = \sum_{i=1}^N \frac{e_i(\mathbf{x})}{|x_i| + \varepsilon}, \tag{9}$$

where ε is a very small constant to avoid numerical errors. The regularization term reflects the amplitude ratio of the perturbations to the original waveform at sample-point level. As a part of the loss function (5), when one tries to minimize (5), the regularization term would also be minimized. That is, one allows a large $e_i(\mathbf{x})$ given a large $|x_i|$, but penalizes the large $e_i(\mathbf{x})$ when $|x_i|$ is small. This process tends to produce perturbations with a similar shape of the input signal, as denoted in (6) of Fig. 2. The intuitive interpretation of (9) is putting more perturbations on the segments of \mathbf{x} with high energy, which would reduce the perception on the reference sound $e(\mathbf{x})$ when playing together with the masking sound \mathbf{x} .

Differential evolution for black-box optimization

Differential evolution (DE) is a kind of evolutionary algorithm for solving complex multi-modal optimization problems [37]. DE does not use any information of the system for optimizing and is thus suitable for black-box optimization in adversarial attacks. Moreover, the objective function involved does not have to be differentiable nor have an analytical form [38]. Thus, it matches the goal to solve (5) in this paper.

The key step of DE lies in the population selection that keep the diversity [39]. In specific, during each iteration another set of candidate solutions (children) is generated according to the current population (parents). Then the children are compared with their corresponding parents, surviving only if they are more fitted (possess higher fitness value, i.e., the objective function (5) in this paper) than their parents. In such a way, by only comparing the fitness of parent and his child, the goal of keeping diversity and improving optimization can be achieved simultaneously [40].

Experimental settings

Data sets and speaker recognition models

Experimental evaluation of the proposed methods is conducted on the commonly used data sets for speaker recognition, say VoxCeleb1 [41], VoxCeleb2 [42] and LibriSpeech [43]. VoxCeleb1 is a large-scale text-independent speaker

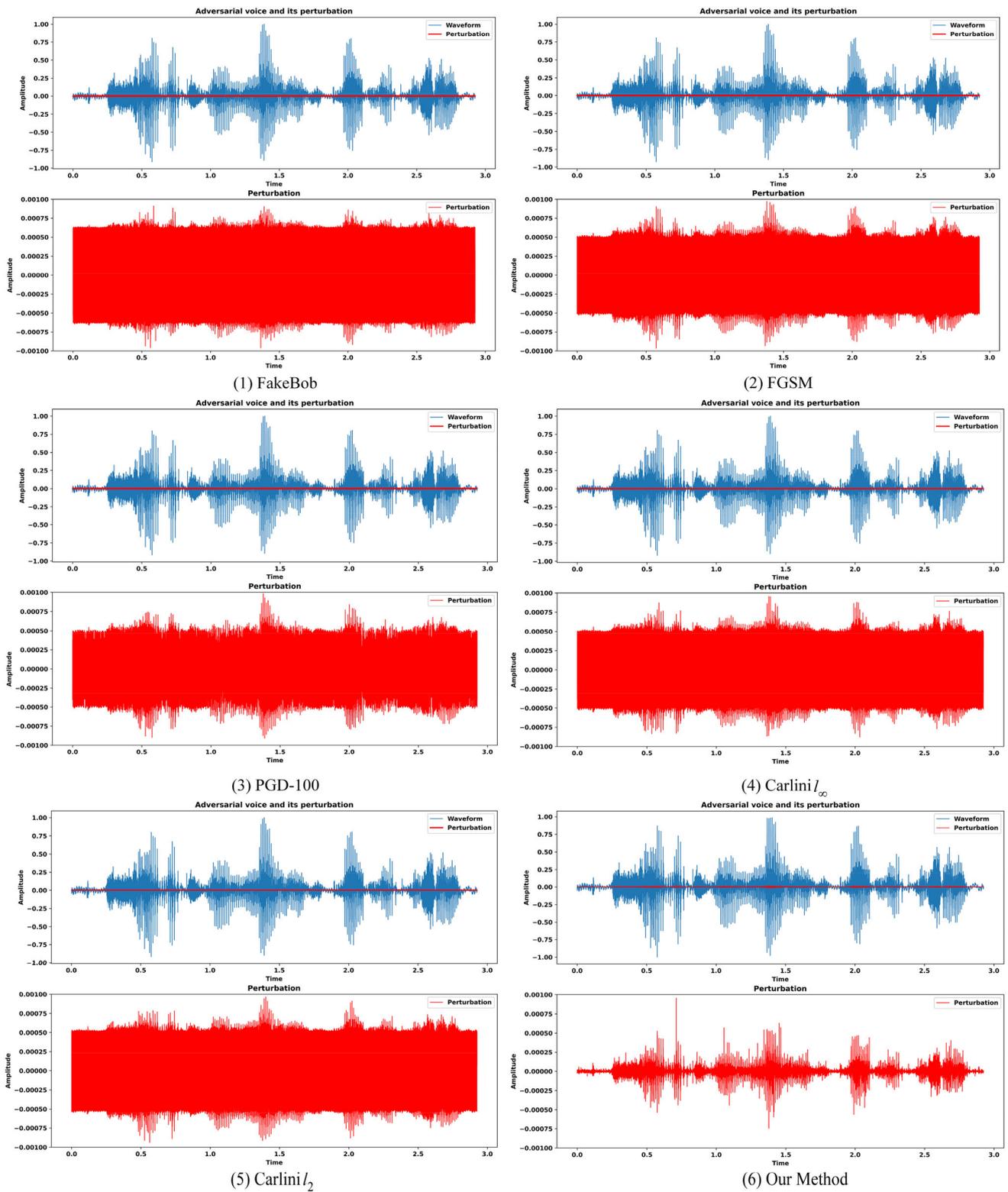


Fig. 2 Comparison of the proposed method w.r.t. the baselines on adversarial attacks of CSI. In each subgraph, the top one is an overview of the adversarial waveform (i.e., the blue line represents perturbed waveform and the red line represents the perturbations), while the bottom one is

the zoomed-in view of the crafted perturbations. It is seen that our proposed method puts perturbations according to the energy distribution of the waveform to improve imperceptibility

Table 1 The distribution of *Imposter* and *Test* sets

Dataset	Gender	Speaker IDs
Imposter set	M	61, 260, 672, 908, 1089, 1188, 1320, 2300, 2830, 4077
	F	121, 237, 1221, 1284, 1580, 1995, 2094, 2961, 3570, 3575
Test set	M	5105, 5639, 6930, 7021, 7127, 7176, 7729, 8224, 8230, 8455
	F	3729, 4446, 4507, 4970, 4992, 5142, 5683, 6829, 8463, 8555

Table 2 Configuration of speakers and adversarial examples for the three tasks (*intra-gender* + *inter-gender* in the last column)

Task	Source speaker (size)	Target speaker (size)	Num of examples
CSI	Test set (20)	Test set (20)	900 + 1000
ASV	Imposter set (20)	Test set (20)	1000 + 1000
OSI	Imposter set (20)	Test set (20)	1000 + 1000

identification data set collected in the wild, which is extracted from videos uploaded to YouTube and consists of 153,516 utterances of 1251 speakers. VoxCeleb2 has over 1 million utterances for over 6000 celebrities extracted from videos uploaded to YouTube. VoxCeleb1 and VoxCeleb2 provide a large amount of the diversity of speaker characteristics. LibriSpeech is derived from audiobooks that are part of the LibriVox project and has 1000 h of speech sampled at 16 kHz.

In this paper, the data from VoxCeleb1 (153,516 utterances from 1251 speakers) and VoxCeleb2 (1,128,246 utterances from 6112 speakers) is utilized to train the models of GMM-UBM, *i-vector* and *x-vector* for automatic speaker recognition by following the recipe provided in Kaldi [44]. This data set is named by *Train-1 Set* (1,281,762 utterances from 7363 speakers in total).

The *Imposter Set* consists of 100 utterances of 20 speakers from the test-clean subset of LibriSpeech with 5 utterances per speaker. The *Test Set* is configured in the same way but with 20 different speakers each with 5 utterances, from the test-clean subset in LibriSpeech too. In either *Imposter Set* or *Test Set*, the gender ratio is 1:1, i.e., 10 males and 10 females are considered. The speaker IDs involved are shown in Table 1 where “M” and “F” denotes male and female, respectively.

The roles of the sets of *Train-1*, *Imposter* and *Test* will be presented in Table 2 and “[Experimental design](#)” section.

Experimental design

When conducting adversarial attacks, besides the three typical tasks presented in “[The proposed methods](#)” section, genders are also considered, where both intra-gender and inter-gender attacks are configured and evaluated. Therefore, there are six combinations in total.

To comprehensively evaluate the performance of adversarial attacks including our proposed one, three groups of experiments are designed for the following purposes.

(1) *Matched attacks*. In this group of experiments, we assume the attacker uses the same speaker recognition model the victim is using. That is, the model f to obtain an adversarial sample in (6)–(8) matches the model to be attacked. In the experiments, algorithms recently proposed in [13, 14] are taken as baselines for comparison. To increase the representativeness of the experiments, the source speakers of CSI and ASV/OSI are designated to be different, i.e., source speakers in CSI are from the Test Set and those in ASV/OSI are from the Imposter Set. As shown in Table 2, for CSI, the source and target speakers are both from the *Test Set*. Therefore, 1900 adversarial examples will be crafted and evaluated in total for CSI. In this setting, $P_{10}^2 \times 5 \times 2 = 900$ examples are for intra-gender and $10 \times 10 \times 5 \times 2 = 1000$ ones are for inter-gender, where P_{10}^2 is the permutation. For OSI and ASV, the source and target speakers are taken from the *Imposter Set* and the *Test Set*, respectively. In either task, there are 2000 adversarial tests in total, where half of them are intra-gender and the remaining ones are inter-gender, as summarized in Table 2.

(2) *Transferability*. In this group of experiments, we assume the attacker does not know the speaker recognition model the victim is using, which means the model f in (6)–(8) is different from that to be attacked. To evaluate the transferability of the adversarial examples crafted by our algorithm, we treat GMM-UBM, *i-vector* and *x-vector* trained on *Train-1* (system A, B and C, respectively) as the source systems f . Other systems with different architectures or different training set or both are taken as targeted victim systems. A portion of LibriSpeech (train-clean-100 subset) is utilized to train the target victim models of GMM-UBM, *i-vector* and *x-vector*, named by *Train-2 Set* (28,539 utterances from 251 speakers). The detailed configurations are shown in Table 3 where I, II, VI, VII, XI and XII are attacks cross different models. III, IX and XV are cross training data. IV, V, VIII, X, XIII and XIV are cross both model and data.

(3) *Robustness*. As shown in (6)–(8), there are three key hyper parameters in our proposed method, i.e., the maximum number of data points to be modified d , the maximum

Table 3 Experimental configuration for evaluating transferability

Source system	Target system	Attack ID
GMM-UBM ¹	<i>i-vector</i> ¹	I
	<i>x-vector</i> ¹	II
	GMM-UBM ²	III
	<i>i-vector</i> ²	IV
	<i>x-vector</i> ²	V
<i>i-vector</i> ¹	GMM-UBM ¹	VI
	<i>x-vector</i> ¹	VII
	GMM-UBM ²	VIII
	<i>i-vector</i> ²	IX
	<i>x-vector</i> ²	X
<i>x-vector</i> ¹	GMM-UBM ¹	XI
	<i>i-vector</i> ¹	XII
	GMM-UBM ²	XIII
	<i>i-vector</i> ²	XIV
	<i>x-vector</i> ²	XV

Superscripts 1 or 2 mean a speaker recognition system is trained on Train-1 or Train-2, respectively

amplitude to be modified ζ , and the regularization parameter λ . The parameter d controls the amount of perturbation, ζ is a parameter which controls the maximum amplitude of perturbation, while the regularization parameter λ weights the amplitude of perturbation. Therefore, it is necessary to analyze the influence of the parameters and to provide suggestions on the choices of them.

Evaluation metrics

To give a comprehensive quantitative evaluation on adversarial audio samples, the following three metrics are introduced to measure *efficacy*, *quality* and *imperceptibility*, respectively.

- **Targeted attack success rate (TASR)**—TASR reflects the efficacy of adversarial examples on attacking speaker recognition systems. It is defined as the ratio between the number of successful attacks and the total number of attempts.
- **Signal-to-noise ratio (SNR)**—SNR is proportional to the log ratio between the power of the source voice \mathbf{x} over the power of the perturbation $e(\mathbf{x})$. High SNR reflects the high quality of an adversarial sample, which is consistent with the imperceptibility as motivated in this paper.
- **Perceptual evaluation of speech quality (PESQ)**—PESQ is a commonly used metric for evaluating speech enhancement. In this paper, given the assumption that \mathbf{x} is a clean signal, $\hat{\mathbf{x}}$ would better also hold a high perceptual quality.

Otherwise, the difference between $\hat{\mathbf{x}}$ and \mathbf{x} can be noticed and the adversarial attack would fail by simply listening to the sound.

Results and analysis

In this section, we report and analyze the results of the experiments designed in “[Experimental design](#)” section.

Evaluation of matched attacks

Parameter settings

The threshold θ in ASV and OSI can be estimated by the approach proposed in [14]. The thresholds for different tasks of the same system keep unchanged. As a trade-off between effectiveness and imperceptibility, we choose $d = L/3$, $\lambda = 0.005$ and $\zeta = 0.003$ in the experiments unless otherwise specified, where L represents the length of the utterance. The analysis of choosing λ and d will be discussed in Sect. D.

Performance on adversarial attacks

The results for GMM-UBM, *i-vector* and *x-vector* are shown in Tables 4, 5 and 6, respectively. *No effort attack* means utilizing the unmodified source voice to attack a speaker recognition system. It can be seen that our proposed method is effective on deceiving speaker recognition systems by yielding over 69% success rates on all the attempts even on the most difficult one, OSI. In terms of SNR and PESQ, the average SNR value is around 40 dB and the average PESQ value is around 4.00, which indicate high quality of the adversarial samples and thus high imperceptibility. Inter-gender attacks are slightly difficult than intra-gender ones due to the difference between male voices and female voices, as expected.

By comparing the values reported in Tables 4, 5 and 6, it is observed that *x-vector* is the most difficult system to be attacked, the easier ones are *i-vector* and GMM-UBM subsequently. By comparing the rows of Tables 4, 5 and 6, attacking OSI is the most difficult task compared with attacking CSI and ASV. However, the high average SNR and PESQ scores show that the adversarial samples crafted by our methods are not easily perceived by human listening as well as holding high success rates (i.e., TASR’s) on attacking. These results demonstrate the effectiveness of our method.

We have also compared the proposed method with recently proposed baselines on CSI tasks with *x-vector* as the victim speaker recognition model, e.g., Fast Gradient Sign Method (FGSM) in [13], FakeBob in [14] and PGD-100, C&W l_∞ and C&W l_2 in [13]. Unlike the white-box configuration of FGSM, PGD-100, C&W l_∞ , and C&W l_2 in the

Table 4 Performance of the proposed method for matched attacks on GMM-UBM with $d = L/3$, $\lambda = 0.005$ and $\zeta = 0.003$

Task	No effort attack	Intra-gender attack			Inter-gender attack		
	TASR (%)	TASR (%)	SNR (dB)	PESQ	TASR (%)	SNR (dB)	PESQ
CSI	0.00	92.89	50.29	4.46	89.30	46.27	4.40
ASV	0.00	88.20	45.97	4.36	82.60	42.01	4.27
OSI	0.00	84.30	44.24	4.31	82.80	42.66	4.29

Table 5 Performance of the proposed method for matched attacks on i -vector with $d = L/3$, $\lambda = 0.005$ and $\zeta = 0.003$

Task	No effort attack	Intra-gender attack			Inter-gender attack		
	TASR (%)	TASR (%)	SNR (dB)	PESQ	TASR (%)	SNR (dB)	PESQ
CSI	0.00	82.67	42.34	4.25	80.40	42.02	4.18
ASV	0.00	81.00	45.37	4.34	84.60	44.61	4.28
OSI	0.00	77.40	41.25	4.12	76.50	40.13	4.06

Table 6 Performance of the proposed method for matched attacks on x -vector with $d = L/3$, $\lambda = 0.005$ and $\zeta = 0.003$

Task	No effort attack	Intra-gender attack			Inter-gender attack		
	TASR (%)	TASR (%)	SNR (dB)	PESQ	TASR (%)	SNR (dB)	PESQ
CSI	0.00	74.56	39.81	4.09	71.50	38.54	4.07
ASV	0.00	78.50	41.50	4.14	74.10	41.09	4.15
OSI	0.00	71.30	36.86	3.92	69.90	36.81	3.90

Table 7 Our method w.r.t. recently proposed baselines on matched CSI using x -vector with $d = L/3$, $\lambda = 0.005$ and $\zeta = 0.003$

Methods	TASR (%)	SNR (dB)	PESQ
FGSM [13]	72.47	31.34	3.07
FakeBob [14]	72.79	32.01	3.16
PGD-100 [13]	73.05	33.59	3.52
C&W l_∞ [13]	72.58	35.12	3.81
C&W l_2 [13]	72.84	32.66	3.74
Our Method	72.95	39.14	4.08

Bold values indicate the best values in the same column

referred works, an additional x -vector model is trained as a substitution of the victim model to fulfill the black-box attack configuration in this paper. The experiments are thus black-box attacks with matched speaker recognition systems between attackers and victims. The results on TASR, SNR and PESQ are shown in Table 7. As seen from the table, our method has greatly improved the quality of adversarial samples on SNR and PESQ with a tiny sacrifice on TASR, which demonstrates the advantage on the imperceptibility of our method.¹

¹ Audio samples can be found here: <https://attackasv.github.io>.

Visualization

Adversarial samples for a successful attack and the corresponding perturbations are shown in Fig. 2. It is seen that the baseline methods tend to add uniform perturbations through the temporal axis, which would be noticed by listening to the non-speech segments carefully. However, our method adds more perturbations on the data points with high energy, while adds less perturbations on the data points with low energy. This will reduce the chances being noticed by human perception. For quantitative evaluation, the background noise is suppressed in the generated adversarial examples by our method as measured by SNR and PESQ in Table 7.

Towards intuitive interpretation of the adversarial process, t-SNE is taken as the dimension reduction tool to visualize the i -vectors in Fig. 3 and x -vectors in Fig. 4 involved in attacking CSI. t-SNE (t-distributed Stochastic Neighbor Embedding) is a technique that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map [45]. In Figs. 3 and 4, source voices are denoted by circles with five different colors with each color for one speaker. Adversarial examples are denoted by triangles with five different colors, where the color indicates the source speaker of an adversarial example. It can be seen from the figures that targeted adversarial attack pulls source voice into the cluster of the target speaker. In addition, x -vector clusters in Fig. 4 hold larger margins than those of i -vectors in Fig. 3,

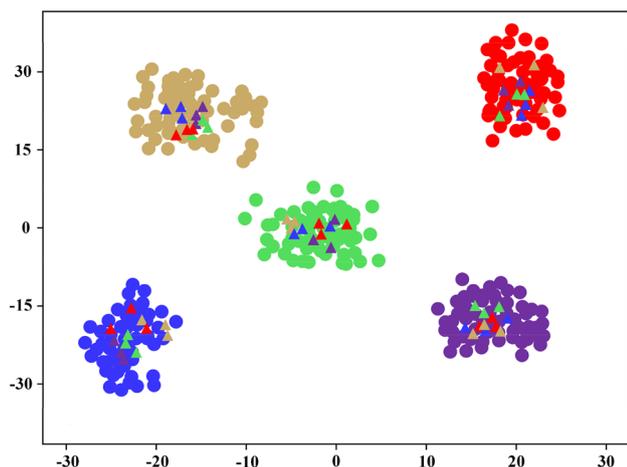


Fig. 3 A t-SNE figure to visualize *i-vectors* of source utterances and adversarial examples. Source voices are denoted by circles with five different colors with each color for one speaker. Adversarial examples are denoted by triangles with five different colors where the color indicates the source speaker of an adversarial example (Better viewed in color)

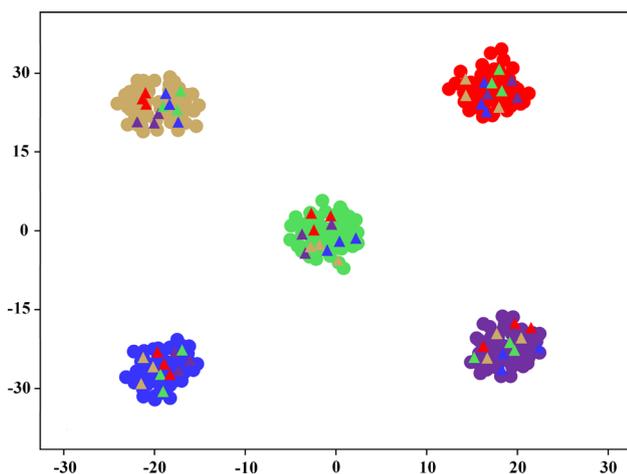


Fig. 4 A t-SNE figure to visualize *x-vectors* of source utterances and adversarial examples. Source voices are denoted by circles with five different colors with each color for one speaker. Adversarial examples are denoted by triangles with five different colors where the color indicates the source speaker of an adversarial example (Better viewed in color)

which explains the drop of TASR's from attacking *x-vector* to attacking *i-vector* from Tables 5 to 6.

Evaluation of transferability

Transferability cross model or data

The transferability of our method is evaluated as shown in Table 8. Though degradation of performance is observed in all the cross-model (i.e., I, II, VI, VII, XI and XII in Table

8), cross-data (i.e., III, IX and XV in Table 8) and cross-model-and-data (i.e., IV, V, VIII, X, XIII and XIV in Table 8) attacks, over 59% TASR can still be obtained with relatively high SNRs and PESQs.

The degradation is consistent with the results reported in other domains of adversarial attacks, where the distinction of the substitute model from the target model has a negative impact on black-box attacks, even the target model shares the same classifier as the source one but is trained on different data sets. As seen from Table 8, the adversarial samples trained on *x-vector* system have stronger transferability than GMM-UBM and *i-vector*, which reveals that it is better to choose a substitute model with high performance on speaker recognition when generating adversarial examples. It is also observed that the match of models is much more important than the match of training data to conduct successful attacks. In an attacker's view, he/she should learn as much as possible knowledge of the target victim model. While in a defender's view, any information on the speaker recognition model should be protected carefully, including the model of choice, configuration of models, parameters, and training data etc.

Transfer attack to the commercial system Microsoft Azure

Microsoft Azure is a cloud-based operating system supported by Microsoft. The main goal of Microsoft Azure is to provide a platform for developers to develop applications that can run on cloud servers, data centers, the Web, and PCs [46]. It supports both the ASV and OSI tasks via online API (Application Programming Interface). Since ASV task on Microsoft Azure is text-dependent, we have only tested the attack on OSI, where 20 speakers from the *Test Set* are enrolled in the OSI system of Microsoft Azure. The baseline performance of this OSI system is tested by utilizing 50 original voices from *Imposter Set* and the TASR is 0%. We then attack Microsoft Azure using 50 adversarial voices crafted using GMM-UBM¹, *i-vector*¹ and *x-vector*¹ systems as the source systems in Table 3, where the TASRs are 42%, 24% and 30%, respectively. The top TASR is from GMM-UBM, which may indicate that the speaker recognition API on Microsoft Azure shares much similarities with GMM-UBM.

Evaluation of parameter sensitivity

In this section, we discuss the selection of parameters that control the strength and length of adversarial samples, and verify the validity of DE algorithm.

Table 8 Evaluation of transferability for targeted attacks on CSI task with $d = L/3$, $\lambda = 0.005$ and $\zeta = 0.003$

Source	Target	TASR (%)	SNR (dB)	PESQ	Attack ID
GMM-UBM ¹	<i>i-vector</i> ¹	75.89	40.02	4.11	I
	<i>x-vector</i> ¹	63.74	35.28	3.90	II
	GMM-UBM ²	88.00	46.30	4.41	III
	<i>i-vector</i> ²	71.47	38.34	4.06	IV
	<i>x-vector</i> ²	59.58	34.19	3.88	V
<i>i-vector</i> ¹	Mean	71.74	38.83	4.07	
	GMM-UBM ¹	86.32	45.61	4.34	VI
	<i>x-vector</i> ¹	66.26	36.25	3.92	VII
	GMM-UBM ²	82.11	42.47	4.25	VIII
	<i>i-vector</i> ²	77.63	40.14	4.10	IX
<i>x-vector</i> ¹	<i>x-vector</i> ²	64.37	35.96	3.93	X
	Mean	75.34	40.09	4.11	
	GMM-UBM ¹	87.84	45.12	4.31	XI
	<i>i-vector</i> ¹	75.58	39.53	4.04	XII
	GMM-UBM ²	85.16	44.39	4.18	XIII
<i>i-vector</i> ²	<i>i-vector</i> ²	73.53	38.90	4.09	XIV
	<i>x-vector</i> ²	68.42	37.11	3.90	XV
	Mean	78.11	41.01	4.11	

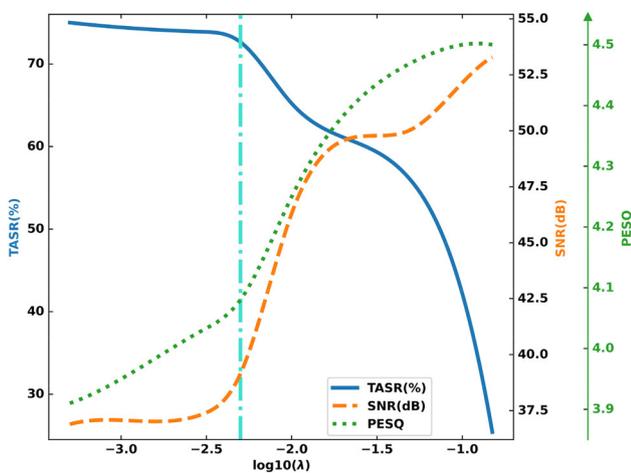


Fig. 5 Parameter sensitivity with respect to λ on CSI with *x-vector* as the speaker recognition model. The vertical line indicates the performance with $\lambda = 0.005$. d and ζ are fixed to $L/3$ and 0.003, respectively

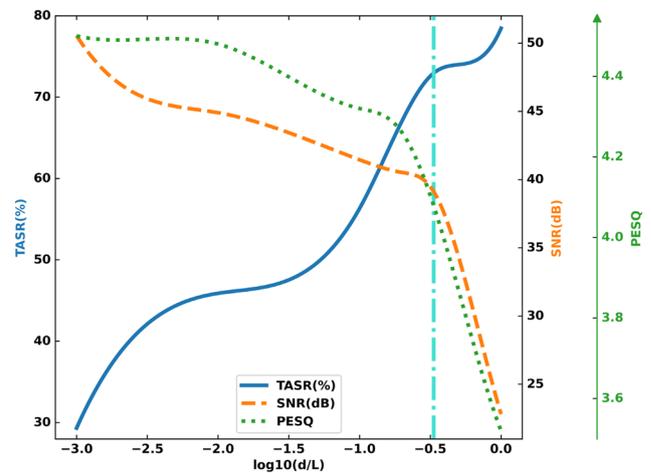


Fig. 6 Parameter sensitivity with respect to d on CSI with *x-vector* as the speaker recognition model. The vertical line indicates the performance with $d = L/3$. λ and ζ are fixed to 0.005 and 0.003, respectively

Results of changing parameters λ , d and ζ

Figures 5, 6 and 7 show the results of parameter sensitivity regarding to λ , d and ζ on CSI with *x-vector* as the speaker recognition model. In summary, three constant variables, $d = L/3$, $\lambda = 0.005$, and $\zeta = 0.003$ are taken for these experiments, where we change one variable by fixing the remaining two.

Higher λ and lower d or ζ tend to yield adversarial samples with higher quality (i.e., higher SNR or PESQ) with some sacrifice on attacking efficiency (i.e., lower TASR). It is found

that $\lambda = 0.005$, $d = L/3$ and $\zeta = 0.003$ made a relatively good balance between efficiency and quality.

The role of the regularization term

Table 9 shows the results with and without the regularization term in (9). The regularization term plays a role of optimizing the location of perturbations to be added, which is important to reduce the perception on perturbation noises and hence to improve the quality of adversarial samples. When removing

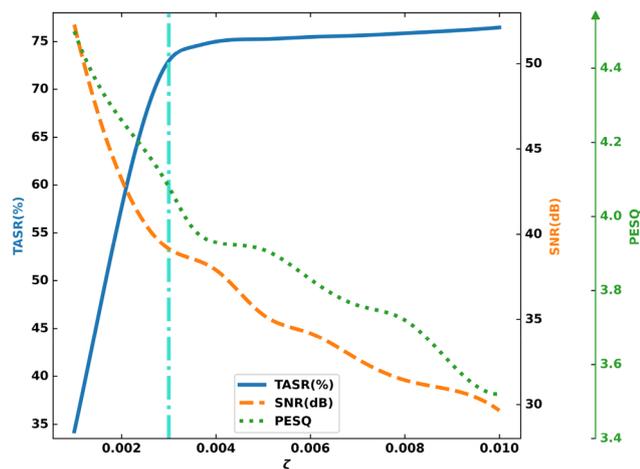


Fig. 7 Parameter sensitivity with respect to ζ on CSI with x -vector as the speaker recognition model. The vertical line indicates the performance with $\zeta = 0.003 \cdot \lambda$ and d are fixed to 0.005 and $L/3$, respectively

Table 9 The roles of regularization when conducting matched attacks on x -vector (with/without regularization)

Tasks	TASR (%)	SNR (dB)	PESQ
CSI	72.95/75.00	39.14/35.62	4.08/3.88
ASV	76.30/78.90	41.30/37.19	4.15/3.92
OSI	70.30/74.80	36.84/34.26	3.91/3.78

this term, the quality of the generated adversarial examples drops significantly in SNR and PESQ, with a slight increase of TASR, as shown in Table 9. These results demonstrate the effectiveness of regularization term on generating imperceptible adversarial samples as motivated in the Introduction.

Table 10 Results of white-box attacks

Tasks	Attack algorithm	TASR (%)	SNR (dB)	PESQ
CSI	FGSM [13]	97.28	33.17	3.25
	PGD-100 [13]	100.00	35.24	3.63
	C&W l_∞ [13]	100.00	38.21	4.00
	C&W l_2 [13]	100.00	35.66	3.87
ASV	FGSM [13]	98.59	34.52	3.31
	PGD-100 [13]	100.00	36.89	3.67
	C&W l_∞ [13]	100.00	39.23	4.08
	C&W l_2 [13]	100.00	38.11	4.03
OSI	FGSM [13]	96.11	31.47	3.04
	PGD-100 [13]	100.00	34.28	3.55
	C&W l_∞ [13]	100.00	37.57	3.96
	C&W l_2 [13]	100.00	35.93	3.88

Replacing black-box attacks by white-box ones

Table 10 shows the performance of white-box attacks using the differential evolution algorithm in the white-box versions of FGSM, PGD and C&W. Both the attacking model and the victim model is x -vector, where the TDNN is adopted to compute gradients. By comparing the numbers in Table 10 and those in Table 7, the TASR values of white-box attacks have been greatly improved compared with respect to them of black-box attacks. However, in real-world settings, attackers can seldomly know the configurations of the victim systems, i.e., the attacks are actually black-box ones.

Replacing waveform-level perturbations by feature-level ones

Since the interfaces of most ASV systems are speech waveforms, a substitution of waveform-level perturbation is to compute feature-level perturbations first and to transform the features back into waveforms subsequently. The results reported in Table 11 are obtained by computing perturbations on MFCC [8] and LPMS [9] and transforming the perturbed MFCC's and LPMS's back to waveforms (by copying the phase information of the original voice). The victim model is x -vector. Deteriorated performance is observed comparing to our proposed waveform-level method. It is worth noting that the regularization term in Eq. (9) is designed on waveform level, so we have to omit this term when conducting the adversarial experiments on the feature-levels. One may argue that additional regularization terms can be invented on these features, but they are out the scope of our paper.

The extracting processes of MFCC and LPMS can actually lose some information when conducting the transforms. The feature-level perturbations can be destroyed when passing

Table 11 The performance of the attacks using different speaker features on x -vector

Tasks	Speaker feature	TASR (%)	SNR (dB)	PESQ
CSI	Waveform	72.95	39.14	4.08
	MFCC	64.87	30.26	2.98
	LPMS	59.25	24.19	2.56
ASV	Waveform	76.30	41.30	4.15
	MFCC	69.21	29.84	2.86
	LPMS	60.10	26.53	2.74
OSI	Waveform	70.30	36.84	3.91
	MFCC	58.77	28.49	2.82
	LPMS	53.26	23.18	2.46

Table 12 Abbreviations in the paper

Abbreviations	Full name
API	Application programming interface
ASV	Automatic speaker verification
BIM	Basic iterative method
CSI	Close-set identification
C&W	Carlini and Wagner attacks
DE	Differential evolution
FGSM	Fast gradient sign method
GMM-UBM	Gaussian mixture model-universal background model
JSMA	Jacobian-based saliency map attack
LPMS	Log power magnitude spectrum
MFCC	Mel-frequency cepstral coefficient
OSI	Open-set identification
PESQ	Perceptual evaluation of speech quality
PGD	Project gradient descent
SNR	Signal-to-noise ratio
s.t	Subject to
TASR	Targeted attack success rate
TDNN	Time delay neural networks
t-SNE	t-Distributed stochastic neighbor embedding
$\ \cdot\ _p$	p-norm

through the transform, as also analyzed in [47]. The abbreviations appeared in the paper are shown in Table 12.

Conclusions

This paper proposed a new method to conduct imperceptible black-box waveform-level targeted adversarial attacks against speaker recognition systems. Inspired by auditory masking, the imperceptibility was promoted by only modifying a part of selected sample points with a constraint

on modification amplitudes. Extensive experiments conducted on three typical tasks and state-of-the-art models of speaker recognition were conducted, and the results showed effectiveness of the proposed method. Transferability and robustness of the methods were evaluated and analyzed, which shed lights on both attackers and defenders on designing secure speaker recognition systems.

Future works may explore adaptive attack methods to improve the efficiency of the adversarial samples [48]. In addition, the adversarial attacks currently take several minutes to be conducted, which is infeasible for real-time implementations. Therefore, future attacks will have to be computationally efficient while retaining their robustness [49].

Acknowledgements This work was supported by the Natural Science Foundation of Jiangsu Province (BK20180080) and the National Natural Science Foundation of China (62071484, U20B2047).

Declarations

Conflict of interest The authors declare that they do not have any conflicts of interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Ren H, Song Y, Yang S, Situ F (2016) Secure smart home: a voiceprint and internet-based authentication system for remote

- accessing. In Proc. 2016 11th international conference on computer science and education (ICCSE), Nagoya, Japan, Aug. 2016, pp 247–251
2. Granqvist F, Seigel M, van Dalen R, Cahill A, Shum S, Paulik M (2020) Improving on-device speaker verification using federated learning with privacy. In: Proc. 2020 21th annual conference of the international speech communication association (INTERSPEECH), Shanghai, China, Oct. 2020
 3. Hansen JH, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. *IEEE Signal Process Mag* 32(6):74–99
 4. Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N, Yamagishi J, Lee KA (2017) The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection. In: Proc. 2017 18th annual conference of the international speech communication association (INTERSPEECH), Stockholm, Sweden, Aug. 2017, pp 2–6
 5. Todisco M, Wang X, Vestman V, Sahidullah M, Delgado H, Nautsch A, Yamagishi J, Evans N, Kinnunen T, Lee KA (2019) ASVspoof 2019: future horizons in spoofed and fake audio detection. In: Proc. 2019 20th annual conference of the international speech communication association (INTERSPEECH), Graz, Austria, Sep. 2019
 6. Lorenzo-Trueba J, Yamagishi J, Toda T, Saito D, Villavicencio F, Kinnunen T, Ling T (2018) The voice conversion challenge 2018: promoting development of parallel and nonparallel methods. In: Proc. 2018 19th annual conference of the international speech communication association (INTERSPEECH), Hyderabad, India, Sep. 2018
 7. Voice Conversion Challenge (2020) Accessed Oct. 2020. <https://vc-challenge.org>
 8. Kreuk F, Adi Y, Cisse M, Keshet J (2018) Fooling end-to-end speaker verification with adversarial examples. In: Proc. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), Calgary, AB, Canada, Apr. 2018, pp 1962–1966
 9. Li X, Zhong J, Wu X, Yu J, Liu X, Meng H (2020) Adversarial attacks on GMM I-vector based speaker verification systems. In: Proc. 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), Barcelona, Spain, May 2020, pp 6579–6583
 10. Li Z, Shi C, Xie Y, Liu J, Yuan B, Chen Y (2020) Practical adversarial attacks against speaker recognition systems. In: Proc. 21st international workshop on mobile computing systems and applications (ACM Hot Mobile), Austin, Texas, USA, Mar. 2020, pp 9–14
 11. Xie Y, Shi C, Li Z, Liu J, Chen Y, Yuan B (2020) Real-time, universal and robust adversarial attacks against speaker recognition systems. In: Proc. 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), Barcelona, Spain, May 2020, pp 1738–1742
 12. Wang Q, Guo P, Xie L (2020) Inaudible adversarial perturbations for targeted attack in speaker recognition. In: Proc. 2020 21th annual conference of the international speech communication association (INTERSPEECH), Shanghai, China, Oct. 2020
 13. Jati A, Hsu CC, Pal M, Peri R, Abd Almageed W, Narayanan S (2021) Adversarial attack and defense strategies for deep speaker recognition systems. *Comp Speech Lang* 68(101199)
 14. Chen G, Chen S, Fan L, Du X, Zhao Z, Song F, Liu Y (2021) Who is real bob? Adversarial attacks on speaker recognition systems. In: Proc. 2021 IEEE symposium on security and privacy workshops (SPW), San Francisco, CA, USA, May 2021
 15. Abdullah H, Garcia W, Peeters C, Traynor P, Butler KRB, Wilson J (2019) Practical hidden voice attacks against speech and speaker recognition systems. In: Proc. Network and Distributed Systems Security (NDSS), San Diego, United States, Feb. 2019
 16. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. In: Proc. 2nd international conference on learning representations (ICLR), Banff, Canada, Apr. 2014
 17. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: Proc. 3rd international conference on learning representations (ICLR), Toronto, Canada, Jul. 2015
 18. Kurakin A, Goodfellow I, Bengio S (2017) Adversarial examples in the physical world. In: Proc. 5th international conference on learning representations (ICLR), Toulon, France, Apr. 2017
 19. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016) The Limitations of Deep Learning in Adversarial Settings. In: Proc. 2016 IEEE european symposium on security and privacy (Euro S&P), Saarbrücken, Germany, Mar. 2016, pp 372–387
 20. Su J, Vargas DV, Sakurai K (2019) One pixel attack for fooling deep neural networks. *IEEE Trans Evol Comput* 23(5):828–841
 21. Carlini N, Wagner D (2017) Towards Evaluating the Robustness of Neural Networks. In: Proc. 2017 symposium on IEEE security and privacy workshops (SPW), San Jose, CA, USA, May 2017, pp 39–57
 22. Moosavi-Dezfooli SM, Fawzi A, Frossard P (2016) DeepFool: a simple and accurate method to fool deep neural networks. In: Proc. 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, Nevada, USA, Jun. 2016, pp 2574–2582
 23. Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P (2017) Universal Adversarial Perturbations. In: Proc. 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, Hawaii, USA, Jul. 2017, pp 1765–1773
 24. Das RK, Tian X, Kinnunen T, Li H (2020) The attacker’s perspective on automatic speaker verification: an overview. In: Proc. 2020 21th annual conference of the international speech communication association (INTERSPEECH), Shanghai, China, Oct. 2020
 25. Zhang Z, Geiger J, Pohjalainen J, Mousa AED, Jin W, Schuller B (2018) Deep learning for environmentally robust speech recognition: an overview of recent developments. *ACM Trans Intell Syst Technol (TIST)* 9(5):1–28
 26. Safavi S, Gan H, Mporas I, Sotudeh R (2016) Fraud detection in voice-based identity authentication applications and services. In: 2016 IEEE 16th international conference on data mining workshops (ICDMW), Barcelona, Spain, Dec. 2016, pp 1074–1081
 27. Yuan X, He P, Zhu Q, Li X (2019) Adversarial examples: attacks and defenses for deep learning. *IEEE Trans Neural Netw Learn Syst* 30(9):2805–2824
 28. Qin Y, Carlini N, Goodfellow I, Cottrell G, Raffel C (2019) Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: Proc. 2019 36th international conference on machine learning (PMLR), Long Beach, California, 2019
 29. Schonherr L, Kohls K, Zeiler S, Holz T, Kolossa D (2019) Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. In: Proc. 2019 network and distributed system security symposium (NDSS), San Diego, California, Feb. 2019
 30. Ilyas A, Engstrom L, Athalye A, Lin J (2018) Black-box adversarial attacks with limited queries and information. In: Proc. 2018 35th international conference on machine learning (ICML), Stockholm, Sweden, Jul. 2018, pp 2137–2146
 31. Wilkinghoff K (2020) On open-set speaker identification with I-vectors. In: Proc. Odyssey 2020 the speaker and language recognition workshop, Tokyo, Japan, May 2020, pp 408–414
 32. Liu T, Guan S (2014) Factor analysis method for text-independent speaker identification. *J Softw (JSW)* 9(11):2851–2860
 33. Snyder D, Garcia-Romero D, Povey D, Khudanpur S (2017) Deep neural network embeddings for text-independent speaker verification. In: Proc. 2017 18th annual conference of the international speech communication association (INTERSPEECH), Stockholm, Sweden, Aug. 2017, pp 999–1003
 34. Cumani S, Plchot O, Laface P (2014) On the use of I-vector posterior distributions in probabilistic linear discriminant analysis. *IEEE/ACM Trans Audio Speech Lang Process* 22(4):846–857

35. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) X-vectors: robust DNN embeddings for speaker recognition. In: Proc. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), Calgary, AB, Canada, Apr. 2018, pp 5329–5333
36. Gelfand SA (2017) Hearing: An introduction to psychological and physiological acoustics, 6th edn. CRC Press, Boca Raton, FL, USA
37. Opara KR, Arabas J (2019) Differential evolution: a survey of theoretical analyses. *Swarm Evol Comput* 44:546–558
38. Das S, Mullick SS, Suganthan PN (2016) Recent advances in differential evolution—an updated survey. *Swarm Evol Comput* 27:1–30
39. Mashwani WK (2014) Enhanced versions of differential evolution: state of the art survey. *Int J Comput Sci Math* 5(2):107–126
40. Tang L, Dong Y, Liu J (2015) Differential evolution with an individual-dependent mechanism. *IEEE Trans Evol Comput* 19(4):560–574
41. Nagrani A, Chung JS, Zisserman A (2017) VoxCeleb: a large-scale speaker identification dataset. In: Proc. 2017 18th conference of the international speech communication association (INTERSPEECH), Stockholm, Sweden, Aug. 2017, pp 2616–2620
42. Chung JS, Nagrani A, Zisserman A VoxCeleb2: deep speaker recognition. In: Proc. 2018 19th conference of the international speech communication association (INTERSPEECH), Hyderabad, India, Sept. 2018, pp 1086–1090
43. Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an ASR corpus based on public domain audio books. In: Proc. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), Brisbane, Australia, Apr. 2015, pp 5206–5210
44. Kaldi. Accessed: Nov. 2019. <https://github.com/kaldi-asr/kaldi>
45. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(11)
46. Microsoft Azure. Accessed: Mar. 2022. <https://azure.microsoft.com/zh-cn>
47. Gong Y, Poellabauer C (2018) Crafting adversarial examples for speech paralinguistics applications. In: Proc. of dynamic and novel advances in machine learning and intelligent cyber security (DYNAMICS) Workshop, San Juan, Puerto Rico, USA, 2018
48. Tramer F, Carlini N, Brendel W, Madry A (2020) On adaptive attacks to adversarial example defenses. *Adv Neural Inf Process Syst* 33:1633–1645
49. Carlini N, Mishra P, Vaidya T, Zhang Y, Sherr M, Shields C, Wagner D, Zhou W (2016) Hidden voice commands. In: 25th USENIX security symposium (USENIX Security 16), Austin, TX, USA, Aug. 2016, pp 513–530

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.