Learning to Fool the Speaker Recognition

JIGUO LI, Institute of Computing Technology, Chinese Academy of Sciences & University of Chinese Academy of Sciences, China
XINFENG ZHANG, University of Chinese Academy of Sciences, China
JIZHENG XU, Bytedance Inc, China
SIWEI MA*, Peking University, China
WEN GAO, Peking University, China

Due to the widespread deployment of fingerprint/face/speaker recognition systems, the risk in these systems, especially the adversarial attack, has drawn increasing attention in recent years. Previous researches mainly studied the adversarial attack to the vision-based systems, such as fingerprint and face recognition. While the attack for speech-based systems has not been well-studied yet, although it has been widely used in our daily life. In this paper, we attempt to fool the state-of-the-art speaker recognition model and present *speaker recognition attacker*, a lightweight multi-layer convolutional neural network to fool the well-trained state-of-the-art speaker recognition model by adding imperceptible perturbations onto the raw speech waveform. We find that the speaker recognition system is vulnerable to the adversarial attack, and achieve a high success rate on both the non-targeted attack and targeted attack. Besides, we present an effective method by leveraging a pretrained phoneme recognition model to optimize the speaker recognition attacker to obtain a trade-off between the attack success rate and the perceptual quality. Experimental results on the TIMIT and LibriSpeech datasets demonstrate the effectiveness and efficiency of our proposed model. And the experiments for frequency analysis indicate that high frequency attack is more effective than low frequency attack, which is different from the conclusion drawn in previous image-based works. Additionally, the ablation study gives more insights into our model.

CCS Concepts: • **Computing methodologies** → **Neural networks**; **Multi-task learning**; **Speech recognition**.

Additional Key Words and Phrases: Audio forensics, adversarial attack, deep neural network

ACM Reference Format:
Jiguo Li, Xinfeng Zhang, Jizheng Xu, Siwei Ma, and Wen Gao. 2021. Learning to Fool the Speaker Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (January 2021), 22 pages. https://doi.org/10.1145/3468673

*Siwei Ma (swma@pku.edu.cn) is the corresponding author.

Authors' addresses: Jiguo Li, Institute of Computing Technology, Chinese Academy of Sciences & University of Chinese Academy of Sciences, Beijing, China, jiguo.li@vipl.ict.ac.cn; Xinfeng Zhang, University of Chinese Academy of Sciences, Beijing, China, xfzhang@ucas.ac.cn; Jizheng Xu, Bytedance Inc, Beijing, China, xujizheng@bytedance.com; Siwei Ma, Peking University, Beijing, China, swma@pku.edu.cn; Wen Gao, Peking University, Beijing, China, wgao@pku.edu.cn.

1551-6857/2021/1-ART1 $15.00
https://doi.org/10.1145/3468673

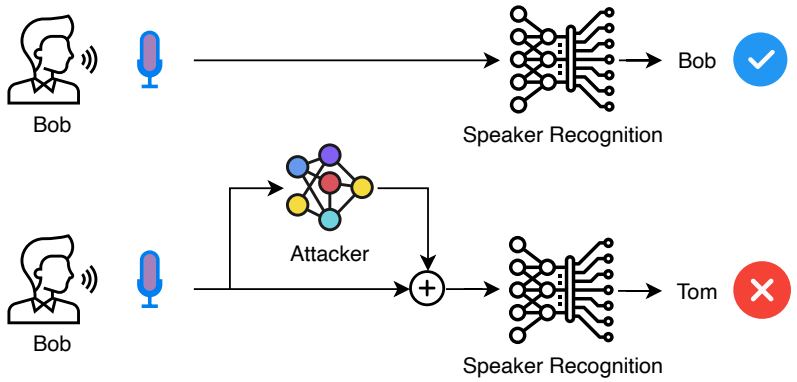

Fig. 1. Illustration of the speaker recognition attacker. An attacker network is used before the speaker recognition model to add perturbations on the input speech to spoof the following speaker recognition model.

1 INTRODUCTION

Deep neural network based biometric systems, such as fingerprint/face/speaker recognition, have been widely deployed in our daily life. Meanwhile, finding the weakness and adversarially attacking these recognition systems also draw more and more attention. Although many works have been done on vision-based systems [6, 18, 44], the adversarial attack to the speaker recognition systems has not been well-studied. There are two main applications of attacking the speaker recognition systems and finding the adversarial examples: (1) disturbing the speaker recognition systems when they are not wanted; (2) helping improve the performance and robustness of the speaker recognition systems. In this work, we focus on adversarially attacking the speaker recognition models and present an attack model as well as its optimization method to attack the well-trained state-of-the-art deep speaker recognition model by adding the perturbations on the input speech, as illustrated in Fig. 1.

Attacking the deep neural networks (DNNs) has become an emerging topic along with the development of DNNs since the weaknesses of DNNs have been found by Szegedy *et al.* [44]. On the vision tasks, some optimization methods, such as L-BFGS [44], Adam [9], or genetic algorithm [43], are used to modify the input pixels to obtain the adversarial examples. But these methods need the gradient or iterations during the testing phase, which are not practical in realistic scenarios. Baluja *et al.* [6] proposed adversarial transformation networks (ATNs), which create a separated attacker network, to transform all inputs into adversarial samples. ATNs are fast and need neither gradient nor iteration in the inference stage, so they have the potential for practical unillities.

Motivated by primary works on the adversarial attack for vision tasks, some methods are proposed to attack the automatic speech recognition (ASR) model. Alzantot *et al.* [3] proposed to attack the ASR model via a gradient-free genetic algorithm to generate adversarial examples iteratively. However, different from images, the psychoacoustic model shows that no difference will be perceived by humans if the distortion is under certain hearing thresholds. Therefore, Schonherr *et al.* [38] proposed to optimize the attack with the psychoacoustic model, and add perturbations under the hearing threshold. Following this work, Szurley *et al.* [45] demonstrates a physically realizable audio adversarial attack model, showing that the attack model is feasible in the real scenarios.

Our work has two major different aspects from previous speech based attack works. On one hand, our work mainly focuses on attacking the speaker recognition model, which has not been well-studied, because most of the previous works focused on the ASR model. On the other hand, our model is based on ATNs, which need neither gradient nor iteration during the testing phase, while all the previous works are iterative or gradient-dependent in the inference phase. Besides, our work is different from the existing literatures about speaker recognition attack and defense [22]. Because they mainly focused on the replaying based attack, while our work is based on learning techniques and attacks the well-trained speaker recognition model by adding the perturbations on the input speech data. Recently, Gong *et al.* [17] and Kreuk *et al.* [23] attempted to attack the speaker recognition model using FGSM. However, FGSM-based methods need gradients in the testing phase, which are not practical in real scenarios. In this paper, we attempt to use an attacker network to fool the following speaker recognition model by adding adversarial perturbations on the input speech data. The main contributions of this paper can be summarized as follows:

- A lightweight attack network is designed and used to attack the state-of-the-art speaker recognition models with high success rate [25]. Moreover, our attack method needs neither gradient nor iteration in the testing stage, so it is fast enough for the real-time attack. Due to the lack of systematical work for speaker recognition adversarial attacks, our work provides a benchmark for this emerging topic.
- An optimization method is proposed to train the attack model. Experimental results show that our method can achieve a better trade-off between the attack success rate and the perceptual quality on both the non-targeted and targeted attack than traditional optimization methods.
- Elaborate experiments are conducted to demonstrate the effectiveness of our proposed methods. Comparison experiments with prior works show that our proposed method can achieve better perceptual quality when a similar attack success rate is obtained. Moreover, by analyzing the frequency domain properties of the adversarial perturbations for speaker recognition, we find that high frequency attack can achieve better attack performance than low frequency attack, which is different from the conclusion drawn in image classification attack.

The rest of this paper is organized as follows: Section 2 briefly reviews the related works on image-based attack, speech-based attack, and speaker recognition; Section 3 introduces our proposed attack model and its optimization method; Section 4 presents the experimental results of our model on both non-targeted and targeted attack, as well as the frequency domain analysis; Finally, Section 5 concludes this paper and introduces some future directions about the adversarial attack for speaker recognition.

2 RELATED WORKS

Adversarial training can be used for training to obtain more robust model [5] or attacking the well-trained model to find the model's weakness [10]. In this section, we mainly focus on the white-box adversarial attack related to our work. The image-based attack, speech-based attack, and speaker recognition are reviewed.

2.1 Attack for Image-based Systems

Image-based attack has been explored for both conventional machine learning models (*e.g. SVM, shallowed ANNs, Naive Bayes*) [7, 11] and DNNs [18, 44]. Here we mainly review some works related to our work. Szegedy *et al.* [44] first demonstrated the existence of adversarial

examples that can spoof the deep learning classification models, while the perturbations in the adversarial example are almost imperceptible for humans. To find the adversarial example, given the pair data (v, y), where v is the input image and y is the class label, the problem is formulated as:

$$\min_{v'} \|v' - v\|_p \quad \text{s.t. } f(v') = y' \neq y, \tag{1}$$

where f is a deep image classification model, y' is the class prediction for v', v' is the adversarial example respect to v. This problem formulation is widely used in the following studies about the adversarial attack. The existence of the adversarial examples indicates that the feature space defined by deep learning models is not well-matched with the human perceptual space [2]. Szegedy *et al.* [44] firstly found the adversarial examples by using box-constrained L-BFGS [15]. Subsequently, Fast Gradient Sign Method (FGSM) [18] was proposed to generate adversarial examples in one-step backpropagation. Basic Iterative Method (BIM) [24] and Projected Gradient Descent (PGD) [26], which iteratively update the input image, are equivalent to 'multi-step' FGSM. These methods generate adversarial examples by restricting ℓ_∞ norm or ℓ_2 norm, while Jacobian-based Saliency Map Attack (JSMA) [31] and one pixel attack [43] aim to restrict the ℓ_0 norm (as few as pixels) of the perturbation. To generate more than a single adversarial example with only a perturbation, Universal Adversarial Networks (UANs) [27] were presented to generate a universal perturbation that can fool a network on any input images with high probability. More attack methods about image recognition can be found in [10]. Besides, to transform any input images into the adversarial examples with a single trained model, rather than updating each input image iteratively, Adversarial Transformation Networks (ATNs) [6] were introduced by training a transformation network to add perturbations on the input images. ATNs are fast to execute because they only need one forward pass without backward, gradient, or any iteration. Moreover, ATNs can attack the target model with a high success rate while the perturbations are almost imperceptible.

More works about the image-based attack are reviewed in recent surveys [2, 46] about deep models adversarial attack.

2.2 Attack for Speech-based Systems

Attack for speech recognition. Along with the deployment of speech recognition systems, the speech-based attack has also been widely studied in recent years. The security concern about speech recognition on mobile phones or home voice assistants had been presented by the communication security community several years ago [13]. Carlini *et al.* [8] proposed to attack the speech recognition model by generating speeches that are unintelligible to human listeners but interpreted as commands by devices. They iteratively revised the input normal speech in the Mel-frequency cepstral (MFC) domain to remove the audio features that are not used in the speech recognition model but a human might use for comprehension [8]. However, the hidden audio commands are nonetheless audible. To make the hidden command inaudible absolutely, Zhang *et al.* [48] modulated the voice commands on the ultrasonic carrier, which is inaudible for humans, to spoof the popular speech recognition systems and achieved a high success rate. Besides, by leveraging the weakness of the DNNs, Yuan *et al.* [47] hid the commands in the randomly selected songs by gradient descent. The crafted songs are unnoticeable to the listeners and can be remotely deliverable through the internet and played by popular devices, which will affect a large amount of speech recognition systems in our daily life. In addition to these researches about the hiding-command attack, studies

on adding imperceptible adversarial perturbations on the input speech to mislead the speech recognition are also conducted in recent years. Alzantot *et al.* [3] proposed to attack the ASR model via a gradient-free genetic algorithm to generate adversarial examples iteratively. However, different from visual images, the psychoacoustic model shows that no difference will be perceived by humans if the distortion is under certain hearing thresholds. Therefore, Schonherr *et al.* [38] proposed to optimize the attack with the psychoacoustic model, and add perturbations under the hearing threshold. Subsequently, Qin *et al.* [32] attacked the speech recognition model iteratively based on the psychoacoustic model, achieving imperceptible adversarial attack for targeted attack.

Attack for speaker recognition. The attack for speaker recognition has been studied for several years, even when the DNNs have not been used in speaker recognition. In the Speaker Antispoofing Competition organized by Biometric group at Idiap Research Institute for the IEEE International Conference on Biometrics: Theory, Applications, and Systems in 2016 (BTAS 2016) [22], the participants were asked to propose the defense methods for three types attack: (1) direct replay attacks when a genuine data is played back; (2) synthesized speech; (3) speech created with a voice conversion algorithm. While this paper aims to find the weaknesses of the speaker recognition models and attack them using learning-based methods, which is different from the previous replaying, synthesized attacks. There are two main applications of attacking the speaker recognition systems and finding the adversarial examples: (1) disturbing the speaker recognition systems when they are not wanted; (2) helping improve the performance and robustness of the speaker recognition systems. In recent years, some learning based attack methods have been proposed to spoof the speaker recognition models. Gong *et al.* [17] and Kreuk *et al.* [23] attempted to attack the speaker recognition model using FGSM. However, FGSM based method needs gradients in the testing phase, which is unpractical in real scenarios and slow for real-time applications.

2.3 Speaker Recognition

Speaker recognition is an active research area in the audio processing community and has been applied in various fields, such as biometric authentication, forensics, and security [14, 34]. The text also can be used in speaker recognition, so-called text-dependent speaker recognition, to decrease the risk to the replaying attack. But in this paper, we mainly discuss the text-independent speaker recognition because we just want to find the weaknesses of the speaker recognition model. Before the DNNs are used in speaker recognition, most of the speaker recognition models are based on the i-vector feature representation [12] of speech segmentation, which improves the performance with a big margin over the Guassian Mixture Model-Universal Background Models (GMM-UBMs) [36]. Snyder *et al.* [41, 42] proposed to classify or verify the speaker using end-to-end neural networks with the hand-crafted features (e.g. FBANK or MFCC) as input. However, the hand-crafted features, designed based on the perceptual experiments, may not be optimal for the speaker recognition. Muckenhirn *et al.* [28] attempted to train the CNN based deep model from the raw signals, and the cumulative frequency response of the first layer shows that the filters emphasize the low frequency information (<500Hz). Also taking the raw signals as input, Jung *et al.* [20] designed a CNN-LSTM based deep model to extract features for speaker verification. Recently, based on the insight that the most critical part of current waveform-based CNNs is the first convolutional layer, SincNet [34] was proposed to replace the convolution filters of the first layer in DNNs with a group of learnable band pass filters. State-of-the-art performance could be achived on TIMIT [16] and LibriSpeech [30] datasets. In the time

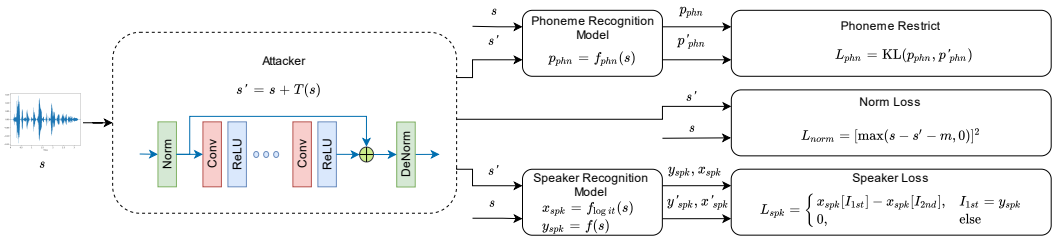

Fig. 2. Illustration of our framework. Our speaker recognition attacker is applied to the raw speech (s) input and generates a speech (s') with perturbations, which can fool the following pretrained speaker recognition model although the perturbations are almost imperceptible. Besides, a pretrained phoneme recognition model is used to help train the attacker network. No finetuning for the pretrained modules is conducted when training the attacker model. (Best view in color.)

domain, given the frequency band $[f_1, f_2]$, the band pass filter can be described as:

$$g[n, f_1, f_2] = 2f_2\mathrm{sinc}(2\pi f_2 n) - 2f_1\mathrm{sinc}(2\pi f_1 n), \tag{2}$$

where $\mathrm{sinc}(x) = \sin x/x$, and f_1, f_2 are the learnable parameters. Compared with the convolution filters, band pass filters reduce the amount of the parameters and give a clear physical meaning. Therefore, in this paper, SincNet is used as the target victim model.

Based on the previous works on the speaker recognition and adversarial attack, we attempt to find the weaknesses and attack the state-of-the-art speaker recognition model. Specifically, a lightweight attacker network is trained to transform the input victim samples into adversarial examples and use these examples to fool the following well-trained speaker recognition model. To optimize our attacker model, a pretrained phoneme recognition model is used to restrict the phoneme information in the adversarial examples. Experimental results show that (1) our model can attack the well-trained speaker recognition model with a high success rate as well as high quality; (2) our proposed optimization method can achieve a good trade-off between the attack success rate and the perceptual quality (PESQ); (3) high frequency attack for speaker recognition can achieve better performance than low frequency attack, which is different with the previous conclusions drawn in image classification adversarial attack [39].

3 SPEAKER RECOGNITION ATTACKER

In this section, we will describe the proposed speaker recognition attacker and its optimization method in detail. The attack problem is formulated as an adversarial optimization. As illustrated in Fig. 2, our proposed speaker recognition attack, a trainable network surrounded by a dotted line, is used to generate additive perturbations to the input speech to attack the following pretrained speaker recognition model. A pretrained phoneme recognition model is used to help train the attacker network by restricting the phoneme information in the adversarial examples. The perturbations are also constrained with norm optimization to make them as small as possible.

Motivation. The basic motivation for why a simple ResNet can well approximate the adversarial perturbation are: (1) there is some acoustic information in a slice of audio that indicates the identity of the speaker, while other information is not related to the identity. So it is feasible to find the acoustic information and ease the identity-related information by learning a model; (2) prior works on image adversarial attack [6] had shown that 2D ResNet

can learn the adversarial perturbation for recognition. Based on these two motivations, we use a ResNet to learn an additive perturbation to attack the well-trained speaker recognition model in our work.

3.1 Formulation

The attacker model aims to transform the input victim sample into an adversarial example with certain perturbations as small as possible because we hope the perturbations are imperceptible for humans. For the non-targeted attack, our attack is successful as long as the prediction is not the same as the label. Given a speech sample s and its speaker label y_{spk}, the non-targeted attack for the speaker recognition model can be formulated as:

$$\arg\min_{s'} L(s, s') \quad \text{s.t.} \quad f(s') = y'_{spk} \tag{3}$$
$$\text{where } s' = T_{f,\theta}(s) \text{ and } y'_{spk} \neq y_{spk},$$

where T is the attack model to transform the input speech s into an adversarial example s', y'_{spk} is the prediction of s', f is a well-trained speaker recognition model, L is a metric function (e.g. the ℓ_2 norm or some other perceptual metric) to measure the distance between two samples. For the targeted attack, our attack is successful as long as the prediction is the target class. Therefore, the targeted attack can be formulated as:

$$\arg\min_{s'} L(s, s') \quad \text{s.t.} \quad f(s') = y'_{spk} \tag{4}$$
$$\text{where } s' = T_{f,\theta}(s) \text{ and } y'_{spk} = y_{target},$$

where y_{target} is the target speaker class.

3.2 Attacker Network

The proposed attacker network is a fully convolution residual network, including 5 convolution blocks plus a global residual connection, as illustrated in Fig. 2. 1-dimensional convolution, batchnorm [19], and ReLU [29] are applied in each convolution block, following the setting in [6]. The kernel size for all convolution layers is set as 3 and the channel is set to be 32. To increase the receptive field, different dilation sizes are used in the different convolution layers, which are $1, 2, 5, 2, 1$, respectively. Besides, we initialize the weights and biases of the last convolution layer as zero so that our model adds no perturbation at the start of the training, which is important for the optimization to keep the perturbations on a small scale. The ablation and comparative study of our network architecture will be given in Section 4.

3.3 Phoneme Restrict

Traditionally, as described in [6], the distance metric (L in Eq. 3) is only ℓ_2 norm, but ℓ_2 norm is mainly consistent with the objective quality, which can be measured by Signal Noise Ratio (SNR). However, the perceptual quality, which can be measured by Perceptual Evaluation of Speech Quality (PESQ), cannot be well modeled by the ℓ_2 norm. In this paper, we propose an insight that one speech consists of two kinds of information: (1) acoustic information that indicates the identity of the speaker, which mainly contributes to the speaker recognition task; (2) phoneme information that indicates the semantic information in the speech, which mainly contributes to the phoneme recognition task. The perturbations we add on the input speech should only modify the acoustic information (part (1)) rather than the phoneme information (part (2)). With a constraint on the change of the phoneme information, we can add perturbations to attack the speaker recognition with minimal distortion, resulting in a better objective and subjective quality. So a pretrained phoneme

recognition is used to impose a *phoneme restrict* between the input sample s and the adversarial example s', as illustrated in Fig. 2. The structure of the phoneme recognition model is based on SincNet [34] and we restrain the distance between the score distributions of the input speech and the adversarial example to minimize the perturbation about the phoneme information. Overall, our distance metric $L(s, s')$ in Eq. 3 contains two items: the norm restrict and the phoneme restrict.

3.4 Optimization

The straightforward method to train the attacker network is gradient ascent, however, in practice, it may fail because a well-trained speaker recognition model propagates back almost zero gradient due to the softmax layer. Motivated by the Wasserstein GAN [4] to optimize the Wasserstein distance between two distributions, we just solve this zero gradient problem by optimizing the adversarial loss directly on the immediate activation before the softmax layer. Moreover, we also need to ensure that the perturbations are imperceptible. Improved ℓ_2 norm with a margin is used to constraint the scale of the perturbations. Besides, we also take the phoneme information into the account via a pretrained phoneme recognition network to optimize the perceptual quality. In summary, we optimize our attacker network from three aspects:

$$L_{total} = L_{spk} + \lambda_{phn} L_{phn} + \lambda_{norm} L_{norm} \tag{5}$$

$$L_{spk} = \begin{cases} x'_{spk}[I_{1st}] - x'_{spk}[I_{2nd}], & I_{1st} = y_{spk} \\ 0, & \text{else} \end{cases} \tag{6}$$

$$L_{phn} = \mathrm{KL}(p_{phn} \| p'_{phn}) \tag{7}$$

$$L_{norm} = [\max(s - s' - m, 0)]^2, \tag{8}$$

where $x' = f_{logit}(s')$ is the intermediate activation before the softmax layer of the speaker recognition model with the input s', p/p' is the output distribution of the softmax layer of the phoneme recognition model with the input s/s', I_{1st}/I_{2nd} is the index of the 1st/2nd largest value in x'_{spk}. In L_{phn}, Kullback–Leibler divergence (KLD) is used to measure the distance between two distributions. In L_{norm}, m is a hyper-parameter to give a margin in which the perturbations are thought to be imperceptible. λ_{phn} and λ_{norm} are used to fuse the three loss items. For the targeted attack, the loss is the same except that L_{spk} is changed to

$$L_{spk,target} = \begin{cases} x'_{spk}[I_{1st}] - x'_{spk}[y_{target}], & I_{1st} \neq y_{target} \\ 0, & \text{else,} \end{cases} \tag{9}$$

where y_{target} denotes the target speaker class.

3.5 Inference

In the training stage, the input speeches with variable length are split into frames with fixed length due to the fully connection layer in the speaker recognition model. However, in the testing stage, the input speeches can be with arbitrary length because our attacker network is a fully convolution residual network. The inference is fast because our attacker network is lightweight with only 5 convolution blocks and small kernel size filters. Besides, the pretrained phoneme recognition model is only used in the training stage.

Table 1. Details for TIMIT and LibriSpeech Dataset.

Dataset	Train Set	Test Set	Speaker Label	Phoneme Label
TIMIT	2310	1386	Yes	Yes
Libri	14481	7452	Yes	No

4 EXPERIMENTAL RESULT

4.1 Experiments Setup

Speaker/Phoneme Recognition Model. We use the state-of-the-art speaker recognition model, SincNet [34], as the target model to be attacked. SincNet replaces the first layer of a CNN with a group of learnable band pass filters. In this way, the network is more interpretable and achieves better performance [33]. Besides, SincNet can be also applied to phoneme recognition[1]. In our experiment, the officially released pretrained SincNet model for speaker recognition is used. The phoneme recognition model is referred from Pytorch-Kaldi [35] and achieves a frame error rate of 26.4% on the TIMIT [16] dataset[2]. Although it is easy to decrease the error rate for phoneme recognition, our aim is to study if the well-trained phoneme recognition model can optimize the perceptual quality of the adversarial example instead of training a good phoneme recognition model. So we did not take much effort to optimize this phoneme recognition model.

Dataset. Following the setting in [34], experiments on TIMIT (462 speakers, train chunk, labels for speakers, and phonemes) [16] and LibriSpeech (2484 speakers, labels for speakers) [30] datasets are conducted to demonstrate the effectiveness of our proposed model. The details for TIMIT and LibriSpeech datasets are shown in Table 1. *Phoneme restrict*, described in Subsection 3.3, can only be evaluated on TIMIT dataset because LibriSpeech dataset has no label for phonemes. The split for the train/test set follows the setting in SincNet [34]. We use the pretrained speaker recognition model released by the authors of TIMIT dataset[3], and we retrain a speaker recognition model for LibriSpeech dataset.

Metric. The signal-to-noise ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) score [37] are used to evaluate the objective and perceptual quality, respectively. SNR is calculated as follows:

$$\mathrm{SNR} = 10\log_{10}[(\sigma_s^2/\sigma_e^2)], \tag{10}$$

where σ_s^2/σ_e^2 is the mean square of the input signal/error. PESQ [37] is an integrated perceptual model to map the distortion to a prediction of subjective mean opinion score (MOS) with range $[-0.5, 4.5]$, which is an ITU-T recommendation technology [1]. For the non-targeted attack, our attack is successful as long as the prediction is not the same as the label, so sentence error rate (SER) is used to measure our attacker's performance. For the targeted attack, our attack is successful as long as the prediction is the target class, so prediction target rate (PTR), which is the percentage of the target class in the prediction over the testing set, is used to measure our attacker's performance.

Trainging Details. The speech sentences, with sampling rate 16k, are split into frames with a 200ms window size and a 10ms stride, following [34]. The data will be normalized

[1]https://github.com/mravanelli/pytorch-kaldi

[2]we use the same train/test split for phoneme and speaker recognition, which is different with the typical split manner for the phoneme recognition

[3]The pretrained model for TIMIT dataset can be found from the GitHub of SincNet: https://github.com/mravanelli/SincNet

Table 2. Non-targeted attack results for TIMIT and LibriSpeech dataset.

Dataset	λ_{phn}	λ_{norm}	SER(%)↑	SNR(dB)↑	PESQ↑
TIMIT	-	-	1.52⋆	-	-
	0	0	99.7	18.56	1.09
	0	1000	96.5	56.39	3.72
	0	2000	86.7	57.79	3.61
	1	1000	**99.2**	57.20	4.20
	5	1000	93.9	58.00	4.25
	10	1000	90.5	**59.01**	**4.28**
Libri	-	-	0.30†	-	-
	0	10	100.0	28.52	2.54
	0	100	99.7	33.72	3.21
	0	200	99.2	35.89	3.43
	0	500	96.9	39.36	3.63
	0	1000	86.2	40.38	3.67
	0	2000	86.4	40.38	3.75

⋆This result is not the same as that in [34], but this model is released by the author of [34].
†This result is our reproduced result, because no model is released for LibriSpeech dataset.

before being feed into the attacker model and de-normalized when they are outputted from the attacker model. The data are clipped into the range $[-1, 1]$ after being added with the perturbation. After finetuning the hyper-parameters, we set $\lambda_{phn} = 1$, $\lambda_{norm} = 1000$ and $m = 0.01$. Adam [21] optimizer with a learning rate of 3×10^{-4} is used to train the attack model for 10 epochs. The data for LibriSpeech are normalized by the author of [34], so we use these data as the raw data and no additional preprocessing is needed. Data, code, and pretrained models have been released on our project main page[4].

Experiments Design. To evaluate our proposed attacker and study the characteristic of the perturbations, the following three groups of experiments are conducted: (1) Non-targeted and targeted attack experiments on both TIMIT and LibriSpeech datasets are conducted to evaluate the effectiveness and efficiency of our proposed attacker. (2) Frequency analysis experiments are conducted by applying different band pass filters on the perturbations, because the frequency characteristic of the speech data is important for human perception. (3) We conduct the ablation studies to give more insights into our proposed attacker.

4.2 Performance Analysis

To evaluate the effectiveness and efficiency of our proposed model, non-targeted and targeted attack experiments on TIMIT and LibriSpeech datasets are conducted.

4.2.1 Non-targeted Attack. The results of non-targeted attack are illustrated in Table 2. The first row for TIMIT/LibriSpeech dataset is the baseline performance of the victim model without attack. For the TIMIT dataset, we use the pretrained model provided by the author of [34], while for the LibriSpeech dataset we reproduce the SincNet.

[4]https://smallflyingpig.github.io/speaker-recognition-attacker/main

Table 3. Targeted attack results for TIMIT/LibriSpeech dataset.

Dataset	Target ID	PTR*(%)↑	SNR(dB)↑	PESQ↑
TIMIT	0	91.4	57.55	3.36
	100	89.3	56.83	3.16
	200	63.3	58.42	3.69
	300	58.7	56.92	3.52
	400	57.6	58.36	3.68
	avg	72.1	57.64	3.48
Libri	0	19.9	39.98	3.50
	500	11.9	39.91	3.35
	1000	49.0	37.57	3.14
	1500	23.1	39.34	3.48
	2000	58.3	38.83	3.30
	avg	32.4	39.13	3.35

*PTR denotes the prediction target rate.

An SER of 0.30% is achieved on the testing set of LibriSpeech dataset (the SER reported in [34] is 0.96%). The rest of the table shows the attack results of our model with different trade-offs between SER with SNR and PESQ by tuning λ_{phn} and λ_{norm}.

Some conclusions can be drawn from the results:

- *Our proposed model successfully attacks the well-trained state-of-the-art speaker recognition model on both TIMIT and LibriSpeech datasets.* On the TIMIT dataset, we achieve an SER of 99.2% on the testing set. Meanwhile, the perturbations are small enough to be imperceptible [5] for humans because the SNR is up to 57.2dB and PESQ is no less than 4.2, indicating the effectiveness of our model. On the LibriSpeech dataset, we achieve an SER of 96.9% on the testing set with an SNR of 39.36dB and a PESQ of 3.63. Without the phoneme restrict on the LibriSpeech dataset, the perturbations may be not imperceptible because the PESQ is only 3.63, however, the perturbations are still small enough with an SNR of 39.36dB.
- *Tuning λ_{phn} works better than tuning λ_{norm} on getting a trade-off between SER and SNR/PESQ, demonstrating the effectiveness of our proposed phoneme restrict method.* On TIMIT dataset, tuning λ_{norm} fails to get a trade-off between SER and PESQ (row 1st, 2nd, 3rd in Table 2). On the LibriSpeech dataset, tuning λ_{norm} can improve the PESQ with the expense of attacking success rate, but the profit is not encouraging although the SER has dropped with a bit margin ($\lambda_{norm} = 200, 500, 1000, 2000$ for LibriSpeech in Table 2.) By comparison, tuning λ_{phn} can improve PESQ with a slight SER drop (2nd, 4th∼6th in Table 2), demonstrating our optimization is effective to achieve a trade-off between the attack success rate and the perceptual quality.
- The objective quality (measured by SNR) changes consistent with the perceptual quality (measured by PESQ), but not proportionately, so they are two different metrics to measure the quality of the audio in different aspects. In our scenarios, human perception is the main concern, so only optimizing the objective metric, such as SNR, is not enough. Better optimization method which can improve the perceptual quality, like our *phoneme restrict*, is necessary to train a good attacker model.

[5]Empirically, the perturbations are imperceptible when PESQ>4.0.

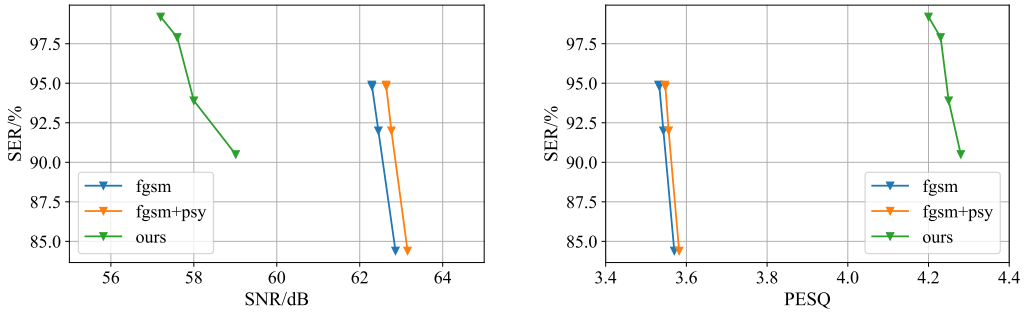

Fig. 3. Comparison results (non-targeted attack) with baseline methods on TIMIT dataset (left: SER-SNR curves, right: SER-PESQ curves, higher curve means better model performance when taking both the attack success rate and signal distortion into account).

4.2.2 Targeted Attack. The results of targeted attack is illustrated in Table 3. The first five rows are the results of the targeted attack for five randomly selected speakers from the TIMIT/LibriSpeech dataset and the last row shows the average performance. Some observations can be obtained from the results:

- *Our proposed attacker model is effective for the targeted attack.* On the TIMIT dataset, our model can attack the speaker recognition model with a PTR of 72.1% on average over the five targets. Meanwhile, the perturbations are small enough because the SNR is up to 57.64dB. On the LibriSpeech dataset, An average PTR of 32.4% with the SNR of 39.13dB and the PESQ of 3.35 can be achieved. Although the PTR of LibriSpeech is not good enough, the quality, especially the perceptual quality, is comparable with that on the TIMIT dataset.
- The PESQ on the TIMIT dataset is not as good as that of the non-targeted attack, indicating that targeted attack is rather more challenging than the non-targeted attack.
- The attack success rate on the LibriSpeech dataset is lower than that on the TIMIT dataset, showing that targeted attack on LibriSpeech is more difficult than that on TIMIT. Because the total speaker number in LibriSpeech dataset (2484) is much more than that in TIMIT dataset (462). So both the objective and subjective quality on the LibriSpeech dataset are rather worse than that on the TIMIT dataset.

4.2.3 Comparison with Baselines. We conducted comparison experiments with two baseline methods: (1) *FGSM-based method [17, 18];* (2) *Psychoacoustic hiding method [32, 38].* Without loss of generality, given the audio sample x and its speaker recognition y, the two baseline methods can be formulated as:

$$x' = x - \text{sign}(\Delta_x f) \times \alpha \tag{11}$$

$$x' = x' - \Delta_{x'} g \times \beta, \tag{12}$$

where x' is the adversarial sample, sign is the sign function, f is the loss function for attacking speaker recognition (such as $f = 1 - \text{CE}(\hat{y}, y)$, CE denotes the cross entropy), g is the psychoacoustic model, α and β are the learning rate for FGSM and psychoacoustic model.

Table 4. Different metrics about computation complexity on our model.

MACs	FLOPs	#params (M)	Model Size (Mb)	RTF(GPU)	RTF(CPU)
0.69×10^6	0.35×10^6	0.064	0.062	0.001*	0.042†

*It is tested on a Nvidia GTX1080TI GPU.
†It is measured on an Inter(R) Core(TM) i7-6700K@3.4GHz CPU.

In our implementation, we set $\alpha = 0.005, \beta = 0/\alpha = 0.005, \beta = 0.001$ for FGSM/FGSM+psy model, respectively, after finetuning the hyperparameters to get strong beselines [6].

Comparison of adversarial attackers is not intuitive because both the attack success rate and the signal distortion should be taken into account. Here we compare different methods by plotting the R-D curve (attack success **R**ate/signal **D**istortion). By adjusting the perturbation level and giving different distortion thresholds, we can obtain different trade-offs between the attack performance and the signal distortion. The untargeted attack results of different attack methods on TIMIT dataset are shown in Fig. 3. We can see from the results: (1) Using SNR as the distortion metric, FGSM based method is a high-performance adversarial attack although its iterative optimization is less efficient. Our method can be seen as a model to fit the FGSM's results with only "one pass" forward process; (2) Our model can achieve much better PESQ than the baseline methods, demonstrating the effectiveness of the proposed optimization method; (3) compared with FGSM based method, the psychoacoustic model can achieve better perceptual fidelity (measured by PESQ) and signal fidelity (measured by SNR) without the loss of attack success rate, indicating its effectiveness for the adversarial attack. Besides, our method is much faster than the baseline methods (0.48s for a sample *vs* 2.52s for a sample, averagely). Summarily, our method can achieve better perceptual quality (measured by PESQ) when achieving a similar attack success rate than the baseline methods, while the computation complexity is much less than the baselines.

4.2.4 Computation Complexity. Our proposed attacker model is fast for testing because (1) it is lightweight with only 5 convolution blocks; (2) it needs neither gradient nor iteration in the testing stage. To verify whether it is fast enough to process the speech data in real-time, we calculate the real-time factor (RTF) over the testing set. RTF is defined as:

$$\text{RTF} = \frac{P}{I}, \tag{13}$$

if it takes time P to process an input speech with duration I. The system is real-time if RTF ≤ 1. Apart from RTF, we also calculate the FLOPs (floating-point operations), the MACs (multiply-accumulate operations), the number of parameters, the saved model size to show that our model is fast enough for real applications. As shown in Table 4, the averaged FLOPs of our model for a test sample is only 0.35×10^6, indicating that this is a lightweight model. The number of parameters is also small, which is only 0.064 Mb. The RTF shows that our model is much faster than real-time because it is more than 20 times faster than the real-time requirement even if it runs in the GPU mode.

4.3 Frequency Domain Analysis

Recently, frequency analysis for the adversarial attack has drawn increase attention since Sharma *et al.* [40] won the first place on both non-targeted and targeted attack in CAAD

[6]We referenced the implementation of these two baseline methods from https://github.com/Trusted-AI/adversarial-robustness-toolbox

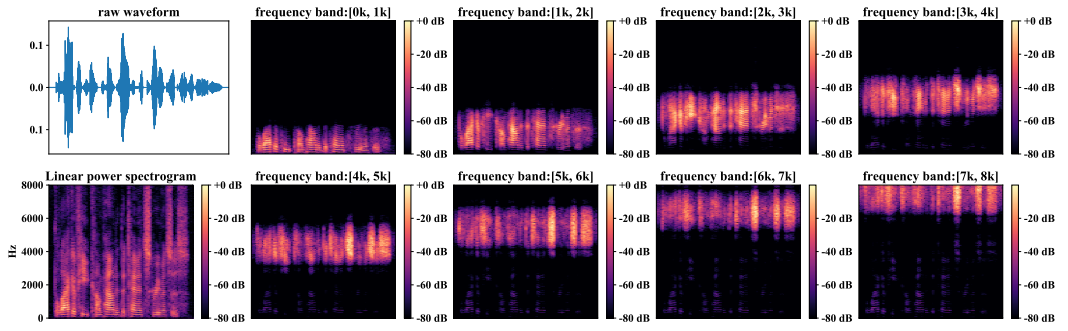

Fig. 4. The results of 8 band pass filters. (row 1, col 1): the raw waveform of a sample in TIMIT dataset; (row 2, col 1): the power spectrogram with a linear scale for the raw sample. (row 1-2, col 2-5): the power spectrogram with a linear scale for the sample data after a specific band pass filter. (Best view in color.)

2018 competition by applying a 2D Gaussian filter on the gradients to smooth the adversarial perturbations. The following work on frequency analysis [39] demonstrated that constraining the perturbations in low frequency can improve the attack performance. However, this conclusion was drawn from the experiments about image classification attack. It may not generalize well to speaker recognition attack.

In this subsection, we would like to explore the following questions: (1) What is the frequency distribution of the perturbations, full-band or focusing on the specific frequency band? (2) Low frequency attack or high frequency attack, which can achieve better attack performance for speaker recognition? (3) What is the influence of the *phoneme restrict* for the perturbations?

4.3.1 Perturbation Frequency Distribution. The perturbation frequency distribution is useful for us to find the weaknesses of the speaker recognition model, and we can attack the well-trained model with perturbations in a restricted frequency band. To get the frequency distribution over the testing set on TIMIT dataset with phoneme restrict, firstly a short-time Fourier transform (STFT) is conducted to transform the perturbation data into a spectrogram with windows length 2048 and hop length 256, resulting in a 1025 dimensions spectrogram. Then we concatenate all the spectrograms along the time dimension and sum the energy along the time dimension to get a 1-dimension feature which represents the additive energy over the whole testing set. Finally, the feature is normalized into a density distribution to get the frequency energy distribution.

As illustrated in Fig. 5, the frequency distribution for the perturbations with *phoneme restrict* on TIMIT dataset is full-band, however, the energies are not uniform and the energy distribution in the high-frequency band is rather more than that in low-frequency band, indicating that the frequency bands do not contribute equally.

A similar conclusion that speaker recognition models do not respond uniformly to the input speech along the frequency has been drawn in [28]. However, they showed that the filters in the well-trained speaker recognition model emphasize the information lying in the low frequency band (below 1k Hz). The results we get here show that, in addition to the low frequency band (below 1k Hz), information lying in the high frequency (above 5k Hz) also

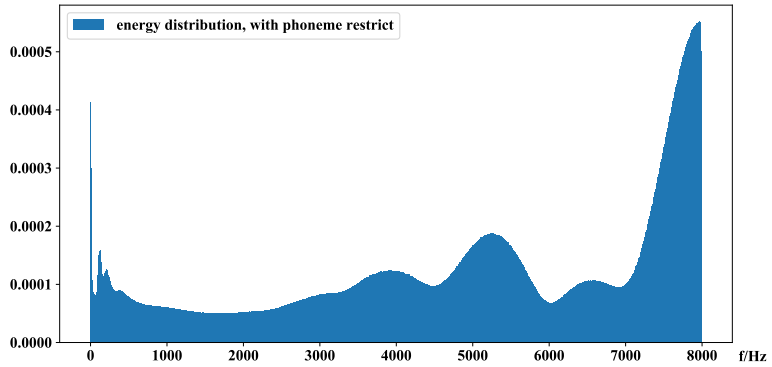

Fig. 5. The frequency distribution over the testing set on the TIMIT dataset. We can see from this figure that the perturbations are full-band, however, the energy is not uniform over the full frequency band.

play an important role in the speaker recognition task, and even more important than that lying in the low frequency band.

4.3.2 Low Frequency Attack or High Frequency Attack Performs Better. To investigate this problem, we need to firstly constrain the frequency of the adversarial perturbations. For frequency domain analysis on image classification, Sharma *et al.* [39] applied a discrete cosine transform (DCT) based mask on the perturbation and masked the perturbations in DCT domain to only preserve the specific frequency band. Different from this DCT based method, which needs to tranform the perturbation into frequency domain, we design a group of band pass filters to process the speech perturbations in time domain, without any transformation or inverse transformation operation. Specifically, 8 band pass filters are designed to restrict the perturbations in the specific frequency bands. We fix the frequency bandwidth as 1k Hz and use a discrete digital filter with 33 points (kernel size), because in our experiments smaller kernel size cannot decay the energy enough at the frequency beyond the specific band. The band pass filters can be formulated as follows:

$$\begin{aligned} g[n, f] &= 2(f + 1000)\text{sinc}(2\pi f n + 2000\pi n) - 2f\text{sinc}(2\pi f n), \\ n &= -16, -15, -14, \ldots, 14, 15, 16, \\ f &= 0, 1000, 2000, \ldots, 7000. \end{aligned} \tag{14}$$

The results after applying the band pass filters are shown in Fig. 4. As illustrated in Fig. 4, the energy beyond the frequency pass band decay by about 80dB, indicating that our band pass filters are effective to restrict the frequency range of the data.

By applying the aforementioned band pass filters to the perturbations, we can get the attack success rate/SNR/PESQ curve along with the frequency. The results are shown in Fig. 6. To show the attack performance changing along with the frequency, we test our model with aligned SNR [7] by scaling the perturbations and plot the SER-Frequency line chart (column 1 in Fig. 6). To take both the attack performance and adversarial example quality into account, we test our model without scaling the perturbations and plot the

[7]we align the SNR to 60dB/42dB for TIMIT/LibriSpeech dataset.

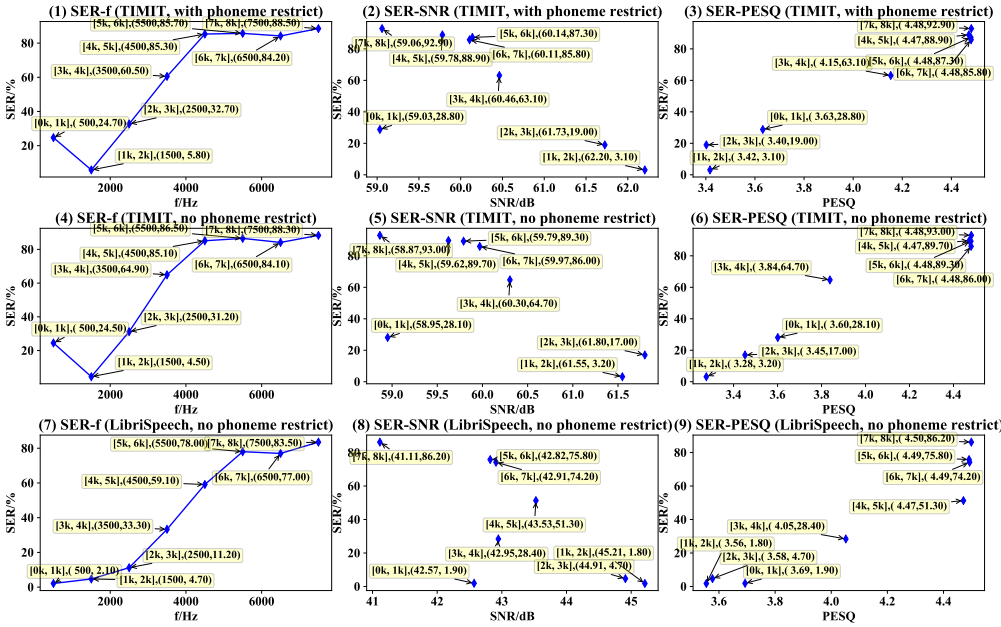

Fig. 6. Non-targeted attack results with different frequency band pass filters applied to the perturbations on TIMIT and LibriSpeech datasets. (Best view in color.)

SER-SNR (column 2 in Fig. 6) and SER-PESQ (column 3 in Fig. 6) scatter diagrams. Some observation can be seen from the figure:

- *Different frequency bands affect the performance of the attacker differently.* On TIMIT dataset, as illustrated in subfigure (1) in Fig. 6, the worst performance is only with an SER of 5.8% given the band pass filter [1k, 2k], while the best performance is with an SER of 88.5% given the band pass filter [7k, 8k]. The SER-Frequency curve is similar to the frequency distribution in Fig. 5, indicating that the energy distribution over the frequency can reflect the contribution to the speaker recognition model of different frequency bands to a certain degree. Similar conclusion can be drawn from the subfigure (4) in Fig. 6, although no *phoneme restrict* is applied in the experiment. On LibriSpeech dataset, the frequency bands affect the attacking performance a lot, too, as shown in subfigure (7) in Fig. 6.
- *Attacking on the frequency bands lower than 3k Hz cannot achieve good attack performance, even hurts the perceptual quality of the adversarial example significantly, although it may give good objective quality.* As shown in subfigure (2), (3), (5), (6), (8), (9) in Fig. 6, attacking with frequency bands lower than 3k can only obtain an SER no more than 29%/5% for TIMIT/LibriSpeech dataset. Moreover, the perceptual quality of these frequency bands (a PESQ no more than 3.7 for both TIMIT and LibriSpeech datasets) are also the three worst ones among the total 8 frequency bands, which indicates that we should avoid using these frequency band if it is possible when attacking the speaker recognition model.
- *High frequency bands (higher than 4k Hz) are able to allow our attacker to fool the speaker recognition model with a high success rate and good perceptual quality, although*

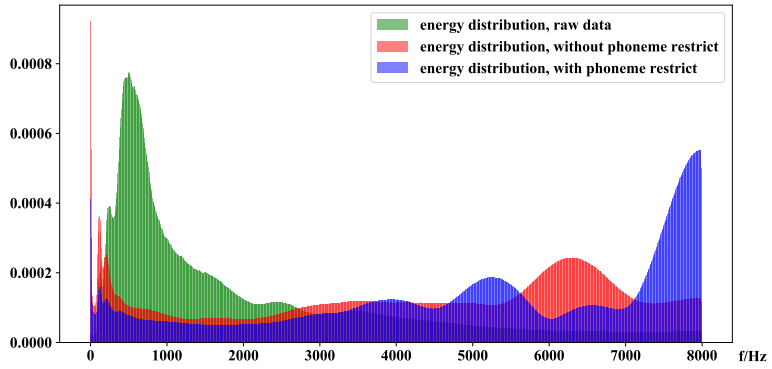

Fig. 7. Energy distribution over the frequency of the raw data, the perturbations with phoneme restrict, and the perturbations without phoneme restrict. (Best view in color.)

the objective quality is not always best. As shown in subfigure (3), (6), (9) in Fig. 6, the frequency bands higher than 4k Hz ([4k, 5k], [5k, 6k], [6k, 7k]) can achieve an SER no less than 85%/51% for TIMIT/LibriSpeech dataset and a PESQ no less than 4.47 for both TIMIT and LibriSpeech datasets, and both the metrics are better than all of the rest frequency bands. Although the objective quality is not always good enough, as shown in the subfigures (2), (5), (8) in Fig. 6, the high frequency bands that higher than 4k Hz should also be highlighted, given that fact that the perceptual quality is the main concern in the attacking.

- *With band pass filters on perturbations, the phoneme restrict does not increase the attacker's performance obviously, nor the perceptual quality, although it achieves significant gain when no band pass filter on the adversarial perturbation.* As shown in the subfigures (1) and (4) in Fig. 6, the performance in the two experiments with or without phoneme restrict are very similar on all the 8 frequency bands. Meanwhile, in the subfigures (3) and (6) in Fig. 6, no obvious gain for PESQ on all the frequency bands except the band [3k, 4k]. So the phoneme restrict will fail if we attack the speaker recognition in a frequency band with a bandwidth about 1k.

Overall, we demonstrate that for speaker recognition adversarial attack, high frequency attack can achieve better performance (both higher attack success rate and better data quality) than low frequency attack, which is different from the pervious conclusion for image classification adversarial attack [39].

4.3.3 Attack with(out) Phoneme Restrict. In the previous two subsections, we have found that: (1) the phoneme restrict can improve the perceptual quality of the adversarial examples when attacking with full-band perturbations; (2) the phoneme restrict does not work when we attack with a certain frequency band. These two conclusions seem contradictory. In this subsubsection, we aim to investigate *how the phoneme restrict affects the perturbations and why it works or fails.*

The frequency distribution of the perturbations, which are the output of our attacker model, is mainly determined by the following three factors: (1) the training/testing data frequency distribution, which are the input of our attacker model; (2) the victim model, which is a well-trained, deep-based speaker recognition model; (3) the optimization method,

which contains three items in our method: L_{spk}, L_{phn} and L_{norm}. To study how the *phoneme restrict* affects the perturbations, we plot the frequency energy distribution of the testing data, the perturbations with phoneme restrict, and the perturbations without phoneme restrict. As illustrated in Fig. 7, both the frequency distributions with and without phoneme restrict are not consistent with the testing data distribution. Some observations can be seen from the results:

- The energy of testing data mainly concentrates on the low frequency that is lower than 3k Hz. This is why we usually analyze the spectrogram of the speech data in the log scale, which highlights the information in low frequency.
- Compared with the testing data distribution, the distribution of the perturbation without phoneme restrict, which bounds the perturbations with only the ℓ_2 norm, focuses more on the high frequency. The reasons may lie: (1) the victim speaker recognition model concentrates more on the high frequency, and to attack the victim model, the perturbations focus more on the high frequency. (2) our attacker model generates perturbations on high frequency more easily than that on low frequency. (3) the task, speaker recognition, is related more with the information in high frequency than that in low frequency.
- Compared with the energy distribution of the perturbations with and without the *phoneme restrict*, we find that by adding phoneme restrict, the perturbations focus much more on the highest frequency band ([7k, 8k]), indicating that the *phoneme restrict* leads the perturbations into the highest frequency band. Given the conclusion drawn above that high frequency attack will achieve better performance and that low frequency attack will hurt the perceptual quality, by leading the perturbations to the high frequency, the phoneme restrict will improve the attacker's performance as well as the perceptual quality.
- The phoneme restrict fails when training with a specific frequency band. This is because the perturbations cannot focus on the high frequency due to the band pass filter, so no obvious gain will be got by adding the phoneme restrict when the perturbations are cut in a certain frequency band.

4.3.4 Looking Deeper into the Energy Distribution. In the above analysis, we have concluded that: high frequency attack can achieve better performance than low frequency attack. However, the energy in high frequency is rather smaller than that in low frequency for the real speech data. One possible reason for why this happens is that *the victim model focuses much more on the high frequency than that on the low frequency*, so high frequency attack will result in a higher success rate than that on low frequency. This may lead to the weakness of the convolution-based models: *the models concentrate on the information on the high frequency but does not make full use of the information on the low frequency.* So better performance is possible if we can take full advantage of the information in low frequency.

4.4 Ablation Study

The ablation study for our proposed attack model is conducted to investigate the influence of the structure, such as the dilation, the kernel size, and the layers. As illustrated in Table 5, some useful conclusions can be drawn from the results:

- The dilation size can affect the performance of our model. Larger dilation can improve the attacking success rate (model Dilation2, Dilation123 in Table 5), but only enlarging all the convolution with the same factor will hurt the perceptual quality of the

Table 5. Ablation study for our proposed attacker on TIMIT dataset.

Model	Dilation	Kernel size	Channel	SER(%)↑	SNR(dB)↑	PESQ↑
Baseline	[1, 1, 1, 1, 1]	[3]×5	[32]×5	93.8	57.96	4.12
Dilation2	[2, 2, 2, 2, 2]	[3]×5	[32]×5	95.3	57.77	3.55
Dilation123	[1, 2, 3, 2, 1]	[3]×5	[32]×5	99.0	56.78	4.24
KS5	[1, 1, 1, 1, 1]	[5]×5	[32]×5	98.6	57.23	4.25
Block4	[1, 1, 1, 1]	[3]×4	[32]×4	93.4	58.42	4.13
Block6	[1, 1, 1, 1, 1, 1]	[3]×6	[32]×6	96.0	57.45	4.15
Block8	[1, 1, 1, 1, 1, 1, 1, 1]	[3]×8	[32]×8	93.9	56.95	4.09

Table 6. Non-targeted attack results on TIMIT dataset for different speaker recognition models.

λ_{phn}	λ_{norm}	SincNet Speaker Model			CNN Speaker Model		
		SER(%)↑	SNR(dB)↑	PESQ↑	SER(%)↑	SNR(dB)↑	PESQ↑
-	-	1.52⋆	-	-	2.16⋆	-	-
1	1000	99.2	57.20	4.20	99.5	57.91	4.22
5	1000	93.9	58.00	4.25	94.1	58.52	4.24
10	1000	90.5	59.01	4.28	92.0	59.97	4.27

⋆ Sentence error rate of speaker recognition without attack.

adversarial examples greatly (model Dilation2). As a comparison, enlarging the dilation with crafted size will improve the SER as well as the PESQ (model Dilation123).

- Enlarging the kernel size of the convolution will improve the attack success rate (model KS5). The reason may be larger kernel size results in a larger receptive field and bigger model capacity. However, a larger kernel size means more parameters and higher computation complexity.
- Deeper networks do not always result in better performance. Compared with the baseline model (model Baseline with 5 blocks), 6 blocks (model Block6) improve the SNR to 96.0%, while 8 blocks (model Block8) decrease back to 93.9%.

Notably, our aim of the ablation study is not to find the optimal hyperparameter but to show different trade-offs between the attack success rate and the signal distortion by tuning the hyperparameters.

4.5 Generalization Capability

In this section, we explore the generalization capability of our proposed method to different speaker recognition models. So far we only demonstrate the effectiveness of our proposed method on the SincNet model. So we design a CNN model by replacing the first layer of the SincNet with a convolutional layer to keep minimal architecture change. As shown in Table 6, our proposed method can achieve similar attack performance on the CNN speaker recognition model when comparing with the SincNet speaker recognition model, indicating the generalization capability of the proposed attack model and optimization method.

5 CONCLUSION AND FUTURE WORKS

In this paper, we have attempted to attack the well-trained state-of-the-art speaker recognition model and proposed an attacker model as well as its optimization method to find

the weaknesses of the well-trained model and analyze the characteristic of the adversarial examples. Both the non-targeted and targeted attack experimental results on TIMIT and LibriSpeech datasets demonstrated the effectiveness and the efficiency of our proposed model. Meanwhile, the comparison experiments on the TIMIT dataset, which has the phoneme labels, showed that our proposed optimization method, which leverages a pretrained phoneme recognition model to restrict the phoneme information of the adversarial examples, can improve the perceptual quality of the adversarial examples. To study the frequency characteristic of the perturbations, the frequency analysis is conducted by applying a crafted band pass filter on the perturbations generated by our proposed attacker model. The experimental results indicated that high frequency attack for speaker recognition performs better than low frequency attack, achieving a higher attack success rate and a better data quality, which is different from the conclusion drawn in image classification adversarial attack. Besides, the ablation study on our proposed model showed that a well-crafted dilation size will improve the attacking success rate as well as the perceptual quality. Our work provides a new benchmark for speaker recognition adversarial attacks.

In the future, we will continue studying the characteristic of the perturbations and attempting to reveal the influence of different frequency bands to the speaker recognition model and exploring the methods to defense the high frequency attack. At present, our proposed model is a kind of white box attacks, and in the next step, we will attempt to attack the well-trained state-of-the-art speaker recognition model in a black box or semi-black box setting, and study the difference between white box and black box attacks.

6 ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation of China under Grants 62025101620 and 61961130392, PKU-Baidu Fund 2019BD003 and High-performance Computing Platform of Peking University.

REFERENCES

[1] 2001. ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. (2001).

[2] Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* 6 (2018), 14410–14430.

[3] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. 2018. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554* (2018).

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 214–223.

[5] Tao Bai, Jinqi Luo, Jun Zhao, and Bihan Wen. 2021. Recent Advances in Adversarial Training for Adversarial Robustness. *arXiv preprint arXiv:2102.01356* (2021).

[6] Shumeet Baluja and Ian Fischer. 2018. Learning to Attack: Adversarial Transformation Networks.. In *AAAI*. 2687–2695.

[7] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*. 1467–1474.

[8] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden voice commands. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 513–530.

[9] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57.

[10] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069* (2018).

[11] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. 2004. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 99–108.
[12] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4 (2010), 788–798.
[13] Wenrui Diao, Xiangyu Liu, Zhe Zhou, and Kehuan Zhang. 2014. Your voice assistant is mine: How to abuse speakers to steal information and control your phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*. ACM, 63–74.
[14] Figen Ertaş. 2000. Fundamentals of Speaker Recognition. In *Journal of Engineering Sciences*. Number 2–3. 185–193.
[15] Roger Fletcher. 2013. *Practical methods of optimization*. John Wiley & Sons.
[16] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n* 93 (1993).
[17] Yuan Gong and Christian Poellabauer. 2017. Crafting adversarial examples for speech paralinguistics applications. *arXiv preprint arXiv:1711.03280* (2017).
[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).
[19] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
[20] Jee-Weon Jung, Hee-Soo Heo, Il-Ho Yang, Hye-Jin Shim, and Ha-Jin Yu. 2018. A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5349–5353.
[21] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).
[22] Pavel Korshunov, Sébastien Marcel, Hannah Muckenhirn, André R Gonçalves, AG Souza Mello, RP Velloso Violato, Flávio O Simoes, M Uliani Neto, Marcus de Assis Angeloni, José Augusto Stuchi, et al. 2016. Overview of BTAS 2016 speaker anti-spoofing competition. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 1–6.
[23] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. 2018. Fooling end-to-end speaker verification with adversarial examples. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1962–1966.
[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).
[25] J. Li, X. Zhang, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao. 2020. Learning to Fool the Speaker Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2937–2941.
[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
[27] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1765–1773.
[28] Hannah Muckenhirn, Mathew Magimai Doss, and Sébastien Marcell. 2018. Towards directly modeling raw speech signal for speaker verification using CNNs. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4884–4888.
[29] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.
[30] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5206–5210.

[31] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 372–387.

[32] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*. PMLR, 5231–5240.

[33] Mirco Ravanelli and Yoshua Bengio. 2018. Interpretable convolutional filters with SincNet. (2018).

[34] Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 1021–1028.

[35] Mirco Ravanelli, Titouan Parcollet, and Yoshua Bengio. 2019. The pytorch-kaldi speech recognition toolkit. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6465–6469.

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

[37] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, Vol. 2. IEEE, 749–752.

[38] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. 2018. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665* (2018).

[39] Yash Sharma, Gavin Weiguang Ding, and Marcus A Brubaker. 2019. On the effectiveness of low frequency perturbations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 3389–3396.

[40] Yash Sharma, Tien-Dung Le, and Moustafa Alzantot. 2018. Caad 2018: Generating transferable adversarial examples. *arXiv preprint arXiv:1810.01268* (2018).

[41] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5329–5333.

[42] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur. 2016. Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 165–170.

[43] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 23, 5 (2019), 828–841.

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[45] Joseph Szurley and J Zico Kolter. 2019. Perceptual Based Adversarial Audio Attacks. *arXiv preprint arXiv:1906.06355* (2019).

[46] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. 2020. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *International Journal of Automation and Computing* 17, 2 (2020), 151–178.

[47] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, XiaoFeng Wang, and Carl A Gunter. 2018. Commandersong: A systematic approach for practical adversarial voice recognition. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 49–64.

[48] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 103–117.