Interpretable Spectrum Transformation Attacks to Speaker Recognition Systems

Jiadi Yao, Hong Luo, Jun Qi, and Xiao-Lei Zhang, *Senior Member, IEEE*

***Abstract*—The success of adversarial attacks on speaker recognition is mainly in white-box scenarios. When applying the adversarial voices that are generated by attacking white-box surrogate models to black-box victim models, i.e. *transfer-based* black-box attacks, the transferability of the adversarial voices is not only far from satisfactory, but also lacks interpretable basis. To address these issues, in this article, we propose a general framework, named spectral transformation attack based on modified discrete cosine transform (STA-MDCT), to improve the transferability of the adversarial voices to a black-box victim model. Specifically, we first apply MDCT to the input voice. Then, we slightly modify the energy of different frequency bands for capturing the salient regions of the adversarial noise in the time-frequency domain that are critical to a successful attack. Unlike existing approaches that operate voices in the time domain, the proposed framework operates voices in the time-frequency domain, which improves the interpretability, transferability, and imperceptibility of the attack. Moreover, it can be implemented with any gradient-based attackers. To utilize the advantage of model ensembling, we not only implement STA-MDCT with a single white-box surrogate model but also with an ensemble of surrogate models. Finally, we visualize the saliency maps of adversarial voices by the class activation maps (CAM), which offer an interpretable basis for transfer-based attacks in speaker recognition for the first time. Extensive comparison results with six representative attackers show that the CAM visualization clearly explains the effectiveness of STA-MDCT and the weaknesses of the comparison methods; the proposed method outperforms the comparison methods by a large margin. Our audio samples are available on the demo website.[1]**

***Index Terms*—Speaker recognition, adversarial examples, adversarial transferability, black-box attacks.**

Manuscript received 2 March 2023; revised 30 October 2023, 14 December 2023, and 12 January 2024; accepted 3 February 2024. Date of publication 8 February 2024; date of current version 23 February 2024. This work was supported in part by the National Science Foundation of China (NSFC) under Grant 62176211 and in part by the Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality, China, under Grant JCYJ20210324143006016 and Grant JSGG20210802152546026. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Omid Sadjadi. *(Corresponding author: Xiao-Lei Zhang.)*

Jiadi Yao and Xiao-Lei Zhang are with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Research & Development Institute of Northwestern Polytechnical University, Shenzhen 710072, China (e-mail: yaojiadi@mail.nwpu.edu.cn; xiaolei.zhang@nwpu.edu.cn).

Hong Luo is with China Mobile (Hangzhou) Information Technology Company Ltd., Hangzhou, China (e-mail: luohong@cmhi.chinamobile.com).

Jun Qi is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China (e-mail: jun-qi@comp.hkbu.edu.hk).

https://sea-yjd.github.io/

Digital Object Identifier 10.1109/TASLP.2024.3364100

I. Introduction

SPEAKER recognition is the task of identifying a person from voices that contain voice characteristics of the speaker [1], [2], [3], and finds its wide applications in real-world scenarios, such as bank trading, remote payment, and criminal investigations. State-of-the-art speaker recognition systems extract speaker embeddings with fixed dimensions to represent the acoustic characteristics of speakers [3], [4]. Prototypical speaker embeddings are i-vectors [5] extracted from Gaussian-mixture-model-based universal background models [1], [2]. In recent years, with the fast development of deep learning, deep speaker embeddings become a new trend. Representative deep embeddings include d-vectors [6], x-vector [3], etc.

Due to the wide applications of speaker recognition, its security is raising widespread concerns. Many attack techniques were developed to make speaker recognition systems easily fail, which can be categorized into three types: spoofing attacks, backdoor attacks, and adversarial attacks. There are main four sub-types of spoofing attacks: replay, voice conversion, impersonation, and text-to-speech synthesis [7], [8]. Notably, the community-driven benchmark *ASVspoof* Challenge series [9] aim to address voice spoofing attacks. Recently, various detections [10], [11] and spoofing countermeasures [12] for spoofing attacks were developing rapidly. Backdoor attacks [13], [14], [15] provide poisoned data to the training data of a victim model in the training stage, and then activate the attack by presenting a specific small trigger pattern to the victim model in the test stage. Adversarial attacks aim to lead speaker recognition systems to wrong decisions by contaminating benign test examples with perceptually indistinguishable structured perturbations. The contaminated examples, a.k.a. adversarial examples, have been proven to significantly undermine deep-learning-based speaker recognition systems.

According to how much information a victim model can be accessed by an attacker, we categorize existing adversarial attacks into two settings, which are white-box attacks and black-box attacks. In the *white-box* setting, the attacker could access all information of the victim model, including the model architecture, parameters, and training data. For example, Villalba et al. [16] demonstrated that gradient-based white-box attacks achieve a high success rate to speaker verification systems. *Black-box* attacks, which assume that the attackers know little about the victim models, are more practical and challenging than their white-box counterparts. Black-box attacks can be

further partitioned into three sub-categories, which are the score-based, decision-based, and transfer-based attacks respectively. For the score-based attacks, the feedback from victim models is continuous, such as the posterior probability. For the decision-based attacks, the feedback is discrete, such as the recognition results. In both cases, a core problem is how to deal with the unknown gradients in the victim models. Existing solutions include gradient estimation and natural evolution. For example, Zhang et al. [17] proposed an adversarial example generation method named VMask, which estimates the gradients according to the difference between the similarity scores of multiple queries, and then uses zeroth-order optimization [18] to solve the gradient-agnostic problem. Chen et al. [19] proposed the FakeBob attack to estimate the gradients through a natural evolutionary strategy [20]. However, the above score-based and decision-based black-box attacks usually require a large number of queries.

To prevent a large number of queries to victim models, *transfer-based* black-box attacks were developed. Transfer-based black-box attacks [16], [19], [21], [22], [23] first generate adversarial examples from *white-box surrogate models*, and then transfer the adversarial examples to *black-box victim models*. Transfer-based methods do not utilize any information about the black-box victim models. Their success relies strongly on the transferability of the adversarial examples, which is the generalization ability of whether an adversarial example generated against a specific model can deceive other models. This article focuses on discussing transfer-based black-box attacks.

Although transfer-based black-box attacks for speaker recognition have received positive effects [16], [24], [25], [26], the following core problems still need further investigation. (i) For a transfer-based attacker, transferability is a desired property of adversarial examples. However, it seems still far from explored. (ii) Many works generate adversarial examples in the time domain [16], which may neglect the difference between the frequency bands of speech signals, while minor changes in frequency components may result in opposite decisions which is a well-known phenomenon. (iii) The success of an attacker cannot be interpreted straightforwardly, e.g. from the time-frequency spectrogram of a speech segment, which makes the design of an attack algorithm mysterious and heuristic.

To address the above issues, inspired by [23], in this article, we propose to improve the transferability of adversarial examples in the time-frequency domain by a novel framework, named *spectrum transformation attack based on modified discrete cosine transform* (STA-MDCT). STA-MDCT first performs MDCT on the input voice and then slightly modifies the energy of different frequency bands to alter the salient regions in the time-frequency domain that may lead the black-box victim model to an error decision. To make the attack effect interpretable, we propose to visualize the saliency maps of adversarial examples via the class activation maps (CAM) [27]. By comparing the saliency maps of a voice before and after being added with an adversarial perturbation, we find that the adversarial examples generated with STA-MDCT are capable of shifting the attention of a black-box victim model, while its counterparts fail to do so, which provides an interpretable basis to transfer-based attacks. To summarize, our main contributions are as follows:

- We propose the STA-MDCT framework. It is a general framework that any gradient-based attacker can be applied with for probably improving its transferability to a black-box victim model. We implement two STA-MDCT variants, one with a single white-box surrogate model, and the other with an ensemble of white-box surrogate models.
- We propose to interpret the attack effect visually by CAM. To our knowledge, it is the first time that the transferability of an attacker can be interpreted visually beyond the final feedback from a victim model in speaker recognition.
- We conducted adversarial attacks on four representative black-box victim speaker recognition systems. Extensive experiments demonstrate that the proposed STA-MDCT outperforms the state-of-the-art adversarial attack algorithms by a large margin in transfer-based attack scenarios. The targeted attack success rate (TASR) can reach up to as high as 70.5%.

The rest of this article is organized as follows. Section II reviews the related work of adversarial attacks. Section III briefly describes some preliminaries. Section IV describes the proposed method in detail. Section V provides an interpretable analysis. Section VI shows the experimental setup, including the dataset, victim models, and evaluation metrics. Section VII and VIII analyze the experimental results of adversarial attacks under speaker verification and speaker identification, respectively. In Section IX, we summarize the article.

II. Related Work

In the following, we briefly make a literature survey of adversarial attacks in white-box and black-box scenarios.

White-box attacks can directly access the gradient of a victim model for generating adversarial examples. Since the first work named Fast Gradient Sign Method (FGSM) [28], a large number of adversarial attack approaches have been proposed, including DeepFool [29], I-FGSM [30], PGD [31], C&W [32], and ACG [33]. Besides, universal adversarial perturbations [34], [35], adversarial perturbations generative networks [36], [37], [38] are also extensively explored for generating real-time and efficient adversarial perturbations.

Black-box attacks can be categorized to score-based, decision-based, and transfer-based black-box attacks. For the score-based attacks, gradient-estimation [18] and natural evolution strategies [20], [39] can be used to adapt perturbations to black-box victim models, given frequent queries to the victim models. For the decision-based attacks, boundary attack [40] aims to find the best disturbance around invisible decision boundaries. In [41] formulates the decision-based attack as a real-valued optimization problem that can be solved by any zeroth order optimization algorithm. HopSkipJumpAttack [42] estimates the gradient directions at decision boundaries using Monte Carlo estimation.

Transfer-based attacks leverage the transferability of adversarial examples. Existing approaches that aim to improve the transferability can be categorized into four classes, which are (i)

optimization-based algorithms, (ii) model augmentation strategy [23], (iii) ensemble learning [21] and meta-learning [43] strategies, and (iv) modification of data distributions, respectively. The optimization-based algorithms aim to stabilize the optimization directions of adversarial perturbations and avoid getting trapped in poor local optima of white-box surrogate models, e.g. Nesterov accelerated gradient [21], [44]. The model augmentation strategy aims to simulate diverse models by applying loss-preserving transformations to inputs. For this purpose, Lin et al. [45] utilizes the scale-invariant property of deep neural networks to calculate the gradients over a set of images with different scales. Dong et al. [46] optimize the perturbation over an ensemble of translated images to mitigate the issues of over-reliance on the surrogate model. Long et al. [23] perform the model augmentation by using DCT in the frequency domain, achieving state-of-the-art transfer-based attacks. The ensemble learning and meta-learning strategies generate transferable adversarial examples by integrating gradient information from multiple white-box surrogate models. Both of the strategies are beneficial in decreasing the gap between the surrogate models and the victim models. The approach of modifying data distributions aims to push the input data away from its original distribution to enhance the adversarial transferability [47].

The proposed STA-MDCT belongs to the second and fourth class of the above categories. Different from the above existing methods, we apply MDCT in the generation process of the adversarial voices for the first time, which improves the transferability and interpretability of adversarial voices in the time-frequency domain.

III. Preliminaries

A. Speaker Recognition

This article considers three representative subtasks of speaker recognition, including *automatic speaker verification* (ASV), *open-set identification* (OSI), and *close-set identification* (CSI). We briefly present the definition of the subtasks as follows.

ASV aims to verify whether an anonymous utterance is pronounced by an enrolled person. A state-of-the-art speaker verification system first extracts a speaker embedding, e.g. x-vector, from an input utterance. Then, in the test phase, it calculates the similarity score between the speaker embeddings of an enrollment speaker $\mathbf{x}^{\text{enroll}}$ and a test speaker $\mathbf{x}$. Finally, it compares the score with a predefined threshold θ:

$$s(\mathbf{x}^{\text{enroll}}, \mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \theta, \tag{1}$$

where $s(\mathbf{x}^{\text{enroll}}, \mathbf{x})$ is the similarity score between the two speakers, H_1 represents the hypothesis that $\mathbf{x}$ is uttered by the enrolled speaker, and H_0 is the opposite hypothesis of H_1.

Speaker identification aims to detect the speaker identity of a test utterance $\mathbf{x}$ from an enrollment database of R speakers $\{r = 1, 2, \ldots, R\}$ $(R > 1)$ by:

$$r^* = \underset{r}{\operatorname{argmax}} \{s(\mathbf{x}_r^{\text{enroll}}, \mathbf{x}) | \forall r = 1, \ldots, R\}, \tag{2}$$

where $s(\mathbf{x}_r^{\text{enroll}}, \mathbf{x})$ denotes the similarity score between $\mathbf{x}$ and the r-th enrollment speaker $\mathbf{x}_r^{\text{enroll}}$. If $\mathbf{x}$ can never be out of the R enrolled speakers, then it is a CSI task; otherwise, it is an OSI task. From the above definitions, one can see that ASV is a special case of the OSI task with $R = 1$. However, given the importance and wide applications of ASV in biometric authentication, this article takes it as a separate task.

B. White-Box Adversarial Attack to Speaker Recognition

An adversarial attacker intentionally crafts a tiny perturbation δ that is indistinguishable from humans, and then combines it with a benign voice $\mathbf{x}$ to produce a new one:

$$\mathbf{x}^{\text{adv}} = \mathbf{x} + \delta \tag{3}$$

which may lead a victim model to wrong decisions according to the attacker's proposal. According to the goal of the attacker, adversarial attacks can be divided into *targeted attacks* and *untargeted attacks*. Targeted attacks involve adversarial examples being classified as a specific target label by a neural network. Untargeted attacks only require adversarial examples to be misclassified.

Given a benign voice $\mathbf{x}$, the problem of searching for an adversarial example $\mathbf{x}^{\text{adv}}$ can be formulated as solving the following constrained optimization problem:

$$\begin{aligned} &\max_{\mathbf{x}^{\text{adv}}} \ \mathcal{L}\left(\mathbf{x}^{\text{enroll}}, \mathbf{x}^{\text{adv}}\right) \\ &\text{s.t.} \ \left\|\mathbf{x}^{\text{adv}} - \mathbf{x}\right\|_p < \epsilon \end{aligned} \tag{4}$$

where $\mathcal{L}(\cdot)$ is a loss function that aims to make the victim model to errors, the p-norm $\|\mathbf{x}^{\text{adv}} - \mathbf{x}\|_p$ represents the perturbation degree that controls the energy difference between the benign voice and the adversarial voice, which is upper-bounded by the predefined *perturbation level* ϵ. If ϵ was set small, then humans may not be able to distinguish the adversarial voice $\mathbf{x}^{\text{adv}}$ from the benign voice $\mathbf{x}$. In the following, we derive the definition of the adversarial attack for each speaker recognition subtask from (4).

To attack an ASV system, a targeted attack, a.k.a. *impersonation attack*, aims to make the ASV system misclassify a non-target trial, i.e. $s(\mathbf{x}^{\text{enroll}}, \mathbf{x}) < \theta$, to a target trial $s(\mathbf{x}^{\text{enroll}}, \mathbf{x}^{\text{adv}}) > \theta$, where a target trial indicates that the enrolled utterance and the test utterance belong to the same speaker identity, while a non-target trial is the opposite. To achieve this goal, the loss function $\mathcal{L}_{\text{imp}}$ can be defined as:

$$\mathcal{L}_{\text{imp}}\left(\mathbf{x}^{\text{enroll}}, \mathbf{x}^{\text{adv}}\right) = s(\mathbf{x}^{\text{enroll}}, \mathbf{x}^{\text{adv}}) - \theta. \tag{5}$$

In contrast, an untargeted attack, a.k.a. *evasion attack*, aims to make the ASV system misclassify a target trial $s(\mathbf{x}^{\text{enroll}}, \mathbf{x}) > \theta$ into a non-target trial $s(\mathbf{x}^{\text{enroll}}, \mathbf{x}^{\text{adv}}) < \theta$. Therefore, its loss function can be $\mathcal{L}_{\text{eva}}$ defined as:

$$\mathcal{L}_{\text{eva}}\left(\mathbf{x}^{\text{enroll}}, \mathbf{x}^{\text{adv}}\right) = -s(\mathbf{x}^{\text{enroll}}, \mathbf{x}^{\text{adv}}) + \theta. \tag{6}$$

In one word, in ASV, targeted attacks are impersonation attacks, while untargeted attacks are evasion attacks. In this article, we discuss the tasks of both the impersonation and evasion attacks to ASV.

To attack a speaker identification system, a targeted attack aims to generate an adversarial voice such that the system may misclassify it as a target speaker; whereas an untargeted attack aims to lead the system to wrong predictions. Because a speaker identification system makes errors easily with the presence of natural noise interference which is similar to the effect from untargeted attacks, in this article, we only discuss the tasks of the *targeted attacks* to CSI and OSI. Suppose the attacker aims to make the victim model wrongly predict the r-th speaker into the t-th target speaker. the objective of CSI is defined as:

$$\mathcal{L}_{\text{CSI}}(\mathbf{x}^{\text{adv}}, \{\mathbf{x}_r^{\text{enroll}}\}_{r=1}^{R}) = s(\mathbf{x}_t^{\text{enroll}}, \mathbf{x}^{\text{adv}}) - \max\{s(\mathbf{x}_r^{\text{enroll}}, \mathbf{x}^{\text{adv}}) | \forall r = 1, \ldots, R, r \neq t\}, \tag{7}$$

and the objective of OSI is defined as:

$$\mathcal{L}_{\text{OSI}}(\mathbf{x}^{\text{adv}}, \{\mathbf{x}_r^{\text{enroll}}\}_{r=1}^{R}) = s(\mathbf{x}_t^{\text{enroll}}, \mathbf{x}^{\text{adv}}) - \max(\max\{s(\mathbf{x}_r^{\text{enroll}}, \mathbf{x}^{\text{adv}}) | \forall r = 1, \ldots, R, r \neq t\}, \theta). \tag{8}$$

where θ is a predefined threshold for the open-set problem of OSI.

In the following, we give a brief description of existing attackers, i.e. optimization algorithms, for the aforementioned adversarial attack objectives. For simplicity, we will omit $\mathbf{x}^{\text{enroll}}$ from $\mathcal{L}(\cdot)$ in the rest of the article, unless otherwise stated.

1) Fast Gradient Sign Method (FGSM): FGSM [28] generates an adversarial example $\mathbf{x}^{\text{adv}}$ by maximizing the loss $\mathcal{L}(\mathbf{x}^{\text{adv}})$ with one-step update to $\mathbf{x}$:

$$\mathbf{x}^{\text{adv}} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x})), \tag{9}$$

where $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x})$ is the derivative of the loss function with respect to $\mathbf{x}$.

2) Iterative Fast Gradient Sign Method (I-FGSM): I-FGSM [48] extends FGSM to an iterative version with a small step size α:

$$\mathbf{x}_{t+1}^{\text{adv}} = \text{Clip}_{\mathbf{x},\epsilon}\left\{\mathbf{x}_t^{\text{adv}} + \alpha \cdot \text{sign}\left(\nabla_{\mathbf{x}_t^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right)\right)\right\}. \tag{10}$$

where $\mathbf{x}_0^{\text{adv}} = \mathbf{x}, t \in \{0, 1, \ldots, T\}$ denotes the t-th iteration with T as the maximum number of iterations, $\alpha = \epsilon/T$, and the function $\text{Clip}_{\mathbf{x},\epsilon}(\cdot)$ constrains the generated adversarial examples to be within the ϵ-ball of $\mathbf{x}$ after each optimization step t, i.e., $\|\mathbf{x}_t^{\text{adv}} - \mathbf{x}\|_\infty < \epsilon$.

3) Projected Gradient Descent (PGD): PGD [31] is similar to I-FGSM, but it performs a random initialization to perturbation and replaces the clip operation in (10) with the projection function.

$$\mathbf{x}_{t+1}^{\text{adv}} = \text{Proj}_{\mathbf{x}+\mathcal{S},\epsilon}\left(\mathbf{x}_t^{\text{adv}} + \alpha \frac{\nabla_{\mathbf{x}_t^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right)}{\left\|\nabla_{\mathbf{x}_t^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right)\right\|_2}\right) \tag{11}$$

where $\text{Proj}_{\mathbf{x}+\mathcal{S},\epsilon}$ is the projection operator of L_p, here, we adopt L_2 norm, i.e., $\|\mathbf{x}_t^{\text{adv}} - \mathbf{x}\|_2 < \epsilon$.

4) Momentum Iterative Fast Gradient Sign Method (MI-FGSM): MI-FGSM [49] integrates the momentum into I-FGSM:

$$\begin{aligned} \mathbf{g}_{t+1} &= \mu \cdot \mathbf{g}_t + \frac{\nabla_{\mathbf{x}_t^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right)}{\left\|\nabla_{\mathbf{x}_t^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right)\right\|_1}, \\ \mathbf{x}_{t+1}^{\text{adv}} &= \text{Clip}_{\mathbf{x},\epsilon}\left\{\mathbf{x}_t^{\text{adv}} + \alpha \cdot \text{sign}\left(\mathbf{g}_{t+1}\right)\right\}. \end{aligned} \tag{12}$$

where $\mathbf{g}_0 = 0$, $\mathbf{g}_t$ is the accumulated gradient at iteration t, and μ is the decay factor where $\mu = 1$ in our experiments.

5) Nesterov Iterative Fast Gradient Sign Method (NI-FGSM): NI-FGSM [45] integrates Nesterov accelerated gradient into I-FGSM:

$$\begin{aligned} \mathbf{x}_t^{\text{nes}} &= \mathbf{x}_t^{\text{adv}} + \alpha \cdot \mu \cdot \mathbf{g}_t, \\ \mathbf{g}_{t+1} &= \mu \cdot \mathbf{g}_t + \frac{\nabla_{\mathbf{x}_t^{\text{nes}}} \mathcal{L}(\mathbf{x}_t^{\text{nes}})}{\left\|\nabla_{\mathbf{x}_t^{\text{nes}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right)\right\|_1}, \\ \mathbf{x}_{t+1}^{\text{adv}} &= \text{Clip}_{\mathbf{x},\epsilon}\left\{\mathbf{x}_t^{\text{adv}} + \alpha \cdot \text{sign}\left(\mathbf{g}_{t+1}\right)\right\}. \end{aligned} \tag{13}$$

6) Auto Conjugate Gradient Attack: Auto conjugate gradient attack (ACG) [33] is based on conjugate gradient descent:

$$\begin{aligned} \mathbf{y}_{t-1} &= \nabla_{\mathbf{x}_{t-1}^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_{t-1}^{\text{adv}}\right) - \nabla_{x_t^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right) \\ \beta_t^{HS} &= \frac{\left\langle -\nabla_{\mathbf{x}_t^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right), \mathbf{y}_{t-1}\right\rangle}{\left\langle \mathbf{s}_{t-1}, \mathbf{y}_{t-1}\right\rangle} \\ \mathbf{s}_t &= \nabla_{\mathbf{x}_t^{\text{adv}}} \mathcal{L}\left(\mathbf{x}_t^{\text{adv}}\right) + \beta_t^{HS} \mathbf{s}_{t-1} \\ \mathbf{x}_{t+1}^{\text{adv}} &= \text{Clip}_{\mathbf{x},\epsilon}\left\{\mathbf{x}_t^{\text{adv}} + \eta_t \cdot \text{sign}\left(\mathbf{s}_t\right)\right\} \end{aligned} \tag{14}$$

where the initial conjugate gradient $\mathbf{s}_0 = 0$, β^{HS} is a parameter calculated from the past search information, the step size η_t is dynamically adjusted and initialized by $\eta_0 = 2\epsilon/T$. Particularly, when the number of iterations reaches a predefined value or the loss no longer drops, η is halved.

From the above formulation, we see that ACG updates the search points in broader directions than the steepest gradient direction as that in FGSM, which may improve the transferability of the generated adversarial examples.

IV. Spectrum Transformation Attack Based on Modified Discrete Cosine Transform

In this section, we first present the STA-MDCT framework in Section IV-A, and then describe the spectrum transformation in detail in Section IV-B. Finally, in Section IV-C, we present two implementations of STA-MDCT, one with a single attacker and the other with an ensemble of attackers.

A. Framework of the Transfer-Based STA-MDCT Attack

Existing works usually apply loss-preserving transformations in the time domain, which might overlook the difference between frequency bands of speech signals, while slight variations in frequency components could lead to distinctly different decisions. Given the same input, different victim models attend to different frequency bands and spectrum features of the input for making a decision [27], [50]. Therefore, a perfect attacker trained from a white-box surrogate model in the time domain may have weak transferability to a black-box victim model that

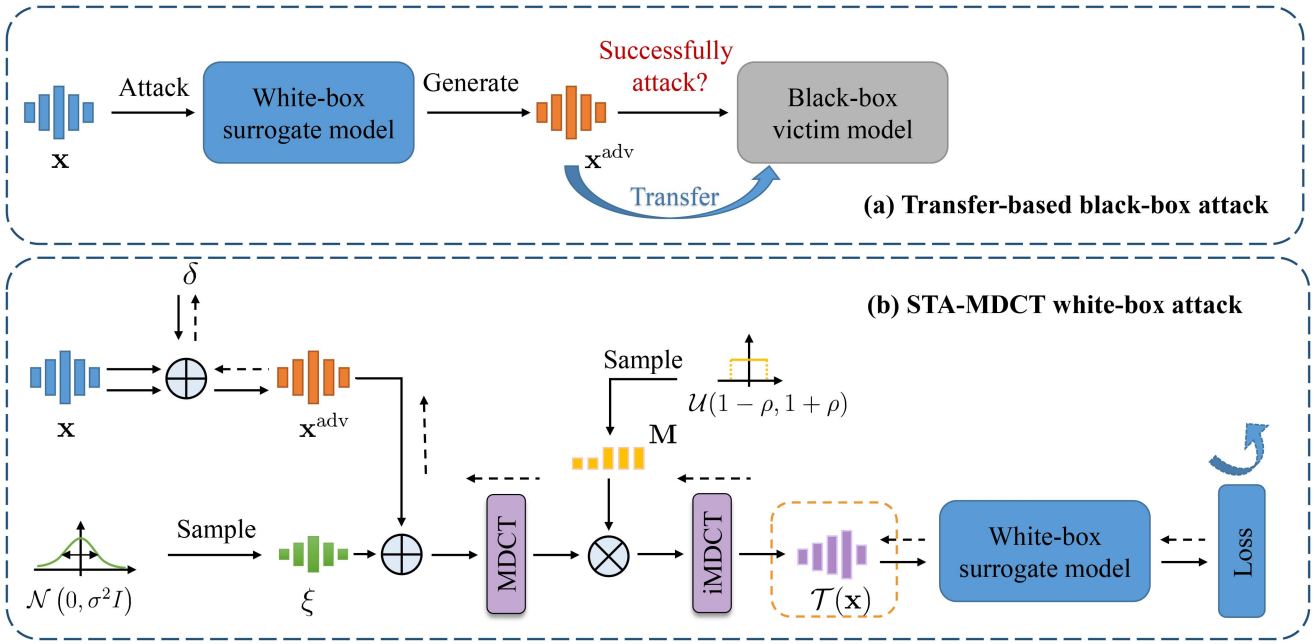


Fig. 1. Framework of the STA-MDCT transfer-based attack, where $\mathbf{x}$ is a benign voice, $\mathbf{x}^{\text{adv}}$ is an adversarial voice from $\mathbf{x}$, δ is the adversarial perturbation, ξ is a random noise signal whose sample points are sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2 I)$, $\mathbf{M}$ is a random matrix whose elements are sampled from a uniform distribution $\sim \mathcal{U}(1-\rho, 1+\rho)$, and $\mathcal{T}(\mathbf{x})$ is the spectrum transformation of $\mathbf{x}$.

has different attention properties in the time-frequency domain from the surrogate model.

To address this issue, Long et al. propose a spectrum transformation based on DCT to diversify input images. In contrast to DCT, MDCT is more suitable for feature extraction and signal analysis of audio signals because it avoids the time-domain aliasing introduced by DCT. Therefore, we propose STA-MDCT based on MDCT to explore the correlation between the victim models in the time-frequency domain. Its framework is illustrated in Fig. 1. As shown in Fig. 1(a), a transfer-based attacker first attacks a white-box surrogate model for generating transferable adversarial examples, and then uses them to attack a black-box target model. As shown in Fig. 1(b), for each optimization iteration, STA-MDCT first applies a spectrum transformation $\mathcal{T}(\mathbf{x}^{\text{adv}})$ based on MDCT and inverse MDCT (iMDCT) to $\mathbf{x}^{\text{adv}}$, and then reallocates the energy of the frequency bands of $\mathbf{x}^{\text{adv}}$ according to the gradient information from the white-box surrogate model for improving the transferability of $\mathbf{x}^{\text{adv}}$.

B. Spectrum Transformation

The spectrum transformation $\mathcal{T}(\mathbf{x}^{\text{adv}})$ is defined as follows:

$$\begin{aligned}\mathcal{T}(\mathbf{x}^{\text{adv}}) &= \text{iMDCT}((\text{MDCT}(\mathbf{x}^{\text{adv}}) + \text{MDCT}(\xi)) \odot \mathbf{M}),\\ &= \text{iMDCT}(\text{MDCT}(\mathbf{x}^{\text{adv}} + \xi) \odot \mathbf{M})\end{aligned} \tag{15}$$

where ξ is a random noise signal whose sample points are sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2 I)$, $\mathbf{M}$ is a random matrix whose elements are sampled from a uniform distribution $\sim \mathcal{U}(1-\rho, 1+\rho)$, the operator $\odot$ refers to the Hadamard product. The key contribution of ξ and $\mathbf{M}$ lies in their ability to enhance the transferability of adversarial examples. Both ξ and $\mathbf{M}$ play a crucial role in manipulating the spectrum saliency map. By leveraging ξ and $\mathbf{M}$ simultaneously, we can effectively simulate a more diverse substitute model and generate transferable adversarial examples. This theory was further confirmed through the ablation study conducted in Section VII-C.

The operator MDCT$(\cdot)$ [51] is defined as:

$$\begin{aligned}X_{\text{MDCT}}(k) = \textstyle\sum_{n=0}^{W-1} x(n)h(n)\cos\left[\frac{\left(2n+1+\frac{W}{2}\right)(2k+1)\pi}{2W}\right],\\ \forall k = 0, 1, \ldots, \tfrac{W}{2}-1, \quad \forall n = 0, 1, \ldots, W-1\end{aligned} \tag{16}$$

and the iMDCT operator iMDCT$(\cdot)$ is defined as:

$$\begin{aligned}x(n) = \tfrac{2}{W}h(n)\textstyle\sum_{k=0}^{\frac{W}{2}-1} X_{\text{MDCT}}(k)\cos\left[\frac{\left(2n+1+\frac{W}{2}\right)(2k+1)\pi}{2W}\right],\\ \forall k = 0, 1, \ldots, \tfrac{W}{2}-1, \quad \forall n = 0, 1, \ldots, W-1\end{aligned} \tag{17}$$

where $h(n)$ represents the Kaiser-bessel-derived window, W denotes the window length of the transformation.

Note that, MDCT is a linear orthogonal lapped transform, based on the time domain aliasing cancellation. In this article, the adjacent frames produced by MDCT has an overlap of 50%. This enables a smooth transition between the time domain and frequency domain, contributing to improved time-frequency performance. Additionally, the use of the Kaiser-bessel-derived window in MDCT is better suited to adapt to the frequency characteristics of audio signals.

C. STA-MDCT Implementations

Any gradient-based attackers can be applied to the proposed STA-MDCT framework. Different from (4), the optimization objective of STA-MDCT is formulated as:

$$\max_{\mathbf{x}^{\text{adv}}} \mathcal{L}\left(\mathbf{x}^{\text{enroll}}, \mathcal{T}(\mathbf{x}^{\text{adv}})\right) \quad \text{s.t.} \ \left\|\mathbf{x}^{\text{adv}} - \mathbf{x}\right\|_p < \epsilon \tag{18}$$

Its optimization iteratively operates the following three steps:

- Calculate $\mathcal{T}(\mathbf{x}_t^{\text{adv}})$ which is then back-propagated through the network to obtain the gradient information $\nabla_{\mathbf{x}_t^{\text{adv}}}\mathcal{L}(\mathcal{T}(\mathbf{x}_t^{\text{adv}}))$, where t denotes the t-th iteration.
- Average N gradients to obtain a more stable gradient direction.
- Update the adversarial example $\mathbf{x}_{t+1}^{\text{adv}}$ using an attacker algorithm, e.g. FGSM, I-FGSM, etc.

In this article, we implement two STA-MDCTs, one with a single white-box surrogate model, and the other with an ensemble of white-box surrogate models.

1) STA-MDCT Based on a Single Surrogate Model: We apply I-FGSM [48] to STA-MDCT for attacking a white-box surrogate model. The optimization algorithm is formulated as:

$$\mathbf{x}_{t+1}^{\text{adv}} = \text{Clip}_{\mathbf{x},\epsilon}\left\{\mathbf{x}_t^{\text{adv}} + \alpha \cdot \text{sign}\left(\frac{1}{N}\sum_{i=1}^{N}\nabla_{\mathbf{x}_t^{\text{adv}}}\mathcal{L}\left(\mathcal{T}\left(\mathbf{x}_t^{\text{adv}}\right)\right)\right)\right\} \tag{19}$$

Additionally, we conducted experiments combining STA-MDCT with PGD (STA-MDCT-PGD) to confirm the effectiveness of our proposed transformation.

2) STA-MDCT Based on an Ensemble of Surrogate Models: To improve transferability, we apply I-FGSM to STA-MDCT for attacking an ensemble of white-box surrogate models (ensemble-STA-MDCT). The algorithm is summarized in Algorithm 1, when there is only one white-box model, i.e., $q = 1$, it corresponds to Section IV-C1.

Note that the following STA-MDCT refers to STA-MDCT combined with I-FGSM unless otherwise specified.

Algorithm 1: STA-MDCT.

Input White-box models $F = \{f_1, \ldots, f_q\}$, ensemble weight $w = [w_1, w_2, \ldots, w_q]$, the enroll utterance $\mathbf{x}^{\text{enroll}}$, the testing utterance $\mathbf{x}$ to be attacked, loss function $\mathcal{L}_{f_j}(\cdot)$ for model f_j, max iterations T, max perturbation ϵ, step size α, number of spectrum transformation N, tunning factor $\boldsymbol{\rho}$, std $\boldsymbol{\sigma}$ of noise ξ.

Output An adversarial example $\mathbf{x}^{\text{adv}}$.

$\mathbf{x}_0^{\text{adv}} = \mathbf{x}$;
for $t = 0$ to $T - 1$ **do**
  for all f_j **do**
    Get spectrum transformation output $\mathcal{T}(\mathbf{x}_t^{\text{adv}})$ using: $\mathcal{T}(\mathbf{x}_t^{\text{adv}}) = \text{iMDCT}(\text{MDCT}(\mathbf{x}_t^{\text{adv}} + \xi) \odot \mathbf{M})$;
    Compute the average gradient of the N augmented models: $\mathbf{k}_{f_j} = \frac{1}{N}\sum_{i=1}^{N}\nabla_{\mathbf{x}_t^{\text{adv}}}\mathcal{L}_{f_j}(\mathcal{T}(\mathbf{x}_t^{\text{adv}}))$
  end
  Fuse these gradients: $\mathbf{k} = \sum_{j=1}^{q} w_j \mathbf{k}_{f_j}$;
  Update $\mathbf{x}_{t+1}^{\text{adv}}$ by applying the sign gradient as: $\mathbf{x}_{t+1}^{\text{adv}} = \text{Clip}_{\mathbf{x},\epsilon}(\mathbf{x}_t^{\text{adv}} + \alpha \cdot \text{sign}(\mathbf{k}))$;
end
$\mathbf{x}^{\text{adv}} = \mathbf{x}_T^{\text{adv}}$;
return $\mathbf{x}^{\text{adv}}$

V. Interpretability of STA-MDCT With Saliency Maps

In this section, we first introduce saliency maps in Section V-A, and then apply a special saliency map, named Layer-CAM, to interpret the effectiveness of STA-MDCT in the time-frequency domain directly in Section V-B.

A. Saliency Maps for Speaker Recognition

When making decisions, humans tend to focus on salient parts of an object and allocate their attention appropriately. Class activation map (CAM) is the saliency map of an image produced by a convolutional neural network (CNN) which emphasizes important regions for classifying the image. Several CAMs have been widely used in computer vision, such as the Grad-CAM [52], Grad-CAM++ [53], Score-CAM [54] and Layer-CAM [55]. In speech processing, Li et al. [27] applied CAM to speaker recognition. Their study concludes that only Layer-CAM is a valid visualization tool for speaker recognition. This article discusses Layer-CAM as well.

Here, we make a brief description of Layer-CAM in speaker recognition. Given a CNN-based speaker recognition system, we denote the output feature maps of the final convolutional layer of the CNN as $\mathbf{A}$ and denote the k-th feature map in $\mathbf{A}$ as $\mathbf{A}^k$. Suppose the predicted score of the input $\mathbf{x}$ for the c-th speaker is:

$$y^c = s(\mathbf{x}_c^{\text{enroll}}, \mathbf{x}). \tag{20}$$

We calculate the gradient of y^c with respect to $\mathbf{A}^k$ by:

$$w_{ij}^{kc} = \text{relu}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right). \tag{21}$$

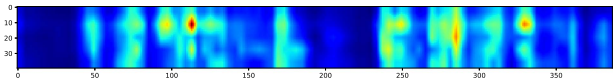
where A_{ij}^k is the (i,j)-th location of $\mathbf{A}^k$ The saliency map $\mathbf{Z}^c$ of the input $\mathbf{x}$ at the location (i,j) is:

$$Z_{ij}^c = \text{relu}\left\{\sum_k w_{ij}^{kc} \cdot A_{ij}^k\right\}. \tag{22}$$

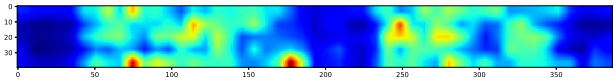
At last, we *normalize* the saliency map $\mathbf{Z}^c$ by:

$$\hat{\mathbf{Z}}^c = \frac{\mathbf{Z}^c - \min \mathbf{Z}^c}{\max \mathbf{Z}^c - \min \mathbf{Z}^c}. \tag{23}$$

As illustrated in Fig. 2, given the same input $\mathbf{x}$, the saliency maps of different models for the same speaker c significantly vary from each other, which clearly reveals that the models have different concerns on the same time-frequency unit. It is precisely due to the differences in spectrum saliency maps among different models that we are able to adjust the data distribution to alter the spectrum saliency maps, thereby simulating more diverse models to generate transferable adversarial examples.

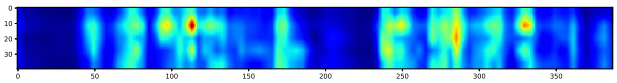


(a) ResNetSE34L

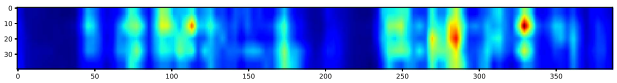


(b) ResNetSE34V2

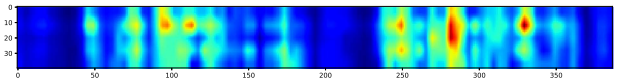
Fig. 2. Saliency maps for the speaker recognition models ResNetSE34L [56] and ResNetSE34V2 [57]. The regions with light and warm colors are critical regions indicating important time-frequency components in making decisions.



(a) Original test voice



(b) Adversarial test voice made by I-FGSM [48]



(c) Adversarial test voice made by STA-MDCT

Fig. 3. Example of saliency maps, given the enrollment voice from the ground-truth speaker.

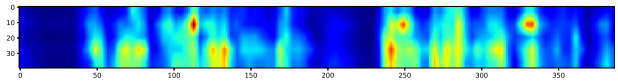
B. Interpretable STA-MDCT With Layer-CAM

Based on the observation in Section V-A, if an attacker could successfully make a victim model shift its attention to the saliency map, then a successful attack may be made. Here we apply it to visually explain the effects of different transfer-based attackers on black-box victim models.

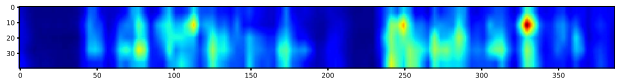
First, we select an utterance from the ground-truth speaker, and generate its adversarial voices by applying I-FGSM and STA-MDCT respectively to the ECAPA-TDNN [58] white-box surrogate model. Then, we apply the adversarial voices to attack the ResNetSE34L [56] black-box victim model. Finally, we calculate the saliency maps of the above voices produced from ResNetSE34 L, given the enrollment voices either from the ground-truth speaker or from the target speaker.

Fig. 3 shows the saliency maps of the test voices, given the enrollment voice from the ground-truth speaker. Comparing Fig. 3(a) with Fig. 3(b), we see that the attentive region of the saliency map of the adversarial voice generated by I-FGSM is quite similar to that generated from the original voice, which indicates that the transfer-based attack based on I-FGSM fails to shift the attention of the black-box victim model. On the contrary, comparing Fig. 3(a) with Fig. 3(c), we see that STA-MDCT effectively shifts the victim model's attention from critical regions to other regions, which may lead to a classification error.

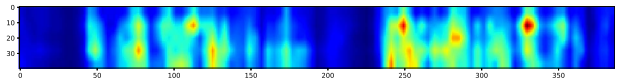
Fig. 4 shows the saliency maps of the test voices, given the enrollment voice from the target speaker. Comparing Fig. 4(a) with Fig. 3(a), we see that, when the victim model verifies the same



(a) Original test voice



(b) Adversarial voice made by I-FGSM [48]



(c) Adversarial voice made by STA-MDCT

Fig. 4. Example of saliency maps, given the enrollment voice from the target speaker.

test voice with different enrollment voices, the critical regions are shifted significantly. Comparing Fig. 4(a) with Fig. 4(b), we see that the attentive region of the saliency map generated by I-FGSM is quite different from that of the original voice, which indicates that I-FGSM fails to make the targeted attack. On the other side, comparing Fig. 4(a) with Fig. 4(c), we see that saliency maps of the original voice and the adversarial voice made by STA-MDCT share similar critical regions, which may lead to a successful targeted attack.

VI. Experimental Setup

In this section, we introduce the datasets, comparison attackers, evaluation metrics, and victim models.

A. Datasets

We first built four speaker recognition systems and then conducted adversarial attack experiments to them. See Section VI-D for the detailed training process of the speaker recognition systems. The adversarial attack experiments were conducted on the VoxCeleb [59] and LibriSpeech [60] datasets:

To attack an ASV system, we arbitrarily selected 1,000 trials, including 40 speakers, from the *Original-Clean* trial list of VoxCeleb1, which includes 500 target trials and 500 non-target trials. Then, an attacker transforms the clean test voices of all trials into adversarial examples which aim to make the ASV system yield opposite predictions from their ground-truth speaker identities.

For a speaker identification system, we first selected 10 speakers randomly, including 5 males and 5 females, from the *test-other* and *dev-other* subsets of LibriSpeech as the enrollment speakers, each of which contains 10 utterances.

To attack the speaker identification system in the CSI setting, we first picked 10 test speakers whose identities are the same as the enrollment speakers. Each of the speakers contains 100 randomly selected utterances that are *different* from the enrollment data of the speaker. Then, we conducted targeted attacks in two scenarios: random target attack and farthest target attack. In the case of a random target attack, an attacker randomly picked a speaker from the 10 enrollment speakers that were different from the ground-truth identity of the utterance as the

target speaker. On the other hand, in the farthest target attack, the target speaker was chosen to be farthest in the embedding space from the ground-truth speaker. Finally, it performed the targeted attack by generating an adversarial example from the clean test utterance, which aims to make the victim system wrongly predict the speaker's identity as the targeted label.

To attack the speaker identification system in the OSI setting, we arbitrarily chose 10 test speakers that were different from the enrollment speakers from the *train-other-500* subset of LibriSpeech. Each of the speakers contains 100 randomly selected utterances. As mentioned earlier, we conducted random target attacks and farthest target attacks.

B. Comparison Adversarial Attackers

The parameter setting of the proposed STA-MDCT was that $N = 20$, $\rho = 0.75$, and $\sigma = 44$ in all experiments unless otherwise stated.

We compare the proposed method with FGSM [28], I-FGSM [48], PGD [31], MI-FGSM [49], NI-FGSM [45] and ACG [33], see Section III-B for the descriptions of the comparison methods. All comparison methods were performed in their best settings or recommended default settings. That is, the maximum perturbation $\epsilon = 40$, the iteration $T = 10$, and the step size $\alpha = \epsilon/T = 4$.

Both the white-box surrogate speaker recognition models and black-box victim models were selected from ResNetSE34 L, ECAPA-TDNN, ResNetSE34V2, and RawNet3.

C. Evaluation Metrics

The attacking effect of an attacker to a victim speaker recognition system was evaluated in terms of the targeted attack success rate (TASR), false acceptance rate (FAR), equal error rate (EER), identification error rate (IER) and normalized minimum detection cost function (minDCF) with $P_{\text{tar}} = 0.05$ and $C_{\text{miss}} = C_{\text{fa}} = 1$, produced from the victim system, where TASR refers to the proportion of the generated adversarial voices that are recognized as the targeted labels, and IER is the proportion of the input voices that are misclassified by the model. The higher the evaluation scores are, the better the attacking effect is.

To measure the stealthiness of the adversarial examples, we used signal-to-noise (SNR), perceptual evaluation of speech quality (PESQ) [61], and the standard L_2 norm. SNR is defined as $\text{SNR} = 10\log_{10}(P_{\mathbf{x}}/P_{\boldsymbol{\delta}})$ where $P_{\mathbf{x}}$ and $P_{\boldsymbol{\delta}}$ are the signal power of the benign voice $\mathbf{x}$ and the power of the perturbation $\boldsymbol{\delta}$ respectively. PESQ first applies an auditory transformation to obtain the loudness spectra of the benign voices and the adversarial voices and then compares both loudness spectra to obtain a metric score with a value in the range of $[-0.5, 4.5]$, see [61] for the details. Larger SNR, higher PESQ and smaller L_2 indicate better stealthiness.

D. Victim Speaker Recognition Systems

First, we trained four representative speaker recognition systems, which are the ResNetSE34L [56], ECAPA-TDNN [58], ResNetSE34V2 [57] and RawNet3 [62] respectively, on the development set of VoxCeleb2 [63]. Then, we applied the speaker recognition systems as the victim models for the tasks of the adversarial attack to ASV, CSI, and OSI, respectively. The parameter settings of the victim models are summarized as follows.

ResNetSE34 L adopts attentive average pooling and uses Angular Prototypical as the loss function [56]. ECAPA-TDNN adopts attentive statistical pooling and AAM-Softmax [64]. ResNetSE34V2 uses attentive statistical pooling and takes the joint loss of the Angular Prototypical loss and softmax loss. RawNet3 uses attentive statistical pooling and AAM-Softmax.

In respect of the input acoustic features, RawNet3 uses raw waveforms as its input. The other three models first extract spectrograms with a hamming window of width 25 ms and step size 10 ms and then apply log Mel-filterbanks to the spectrograms, followed by cepstral mean and variance normalization (CMVN).

To justify the advantage of the victim models, we evaluated them on the VoxCeleb1 *Original-Clean* trial list which contains 37,000 trials from 40 speakers. The evaluation metrics include EER, minDCF, and IER. Table I summarizes the speaker recognition performance of the models. From the table, we see that the models achieve the state-of-the-art performance.

Note that, the optimal decision thresholds θ were determined when the models were evaluated on the VoxCeleb1 *Original-Clean* trial list. They were fixed thereafter, e.g. when we applied the attackers to the victim models.

TABLE I
PERFORMANCE OF THE SPEAKER RECOGNITION MODELS WITHOUT ATTACKERS

Models	ASV			OSI			CSI
	EER (%)	MinDCF	θ	EER (%)	IER (%)	θ	IER (%)
ResNetSE34L [56] + Angular Prototypical	2.179	0.168	0.051	2.8	0	0.72	0
ECAPA-TDNN [58] + AAM-Softmax	1.172	0.08	0.03	1.0	0	0.51	0
ResNetSE34V2 [57] + AP_Softmax	1.023	0.083	0.034	1.2	0	0.57	0
RawNet3 [62] + AAM-Softmax	1.039	0.069	0.03	1.9	0	0.49	0

The term "AP_ Softmax" refers to the joint loss function of the Angular Prototypical loss and Softmax loss. Note that CSI sets θ to 0 as the default.

VII. RESULTS OF ADVERSARIAL ATTACKS TO SPEAKER VERIFICATION

In this section, we first report the comparison results between the STA-MDCT and the comparison attackers with a single white-box surrogate ASV model in Section VII-A, and with an ensemble of white-box surrogate models in Section VII-B. Then, we study the effects of the hyperparameters of STA-MDCT on performance in Section VII-C, and the effect of the SNR budget in Section VII-D, given a single white-box surrogate ASV model.

A. Results With a Single White-Box Surrogate ASV Model

Table II lists the comparison result between the single-model attackers and the proposed STA-MDCT on the ASV task.

TABLE II
PERFORMANCE OF THE COMPARISON ATTACKERS THAT USE THE SAME WHITE-BOX SURROGATE MODEL, ON THE ASV TASK IN TERMS OF EER (%), TASR (%), FAR (%), AND MINDCF

Surrogate Model	Attack	Victim Model: ResNetSE34L				ECAPA-TDNN				ResNetSE34V2				RawNet3						
		EER	TASR	FAR	MinDCF	EER	TASR	FAR	MinDCF	EER	TASR	FAR	MinDCF	EER	TASR	FAR	MinDCF	SNR(dB)	PESQ	L2
ResNetSE34L	FGSM [28]	58.2	59.5	42.4	1.00	9.2	9.5	7.6	0.58	9.4	10.0	6.4	0.63	7.4	7.2	5.0	0.53	30.11	3.12	1.47
	I-FGSM [48]	97.2	95.6	92.6	1.00	16.0	20.5	8.8	0.73	18.6	24.2	7.2	0.81	9.8	13.0	6.0	0.70	37.02	4.13	0.63
	PGD [31]	95.6	93.6	89.0	1.00	16.6	20.7	8.8	0.80	18.4	33.43	7.8	0.84	11.8	14.2	6.0	0.72	33.43	3.64	0.98
	MI-FGSM [49]	96.8	95.8	93.0	1.00	22.6	26.3	10.8	0.89	25.6	30.1	11.6	0.96	16.2	19.4	8.0	0.84	32.61	3.59	1.09
	NI-FGSM [45]	**97.6**	**96.5**	**94.6**	1.00	23.8	27.1	11.4	0.90	26.6	31.6	12.0	0.96	17.0	20.6	8.8	0.85	32.65	3.55	1.10
	ACG [33]	**97.6**	94.2	89.8	1.00	31.8	32.5	14.6	0.93	35.2	38.1	16.2	0.96	23.6	26.6	12.2	0.84	32.19	3.53	1.14
	STA-DCT [23]	90.6	87.6	80.2	1.00	32.0	33.2	16.4	0.95	32.0	34.9	16.0	0.96	26.6	28.2	12.8	0.94	34.65	3.92	0.84
	STA-MDCT	97.2	96.1	93.6	1.00	44.0	44.2	**24.6**	0.98	43.4	46.9	**24.8**	0.99	37.0	37.0	**20.0**	0.98	34.18	3.93	0.89
	STA-MDCT-PGD	96.4	95.6	92.4	1.00	**45.4**	**45.2**	24.2	1.00	**44.8**	**47.5**	24.2	0.99	**38.0**	**38.1**	19.8	0.98	35.24	4.08	0.78
ECAPA-TDNN	FGSM [28]	19.8	22.8	13.6	0.92	62.0	59.5	52.0	1.00	13.2	15.7	8.6	0.83	18.0	20.0	13.0	0.91	30.12	3.10	1.47
	I-FGSM [48]	32.0	35.6	20.6	0.98	98.0	98.0	98.0	1.00	26.4	29.8	13.4	0.96	40.0	38.9	21.8	1.00	36.30	4.12	0.68
	PGD [31]	34.6	38.0	21.4	0.99	96.8	96.8	96.8	1.00	28.8	33.1	16.0	0.97	45.8	32.6	23.8	1.00	33.26	3.63	1.00
	MI-FGSM [49]	40.2	45.7	25.4	0.99	97.4	97.6	97.2	1.00	36.8	40.2	20.8	0.99	51.6	46.4	26.8	1.00	32.54	3.56	1.09
	NI-FGSM [45]	38.6	44.5	25.2	0.99	97.4	97.7	98.2	1.00	35.8	38.0	20.2	0.99	50.2	45.8	27.6	0.99	32.85	3.55	1.08
	ACG [33]	47.0	50.2	26.0	0.99	**98.8**	**98.7**	**98.4**	1.00	49.0	49.5	27.4	0.99	64.8	59.2	42.0	1.00	31.88	3.49	1.18
	STA-DCT [23]	49.4	53.6	31.0	1.00	93.2	92.6	91.4	1.00	45.4	48.6	26.4	0.99	60.6	54.8	37.0	1.00	33.92	3.88	0.92
	STA-MDCT	**58.8**	**60.6**	37.0	1.00	97.8	97.8	97.6	1.00	58.2	56.5	33.8	1.00	72.2	68.5	54.0	1.00	33.43	3.84	0.97
	STA-MDCT-PGD	**58.8**	**60.6**	**37.4**	1.00	97.4	97.3	97.2	1.00	**59.0**	**57.8**	**34.4**	1.00	**77.0**	**61.3**	**56.4**	1.00	34.58	4.00	0.85
ResNetSE34V2	FGSM [28]	14.4	16.7	11.2	0.82	9.2	9.6	7.8	0.58	43.4	44.0	29.2	0.99	7.6	7.5	5.2	0.53	30.11	3.09	1.47
	I-FGSM [48]	23.8	27.1	17.2	0.98	16.0	19.0	10.6	0.79	96.8	94.9	92.0	1.00	12.2	14.4	8.4	0.73	37.18	4.16	0.62
	PGD [31]	26.0	29.8	18.4	0.98	19.2	21.3	11.0	0.83	94.2	92.6	99.8	1.00	15.2	16.5	8.4	0.75	33.45	3.67	0.98
	MI-FGSM [49]	32.2	36.7	22.0	0.99	24.8	27.3	14.6	0.93	96.6	95.1	92.6	1.00	19.4	22.7	11.4	0.86	32.71	3.60	1.08
	NI-FGSM [45]	31.0	35.7	21.2	0.99	25.0	26.8	14.6	0.95	96.2	95.6	93.2	1.00	19.4	22.4	11.6	0.88	33.10	3.59	1.06
	ACG [33]	34.4	38.5	20.0	0.99	30.4	32.0	17.2	0.93	**97.8**	**96.4**	**93.8**	1.00	24.6	27.4	14.4	0.87	31.79	3.47	1.21
	STA-DCT [23]	43.6	47.8	27.4	0.99	37.8	38.5	23.6	0.96	89.0	87.9	83.0	1.00	31.8	35.3	20.8	0.98	34.69	3.97	0.84
	STA-MDCT	**54.8**	**56.9**	**35.2**	1.00	51.2	**49.9**	**31.8**	1.00	96.6	94.9	92.4	1.00	46.2	44.4	26.8	1.00	34.09	3.99	0.90
	STA-MDCT-PGD	53.0	56.1	**35.2**	1.00	**52.4**	49.7	31.0	0.99	95.6	94.6	91.6	1.00	**46.8**	**44.8**	**27.8**	1.00	35.12	4.15	0.79
RawNet3	FGSM [28]	4.0	4.5	2.8	0.30	1.8	1.9	2.2	0.14	1.2	1.2	1.2	0.07	2.6	2.8	2.0	0.21	30.11	3.14	1.47
	I-FGSM [48]	4.0	4.1	4.4	0.30	2.6	2.4	2.8	0.17	1.6	1.5	1.6	0.11	**19.4**	19.6	**14.8**	0.90	40.00	4.19	0.47
	PGD [31]	2.4	2.4	2.4	0.18	1.2	1.2	1.6	0.06	0.8	0.9	1.2	0.07	**2.4**	2.4	**2.2**	0.19	34.19	3.69	0.92
	MI-FGSM [49]	**5.2**	**5.2**	**4.8**	0.34	**3.4**	3.4	**3.2**	0.20	1.6	1.9	2.2	0.15	15.4	17.0	12.2	0.85	32.99	3.49	1.06
	NI-FGSM [45]	4.4	5.1	4.2	0.31	**3.4**	**3.5**	**3.2**	0.20	**2.0**	**2.0**	**2.4**	0.15	19.0	**20.5**	14.6	0.89	33.37	3.54	1.02
	ACG [33]	3.8	3.9	3.2	0.29	1.8	2.0	2.4	0.14	1.2	1.2	1.4	0.09	4.4	4.9	3.8	0.37	32.42	3.41	1.14
	STA-DCT [23]	3.4	3.5	3.6	0.24	1.4	1.6	1.8	0.11	1.2	1.0	1.2	0.08	3.2	3.4	3.0	0.26	40.09	4.20	0.46
	STA-MDCT	3.0	2.9	3.0	0.19	1.4	1.4	1.8	0.09	1.0	1.0	1.2	0.09	2.6	2.5	2.4	0.18	40.09	4.24	0.47
	STA-MDCT-PGD	2.6	2.5	3.0	0.17	1.0	1.3	1.6	0.06	0.4	0.5	0.6	0.07	1.2	1.4	1.8	0.06	42.46	4.33	0.35

The result in gray color indicates that it is a white-box attack where the victim model is also the surrogate model. The thresholds θ with respect to the TASR and FAR are listed in Table I.

From the comparison, we see that the proposed method consistently outperforms the comparison attackers. For example, the proposed STA-MDCT achieves relative EER improvement of 42.56%, 44.18%, and 31.1%, respectively, over MI-FGSM, NI-FGSM and ACG. Meanwhile, MDCT is more suitable for frequency characteristics of audio signals than DCT. Particularly, adversarial examples generated with the ECAPA-TDNN surrogate ASV model tend to have better transferability to other black-box victim ASV models. Although adversarial examples generated with RawNet3 yield poor transferability to other victim models, the contrary transfer direction works fine, which indicates that adversarial examples generated from spectrum features are better than those generated from raw waves.

We should note that Table II also lists the result of the white-box attacks in gray color, where the surrogate model and victim model are the same. From the result, we see that the attack performance of the proposed method is slightly weaker than ACG and NI-FGSM. It can be explained that the proposed STA-MDCT, which improves the generalization ability of the adversarial examples to new black-box victim models, reduces the overfitting phenomenon of the adversarial examples to the surrogate models where they are generated.

Furthermore, we compared the attack algorithms in a scenario where the False Rejection Rate (FRR) as a constant. The experimental results in Table III demonstrate that the proposed STA-MDCT achieves the highest FAR in the black-box transfer attack scenario.

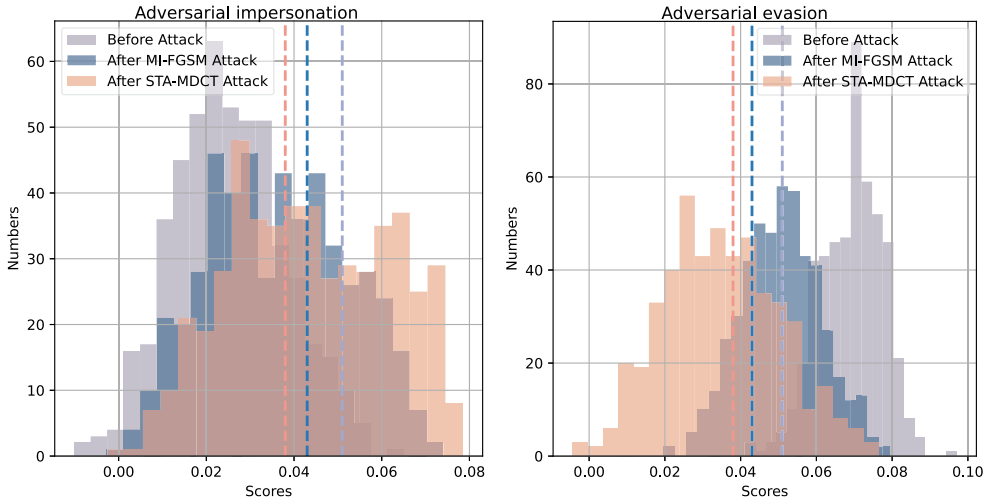
TABLE III
IMPACT OF TRANSFER-BASED ATTACKS ON THE FALSE ACCEPTANCE RATE (FAR) AT AN FRR OF 5% USING A SINGLE SURROGATE MODEL RESNETSE34 L

Victim model	FAR (%)				
	No-attack	PGD	MI-FGSM	ACG	STA-MDCT
ResNetSE34L	1.2	100.0	100.0	100.0	100.0
ECAPA-TDNN	0.0	37.6	56.6	80.8	97.6
ResNet34V2	0.2	36.0	55.2	81.8	92.6
RawNet3	0.2	25.6	41.4	60.2	88.8

Here we need to emphasize the reason why the EER values in Table II can be over 50% as follows. In adversarial attacks on ASV, we conducted evasion and impersonation attacks. When calculating the EER, we took into account both of these scenarios. Therefore, in an ideal situation where all attacks are successful, the EER can be close to 100%. We can also explain it using formulas:

$$FRR = \frac{FN}{FN + TP}, FAR = \frac{FP}{FP + TN} \tag{24}$$

where FN, TP, TN, and FP represent False Negative, True Positive, True Negative, and False Positive, respectively. (FN + TP) represents the number of target trials, while (FP + TN) represents the number of non-target trials, corresponding to evasion and impersonation attacks, respectively. The error rates for both types of attacks range from 0% to 100%. Therefore, the Equal Error Rate (EER) also falls within the range of 0% to 100%.



(a) (b)

Fig. 5. Histograms of the similarity scores of the trials before and after the attacks, where the dashed lines represent the respective thresholds. Left: Adversarial impersonation on non-target trials. Right: Adversarial evasion on target trials.

In addition to EER and attack success rate, we also evaluated the impact of the adversarial attacks by measuring the changes in similarity scores between the target and non-target trials. The experimental results, shown in Fig. 5, demonstrate the effectiveness of the ECAPA-TDNN white-box surrogate model in generating adversarial examples to attack the black-box ResNetSE34 L model. The EER before the attack was 2.179% with a threshold of 0.051, respectively. After the MI-FGSM attack, the EER increased to 40.2% with a threshold of 0.043. Furthermore, the STA-MDCT attack resulted in an EER of 58.8% with a threshold of 0.038. Fig. 5 also illustrates the impact of the adversarial impersonation on increasing the similarity scores of non-target trials, and the impact of the adversarial evasion on decreasing the similarity scores of target trials. Notably, our proposed STA-MDCT achieves significant attack effectiveness in both of the attack scenarios.

B. Results With an Ensemble of White-Box Surrogate ASV Models

Crafting adversarial examples from an ensemble of surrogate models has been shown to be an effective way of improving the transferability of an attacker. Here we conducted an experimental comparison of the attackers with an ensemble of surrogate models that were selected from the four ASV models. The weights of the selected surrogate models were set equal. From the results in Table IV, we observe that our method achieves the highest EER and TASR in all black-box attack scenarios. For instance, the proposed method with the ensemble of the ResNetSE34 L and ResNetSE34V2 surrogate models achieves an EER of 61.6% on the ECAPA-TDNN victim model, which is absolutely 10% higher than the result of the corresponding single white-box surrogate model in Table II.

C. Ablation Study

In this subsection, we investigate the effects of the hyperparameters of STA-MDCT, including the maximum iterations T, number of spectrum transformation N, standard deviation $\boldsymbol{\sigma}$ of noise $\boldsymbol{\xi}$, and tuning factor $\boldsymbol{\rho}$. For tuning each hyperparameter, we fixed the others to their default values. We crafted adversarial examples from the ECAPA-TDNN white-box surrogate model, and applied them to attack the remaining three black-box victim models. The results are summarized in Fig. 6, where a higher EER signifies improved attack performance, while increased SNR and PESQ values indicate enhanced imperceptibility of the adversarial examples. Detailed analysis is provided below. Note that we also show the result of the white-box attack as a reference.

1) Effect of the Maximum Iterations T: From Fig. 6(a), we see that when $T = 1$, the EER and PESQ produced by the proposed method are far from satisfactory; as T increases, both the transferability and the SNR/PESQ of the adversarial examples improves, with a negative effect of the increased computational cost. To balance the two factors, we set $T = 10$ as the default.

2) Effect of the Spectrum Transformation N: From Fig. 6(b), we see that when $N = 1$, a single spectrum transformation yields the worst transferability; the transferability of the adversarial examples is improved when N increases, and tends to be increased slowly when N exceeds 20. This phenomenon indicates that the proposed spectrum transformation can effectively narrow the gap between the white-box surrogate model and the black-box victim models. To balance the computational cost, we set $N = 20$ in this article.

3) Effect of the Standard Deviation σ: From Fig. 6(c), we see that the larger σ is, the higher the magnitude of the adversarial noise will be. Therefore, the EERs of the black-box victim models first increase dramatically, and then drop slightly; the highest EERs appear around $\sigma = 44$. On the other side, as σ increases, the SNR of the adversarial examples decreases continuously; the

TABLE IV
PERFORMANCE OF THE COMPARISON ATTACKERS THAT USE THE SAME ENSEMBLE OF WHITE-BOX SURROGATE MODELS, ON THE ASV TASK IN TERMS OF EER (%), TASR (%), FAR (%), AND MINDCF

Surrogate Model	Attack	Victim Model: ResNetSE34L				ECAPA-TDNN				ResNetSE34V2				RawNet3						
		EER	TASR	FAR	MinDCF	EER	TASR	FAR	MinDCF	EER	TASR	FAR	MinDCF	EER	TASR	FAR	MinDCF	SNR(dB)	PESQ	L2
ECAPA-TDNN & ResNet34V2	FGSM [28]	21.8	25.0	15.6	0.98	47.4	46.6	36.4	0.99	43.6	43.9	29.4	1.00	18.0	19.8	13.0	0.91	30.12	3.09	1.47
	I-FGSM [48]	43.4	49.8	29.0	1.00	95.8	95.7	95.2	1.00	95.0	92.2	87.4	1.00	48.8	45.4	25.4	1.00	36.37	4.12	0.68
	MI-FGSM [49]	48.2	53.3	31.4	1.00	95.0	95.1	94.6	1.00	92.4	90.2	84.2	1.00	56.2	50.5	33.2	1.00	32.82	3.60	1.06
	NI-FGSM [45]	46.0	52.8	32.6	1.00	93.6	93.0	91.0	1.00	95.0	93.0	89.4	1.00	51.2	47.2	29.0	1.00	33.19	3.60	1.05
	ACG [33]	55.2	56.6	32.4	1.00	**97.2**	**96.8**	**96.2**	1.00	**96.4**	**94.1**	**90.6**	1.00	67.0	61.3	42.8	1.00	31.85	3.49	1.19
	STA-DCT [23]	56.6	58.5	36.0	1.00	89.0	86.1	81.6	1.00	86.8	83.6	74.8	1.00	63.8	58.9	42.6	1.00	34.00	3.88	0.91
	STA-MDCT	**65.0**	**64.3**	**40.8**	1.00	96.0	95.8	95.2	1.00	95.4	93.1	88.8	1.00	**73.6**	**70.5**	**56.0**	1.00	33.28	3.87	0.99
ResNetSE34L & ResNet34V2	FGSM [28]	49.6	51.2	36.6	1.00	12.6	13.8	9.6	0.68	38.0	39.1	25.4	0.99	10.8	11.2	7.6	0.72	30.11	3.09	1.47
	I-FGSM [48]	94.2	90.5	83.4	1.00	29.6	32.4	15.6	0.94	94.6	92.1	87.4	1.00	22.6	25.6	11.8	0.88	36.71	4.14	0.66
	MI-FGSM [49]	94.0	90.9	84.6	1.00	36.2	38.3	21.4	0.98	92.2	89.9	84.0	1.00	29.6	31.4	16.6	0.98	32.87	3.62	1.06
	NI-FGSM [45]	92.6	87.5	78.0	1.00	36.4	39.3	21.6	0.97	95.2	**93.9**	**91.2**	1.00	30.6	32.1	16.8	0.99	32.81	3.58	1.08
	ACG [33]	95.4	89.9	81.8	1.00	46.6	46.2	26.2	0.99	**96.6**	93.5	89.4	1.00	39.2	38.1	20.8	0.98	31.89	3.51	1.19
	STA-DCT [23]	86.2	82.2	71.4	1.00	48.4	48.4	29.6	1.00	85.8	82.1	73.6	1.00	42.6	42.0	24.6	1.00	34.37	3.91	0.87
	STA-MDCT	**95.6**	**93.4**	**88.8**	1.00	**61.6**	**57.6**	**38.8**	1.00	94.6	92.2	87.4	1.00	**53.4**	**49.9**	**31.8**	1.00	33.63	3.91	0.95
ResNetSE34L & ECAPA-TDNN	FGSM [28]	54.6	56.0	39.8	1.00	47.2	46.3	35.8	1.00	17.2	19.3	12.6	0.91	16.8	19.4	13.6	0.90	30.11	3.11	1.47
	I-FGSM [48]	94.6	91.4	84.6	1.00	96.0	95.7	95.4	1.00	39.8	43.4	21.6	0.99	46.4	42.5	22.8	1.00	36.25	4.11	0.69
	MI-FGSM [49]	93.4	89.2	81.4	1.00	94.8	94.1	93.4	1.00	45.0	47.6	27.6	1.00	51.4	47.0	30.2	1.00	32.80	3.60	1.07
	NI-FGSM [45]	93.6	88.6	79.4	1.00	96.2	96.3	**96.2**	1.00	45.4	48.9	26.6	0.99	54.6	51.2	32.4	1.00	32.78	3.56	1.08
	ACG [33]	95.6	89.9	81.8	1.00	**97.4**	**97.1**	**96.2**	1.00	59.6	56.3	32.2	1.00	65.2	59.2	40.6	1.00	31.90	3.51	1.18
	STA-DCT [23]	88.0	83.2	72.6	1.00	87.6	85.5	80.4	1.00	50.4	52.4	30.0	1.00	59.4	53.7	36.6	1.00	34.00	3.87	0.91
	STA-MDCT	**95.8**	**93.7**	**88.9**	1.00	95.6	95.4	94.6	1.00	**65.6**	**60.8**	**38.8**	1.00	**69.8**	**66.8**	**51.6**	1.00	33.34	3.84	0.98
ResNetSE34L & ECAPA-TDNN & ResNet34V2	FGSM [28]	49.4	51.9	36.2	1.00	43.3	41.2	29.4	0.98	39.8	41.4	27.2	1.00	18.8	20.5	14.2	0.91	30.11	3.09	1.47
	I-FGSM [48]	93.0	88.3	79.2	1.00	94.6	94.1	93.0	1.00	93.4	90.6	84.6	1.00	53.4	48.3	29.0	1.00	36.21	4.11	0.69
	MI-FGSM [49]	92.2	87.5	78.2	1.00	93.8	92.9	91.2	1.00	91.2	87.6	80.2	1.00	57.2	52.6	35.8	1.00	32.33	3.56	1.12
	NI-FGSM [45]	92.8	90.2	**84.2**	1.00	90.2	87.7	82.8	1.00	91.2	88.3	81.4	1.00	50.6	46.7	28.2	1.00	33.37	3.62	1.04
	ACG [33]	**94.6**	87.9	78.2	1.00	**96.4**	**95.5**	**94.2**	1.00	**96.0**	**92.5**	**87.6**	1.00	68.2	62.0	42.6	1.00	31.86	3.51	1.19
	STA-DCT [23]	86.8	82.2	70.8	1.00	89.8	88.6	84.6	1.00	87.8	83.3	74.4	1.00	71.6	67.2	51.8	1.00	32.59	3.71	1.09
	STA-MDCT	94.4	**90.6**	83.2	1.00	94.8	93.9	92.4	1.00	93.4	90.8	85.0	1.00	**72.8**	**70.1**	**54.8**	1.00	33.27	3.85	0.99

The result in gray color indicates that it is a white-box attack where the victim model is also the surrogate model. The thresholds θ with respect to the TASR and FAR are listed in Table I.

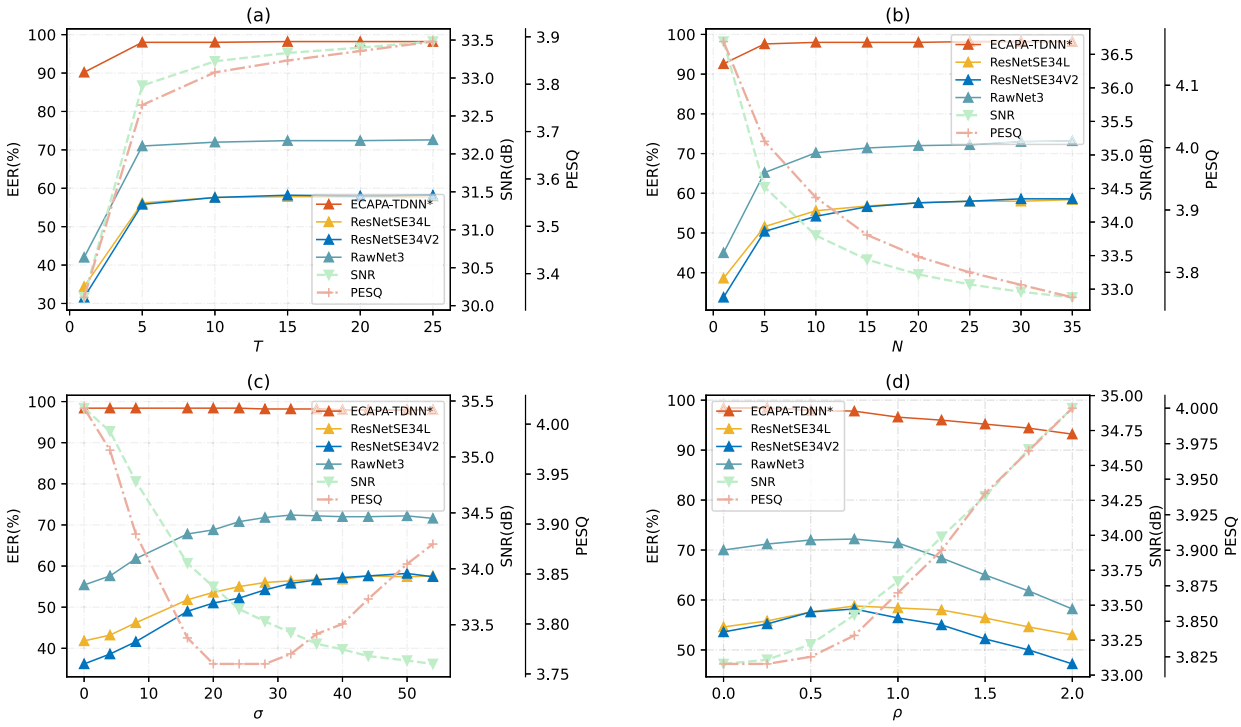


Fig. 6. Effects of the hyperparameters of STA-MDCT on performance in terms of EER (solid line), SNR (dashed line), and PESQ (dotted line). ECAPA-TDNN is used as the white-box surrogate model. The marker "*" indicates the white-box attack.

PESQ first decreases substantially, and then gradually increases. To balance the above three factors, we choose $\sigma = 44$ in this article.

4) Effect of the Tuning Factor ρ: From Fig. 6(d), we see that, as ρ increases, the EER curves of the black-box victim models gradually reach the peak at around $\rho = 0.75$. As ρ continues to increase, the EER curves decrease due to the excessive spectral transformation. On the other side, the SNR and PESQ are increased constantly with the increase of ρ. Consequently, we choose $\rho = 0.75$ in this article.

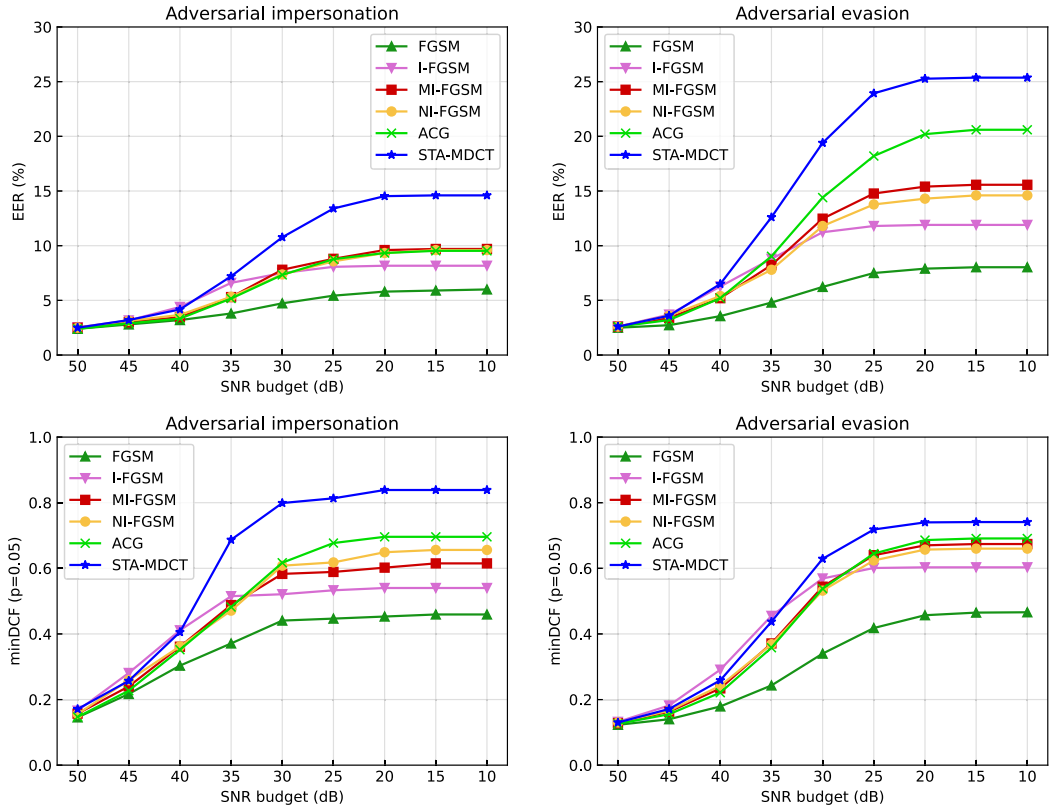


Fig. 7. Performance of the comparison methods with respect to the SNR budget. The black-box victim model is ResNetSE34 L. The white-box surrogate model is ECAPA-TDNN.

D. Effect of the SNR Budget on Performance

All of the above experiments were conducted by setting the perturbation level $\epsilon = 40$, where the SNRs of the adversarial examples were around 33 dB. In this section, we study the attack effect under the situation that the SNR is controlled to be larger than a given threshold b, named the *SNR budget*. To generate a large number of adversarial examples with various SNR levels, we set ϵ to a wide range of $\{5, 10, 20, 30, 40, 50\}$. The white-box surrogate model was ECAPA-TDNN. The black-box victim model was ResNetSE34 L.

To study the effect of the SNR budget b on performance, we count the EER and minDCF statistics of the impersonation attacks and evasion attacks separately as in [16], [65]. Specifically, given an original trial set $\mathcal{O} = \{(\mathbf{x}_i^{\text{enroll}}, \mathbf{x}_i^{\text{test}}) \mid i = 1, 2, \ldots, I\}$ and its corresponding adversarial trial set $\mathcal{A} = \{(\mathbf{x}_i^{\text{enroll}}, \mathbf{x}_i^{\text{adv}}) \mid i = 1, 2, \ldots, I\}$. Suppose $\mathbf{p}_{\text{adv}} = [p_{\text{adv},1}, \ldots, p_{\text{adv},I}]^T$ is a vector describing the SNRs of the adversarial trials. For a given SNR budget b, we can obtain a mixed trial set $\mathcal{M}(b)$ whose elements are defined by:

$$t_i = \begin{cases} (\mathbf{x}_i^{\text{enroll}}, \mathbf{x}_i^{\text{adv}}), & \text{if } p_{\text{adv}},\ i \geq b \text{ and } i \in G \\ (\mathbf{x}_i^{\text{enroll}}, \mathbf{x}_i^{\text{test}}), & \text{otherwise} \end{cases} \quad \forall i = 1, \ldots, I.$$

where G is defined as the set of non-target trials when the task is the impersonation attack to ASV, and defined as the set of target trials when the task is the evasion attack to ASV. Finally, the EER and minDCF are calculated from the mixed trial set $\mathcal{M}(b)$.

Fig. 7 shows the EER and minDCF of the victim model with respect to the SNR budget b, where we have separately summarized the impersonation and evasion attacks. From the figure, we can observe that (i) as the SNR budget b decreases, the EER of the victim model increases for all comparison attackers; (ii) when the SNR budget is below 40 dB, the proposed method achieves a significantly higher EER than the other comparison methods; (iii) when the SNR budget is above 40 dB, the proposed method achieves an EER comparable to I-FGSM, and outperforms the other comparison methods. The experimental phenomena in minDCF are similar to those in EER.

Additionally, we also analyzed the PESQ of the successful attacks of the above adversarial examples. Fig. 8 shows the number of the *successful* adversarial attacks in different PESQ ranges, where the adversarial attacks that were not successful in deceiving the target victim model were discarded. From the figure, we see that the PESQ of the successful adversarial examples generated by the proposed method is higher than all comparison methods, which provides strong evidence that the adversarial perturbations generated by the proposed method are more imperceptible to human.

E. Research on Robustness of Victim Models

In order to study the robustness of the victim models and verify the effectiveness of our proposed method, we perform Specaugment [66] and adversarial training [67] on part of the victim models and attack them in the ASV scenario. The EER

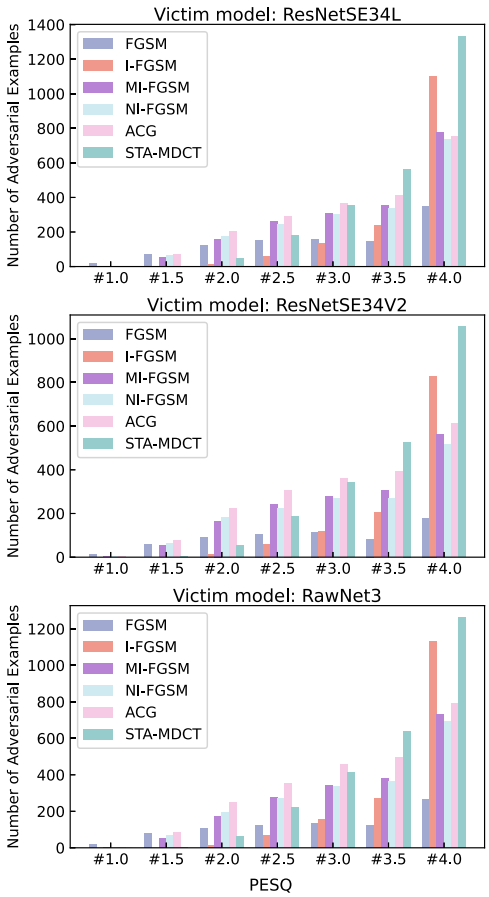


Fig. 8. Histograms of the adversarial examples that can successfully deceive the target victim model in terms of PESQ. The symbol "#i" refers to the range "[i, i+0.5)" in PESQ.

TABLE V
EFFECT OF SPECAUGMENT ON MODEL ROBUSTNESS

EER(%)/TASR(%)	PGD	MI-FGSM	ACG	STA-MDCT
w/o SpecAugment	34.6/38.0	40.2/45.7	47.0/50.2	58.8/60.6
w/ SpecAugment	26.8/30.7	30.6/35.6	35.8/39.2	44.6/52.6

We use adversarial examples generated by ECAPA-TDNN to attack the ResNetSE34L before and after specaugment. 'w/o' stands for without, and 'w/' stands for with.

TABLE VI
EFFECT OF ADVERSARIAL TRAINING (I-FGSM ADVERSARIAL TRAINING) ON MODEL ROBUSTNESS

EER(%)/TASR(%)	PGD	MI-FGSM	ACG	STA-MDCT
w/o Adversarial training	16.6/20.7	22.6/26.3	31.8/32.5	44.0/44.2
w/ Adversarial training	6.4/6.1	7.0/6.8	8.0/7.7	12.0/12.3

We use adversarial examples generated with ResNetSE34L to attack the ECAPA-TDNN model both before and after adversarial training.

before and after specaugment of ResNetSE34 L was 4.242% and 2.179%, respectively. The EER before and after ECAPA-TDNN adversarial training was 1.172% and 3.5%, respectively. We used I-FGSM adversarial training. Each mini-batch carried out an iteration of adversarial examples, during which adversarial examples were generated and used to update model parameters.

TABLE VII
PERFORMANCE OF THE COMPARISON ATTACKERS ON THE CSI TASK IN TERMS OF IER (%) AND TASR (%), WHERE THE AVERAGE SNR OF ADVERSARIAL EXAMPLES IS 33.5 DB

Target	Attack	Victim Model								PESQ
		ECAPATDNN*		ResNetSE4L		ResNetSE34V2		RawNet3		
		IER	TASR	IER	TASR	IER	TASR	IER	TASR	
Random	FGSM [28]	5.7	4.7	1.9	0.6	0.1	0.1	90.0	9.4	2.74
	I-FGSM [48]	**100.0**	**100.0**	9.9	8.2	1.4	1.4	90.0	9.2	3.07
	MI-FGSM [49]	99.8	99.8	9.6	7.7	2.0	2.0	90.0	**11.0**	2.94
	NI-FGSM [45]	**100.0**	**100.0**	9.4	7.6	1.3	1.3	90.0	9.6	2.91
	ACG [33]	99.9	99.9	7.0	5.8	0.8	0.7	**92.2**	10.3	2.98
	STA-MDCT	99.0	98.9	**39.1**	**27.7**	**18.8**	**17.8**	90.0	10.1	3.16
Farthest	FGSM [28]	1.3	0.2	1.7	0.1	0.0	0.0	90.0	7.6	2.74
	I-FGSM [48]	**100.0**	**100.0**	4.0	1.2	0.1	0.1	90.0	**16.0**	3.06
	MI-FGSM [49]	**100.0**	**100.0**	5.5	1.4	0.1	0.1	90.0	7.6	2.93
	NI-FGSM [45]	**100.0**	**100.0**	4.0	0.8	0.1	0.1	90.0	8.1	2.91
	ACG [33]	99.9	99.9	3.7	0.8	0.1	0.1	90.0	8.0	2.98
	STA-MDCT	99.0	98.7	**31.1**	**12.8**	**8.7**	**7.3**	90.0	8.2	3.16

The marker "*" indicates the white-box attack.

TABLE VIII
PERFORMANCE OF THE COMPARISON ATTACKERS ON THE OSI TASK IN TERMS OF IER (%) AND TASR (%), WHERE THE AVERAGE SNR OF ADVERSARIAL EXAMPLES IS 33.5 DB

Target	Attack	Victim Model								PESQ
		ECAPATDNN*		ResNetSE4L		ResNetSE34V2		RawNet3		
		IER	TASR	IER	TASR	IER	TASR	IER	TASR	
Random	FGSM [28]	19.5	18.1	4.9	2.1	1.5	1.0	0.0	0.0	3.06
	I-FGSM [48]	**100.0**	**100.0**	14.4	9.2	7.4	6.0	0.0	0.0	3.29
	MI-FGSM [49]	99.4	99.4	11.6	7.3	6.5	4.6	0.0	0.0	3.32
	NI-FGSM [45]	**100.0**	**100.0**	13.1	7.9	6.4	5.0	0.0	0.0	3.2
	ACG [33]	99.9	99.9	12.5	6.9	6.3	4.5	0.0	0.0	3.26
	STA-MDCT	99.5	99.5	**23.2**	**19.0**	**16.7**	**15.8**	0.0	0.0	3.35
Farthest	FGSM [28]	2.0	0.0	3.5	0.0	0.9	0.0	0.0	0.0	3.06
	I-FGSM [48]	**100.0**	**100.0**	7.2	0.2	2.4	0.0	0.0	0.0	3.28
	MI-FGSM [49]	99.5	99.5	5.9	0.1	1.7	0.0	0.0	0.0	3.30
	NI-FGSM [45]	**100.0**	**100.0**	6.0	0.1	2.4	0.0	0.0	0.0	3.21
	ACG [33]	**100.0**	**100.0**	**7.5**	0.2	**2.9**	0.0	0.0	0.0	3.27
	STA-MDCT	98.3	98.3	7.4	**4.0**	1.7	**0.3**	0.0	0.0	3.33

The marker "*" indicates the white-box attack.

The experiments in Tables V and VI indicate that specaugment and adversarial training can significantly enhance model robustness. Furthermore, our algorithm consistently outperforms other baselines.

VIII. RESULTS OF ADVERSARIAL ATTACKS TO SPEAKER IDENTIFICATION

Tables VII and VIII list the performance of the comparison attackers in the CSI and OSI scenarios, respectively, where we combine each attacker with the single white-box surrogate model ECAPA-TDNN, and then apply the generated adversarial examples to the other three black-box victim systems. From the tables, we see that, when the SNR budget is controlled to be the same around 33.5 dB, the proposed STA-MDCT achieves better attack performance in terms of IER and TASR, as well as higher PESQ than the comparison methods. It's worth noting that the farthest target attack proves to be quite challenging. In addition, the black-box attack performance of all comparison methods is generally poor. It may be caused by the following reasons. First, the dataset used to attack the black-box victim models is different from that of the white-box surrogate model. Second, attacking speaker identification refers to many enrolled speaker identities, making the attack more challenging than attacking ASV. Besides, compared to attacking a CSI system, attacking an OSI system not only needs to maximize the confidence score for determining an adversarial voice to a target label but also has to make the score exceed the predefined threshold θ.

IX. Conclusion

In this article, we propose a spectrum transformation attack method based on a modified discrete cosine transform. It first applies MDCT to the input voices and then slightly modifies the energy of the frequency bands of the transformed voices in the time-frequency domain for capturing the salient regions of the adversarial noise that are critical to a successful attack. Different from existing transfer-based attackers, STA-MDCT generates adversarial examples in the time-frequency domain, which improves the transferability and efficiency of the adversarial attack. Moreover, we also interpret the effectiveness of transfer-based attacks by Layer-CAM. The transferability of transfer-based attackers is observable directly from the critical attention regions of saliency maps. The comprehensive experiments on the ASV, OSI, and CSI tasks demonstrate the effectiveness of the proposed method.

References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1/3, pp. 19–41, 2000.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.

[4] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, 2021.

[5] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 1559–1562.

[6] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4052–4056.

[7] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, 2015.

[8] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," 2020, *arXiv:2004.08849* .

[9] Z. Wu et al., "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 588–604, Jun. 2017.

[10] S. Cui, B. Huang, J. Huang, and X. Kang, "Synthetic speech detection based on local autoregression and variance statistics," *IEEE Signal Process. Lett.*, vol. 29, pp. 1462–1466, 2022 .

[11] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6359–6363.

[12] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6369–6373.

[13] T. Zhai, Y. Li, Z. Zhang, B. Wu, Y. Jiang, and S.-T. Xia, "Backdoor attack against speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 2560–2564.

[14] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 11957–11965.

[15] T. Xu, Y. Li, Y. Jiang, and S.-T. Xia, "Batt: Backdoor attack with transformation-based triggers," 2022, *arXiv:2211.01806*.

[16] J. Villalba, Y. Zhang, and N. Dehak, "x-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," in *Proc. Interspeech*, 2020, pp. 4233–4237.

[17] L. Zhang, Y. Meng, J. Yu, C. Xiang, B. Falk, and H. Zhu, "Voiceprint mimicry attack towards speaker verification system in smart home," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2020, pp. 377–386.

[18] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, 2017, pp. 15–26.

[19] G. Chen et al., "Who is real Bob? Adversarial attacks on speaker recognition systems," in *Proc. IEEE Symp. Secur. Privacy*, 2021, pp. 694–711.

[20] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2137–2146.

[21] J. Zhang et al., "NMI-FGSM-TRI: An efficient and targeted method for generating adversarial examples for speaker recognition," in *Proc. 7th IEEE Int. Conf. Data Sci. Cyberspace*, 2022, pp. 167–174.

[22] Y. Zhang, Z. Jiang, J. Villalba, and N. Dehak, "Black-box attacks on spoofing countermeasures using transferability of adversarial examples," in *Proc. Interspeech*, 2020, pp. 4238–4242.

[23] Y. Long et al., "Frequency domain model augmentation for adversarial attack," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 549–566.

[24] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1962–1966.

[25] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM I-vector based speaker verification systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6579–6583.

[26] A. Jati, C.-C. Hsu, M. Pal, R. Peri, W. AbdAlmageed, and S. Narayanan, "Adversarial attack and defense strategies for deep speaker recognition systems," *Comput. Speech Lang.*, vol. 68, 2021, Art. no. 101199.

[27] P. Li, L. Li, A. Hamdulla, and D. Wang, "Reliable visualization for deep speaker recognition," 2022, *arXiv:2204.03852*.

[28] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015 .

[29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.

[30] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Boca Raton, FL, USA: CRC Press, 2018, pp. 99–112.

[31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[32] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[33] K. Yamamura et al., "Diversified adversarial attacks based on conjugate gradient method," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 24872–24894.

[34] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1765–1773.

[35] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Real-time, robust and adaptive universal adversarial attacks against speaker recognition systems," *J. Signal Process. Syst.*, vol. 93, no. 10, pp. 1187–1200, 2021.

[36] J. Li et al., "Universal adversarial perturbations generative network for speaker recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2020, pp. 1–6.

[37] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Enabling fast and universal audio adversarial attack using generative model," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 14129–14137.

[38] J. Yao, X. Chen, X.-L. Zhang, W.-Q. Zhang, and K. Yang, "Symmetric saliency-based adversarial attack to speaker identification," *IEEE Signal Process. Lett.*, vol. 30, pp. 1–5, 2023.

[39] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM Sigsac Conf. Comput. Commun. Secur.*, 2016, pp. 1528–1540.

[40] W. Brendel, J. Rauber, M. Bethge, and D.-B. Adversarial, "Decision-basedadversarialattacks: Reliableattacksagainstblack-box machine learning models," *Adv. Reliably Evaluating Improving Adversarial Robustness*, 2021, Art. no. 77.

[41] M. Cheng, T. Le, P.-Y. Chen, H. Zhang, J. Yi, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," in *Proc. Int. Conf. Learn. Representations*, 2018.

[42] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in *Proc. IEEE Symp. Secur. Privacy*, 2020, pp. 1277–1294.

[43] Z. Yuan, J. Zhang, Y. Jia, C. Tan, T. Xue, and S. Shan, "Meta gradient adversarial attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7748–7757.

[44] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence o $(1/k^2)$," *Doklady an USSR*, vol. 269, pp. 543–547, 1983.
[45] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. Int. Conf. Learn. Representations*, 2019.
[46] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4312–4321.
[47] Y. Zhu et al., "Toward understanding and boosting adversarial transferability from a distribution perspective," *IEEE Trans. Image Process.*, vol. 31, pp. 6487–6501, 2022.
[48] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Representations*, 2017.
[49] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9185–9193.
[50] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8684–8694.
[51] S. Zhang, W. Dou, and H. Yang, "MDCT sinusoidal analysis for audio signals analysis and processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1403–1414, Jul. 2013.
[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[53] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.
[54] H. Wang et al., "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 24–25.
[55] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021.
[56] J. S. Chung et al., "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020.
[57] Y. Kwon, H. S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: Lessons from VoxSRC 2020," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021.
[58] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020.
[59] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," 2017, *arXiv:1706.08612*.
[60] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.
[61] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Proc.*, 2001, vol. 2, pp. 749–752.
[62] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022.
[63] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
[64] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proc. Interspeech*, 2019, pp. 2873–2877.
[65] Y. Dong et al., "Benchmarking adversarial robustness on image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 321–331.
[66] D. S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.
[67] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," 2021, *arXiv:2102.01356*.

Jiadi Yao received the B.S. degree in electronic science and technology from the North University of China, Taiyuan, China. She is currently working toward the M.S. degree in electronic information from Northwestern Polytechnical University, Xi'an, China. Her research interests include speaker recognition and adversarial attack.

Hong Luo received the master degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2002. She is currently the General Manager of the Home Screen-related Product Department, China Mobile (Hangzhou) Information Technology Company, Ltd., Hangzhou, China. She is active in global and domestic standardization work with 28 authorized patents. Her research interests include mobile communication services, digital home, and artificial intelligence. She is a Member of The Chinese Institute of Electronics, Beijing and China Institute of Communications, Beijing.

Jun Qi received the master degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 2013, the second master degree from Tsinghua University, Beijing, China, in 2017, and the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2022, advised by Prof. Chin-Hui Lee and Prof. Xiaoli Ma. He is currently a Research Assistant Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. He was a Research Intern with Deep Learning Technology Center, Microsoft Research, Redmond, WA, USA, Tencent Al Lab, Bellevue, WA, USA, and MERL, Cambridge, MA, USA. Dr. Qi was the recipient of the first prize in the Xanadu Al Quantum Machine Learning Competition 2019. His ICASSP paper on quantum speech recognition was nominated as the best paper candidate in 2022. Besides, as a Special Session Chair, he organized two special sessions on quantum machine learning for machine learning and signal processing at the venues of ICASSP'23 and ICASSP'24. He also gave three Tutorials on Quantum Neural Networks for Speech and Language Processing at the venues of IJCAI'21, ICASSP'22, and ICASSP'23."

Xiao-Lei Zhang (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Full Professor with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China. He was a Postdoctoral Researcher with the Perception and Neurodynamics Laboratory, The Ohio State University, Columbus, OH, USA. His research interests include speech processing, underwater acoustic signal processing, machine learning, statistical signal processing, and artificial intelligence. He is a Senior Member of IEEE and a Member of IEEE, SPS, and ISCA.