Hindawi
Security and Communication Networks
Volume 2022, Article ID 2666534, 12 pages
https://doi.org/10.1155/2022/2666534

Research Article

Towards Generating Adversarial Examples on Combined Systems of Automatic Speaker Verification and Spoofing Countermeasure

Xingyu Zhang, Xiongwei Zhang, Xia Zou, Haibo Liu, and Meng Sun

Army Engineering University, Nanjing, China

Correspondence should be addressed to Xiongwei Zhang; xwzhang9898@163.com and Xia Zou; zlc1997@163.com

Received 1 April 2022; Accepted 12 July 2022; Published 31 July 2022

Academic Editor: AnMin Fu

The security of unprotected automatic speaker verification (ASV) system is vulnerable to a variety of spoofing attacks where an attacker (adversary) disguises him/herself as a specific targeted user. It is a common practice to use spoofing countermeasure (CM) to improve the security of ASV systems so as to avoid illegal access. However, recent studies have shown that both ASV and CM systems are vulnerable to adversarial attacks. Previous researches mainly focus on adversarial attacks on a single ASV or CM system. But in practical scenarios, ASVs are typically deployed in conjunction with CM. In this paper, we investigate attacking the tandem system of ASV and CM with adversarial examples. The joint objective function is designed to restrict the generating process of adversarial examples. The joint gradient of the ASV and CM system is derived to generate adversarial examples. Fast Gradient Sign Method (FSGM) and Projected Gradient Descent (PGD) are utilized to study the vulnerability of tandem verification systems against white-box adversarial attacks. Through our attack, audio samples whose original labels are spoof or nontarget can be successfully accepted by the tandem system. Experimental results on the ASVSpoof2019 dataset show that the tandem system is vulnerable to our proposed attack.

1. Introduction

Automatic speaker verification (ASV) aims to extract features from given utterances so as to determine whether the utterance belongs to a specific speaker. ASV is undisputedly a crucial technology for biometric identification, which is broadly applied in real-world applications like access control, military, judicial forensics, and surveillance [1]. However, unprotected ASV systems are vulnerable to a variety of spoofing attacks [2]. In spoofing attacks, the attacker usually disguises himself/herself as one of the enrolled speakers by generating spoofing speech [3, 4]. The emergence of spoofing attacks promotes the research of spoofing countermeasures (CM). Whether being independent of ASV or combined with ASV, spoofing countermeasure has become an indispensable part when deploying ASV [5]. In recent years, by following ASVSpoof challenges, the works to address voice spoofing attacks and their defenses have become popular [6–10]. In view of various spoofing scenarios, researchers have proposed lots of effective antispoofing methods [11–15]. The scenarios of both logical access (LA) and physical access (PA) are taken into account in these works. The LA scenario involves fake audios synthesized by modern text-to-speech synthesis (TTS) and voice conversion (VC) models. The PA scenario involves replayed audio signals recorded in reverberant environments under different acoustic configurations. Several teams in ASVSpoof2019 have achieved excellent performance in detecting spoofing and reinforcing robustness of ASV systems under both LA and PA scenarios. Therefore, with the rapid development of spoofing detection, it is becoming common to deploy ASV and CM together.

Adversarial attacks have potential threats to all types of machine learning models [16–19], so they have attracted a lot of attention in different classification tasks. According to whether the attacker has the internal information of ASV

(including model structure, parameters, loss function, and gradient information), adversarial attacks can be divided into white-box attack and black-box attack [20]. In general, white-box attacks have a higher success rate, but black-box attacks are more in line with realistic attack scenarios. In recent years, preliminary progress has been made in adversarial attacks on ASV and CM. Researchers conducted white-box attacks [21–27] or black-box attacks [26–30] on common ASV models. In [31, 32], the vulnerability of the CM system against adversarial examples is also investigated.

Although there have been various works in the field of adversarial attacks on ASV or CM, as far as we know, adversarial attack research on the tandem system of ASV and CM has not yet appeared [2]. As ASV is usually utilized in combination with CM in real scenarios, it is necessary to study the adversarial attack in this kind of tandem system. In the tandem system, ASV and CM systems are trained independently and combined during the validation phase. In order to measure the performance of the tandem system, the tandem detection cost function (t-DCF) is proposed in [33, 34]. The calculation of t-DCF utilizes different kinds of errors generated by two subsystems and assigns different costs to these errors. In this paper, our goal is to enable utterances that should have been rejected by the tandem system to be accepted after an adversarial attack. Because there are two independent subsystems in the tandem system, it is necessary to consider the gradient and loss of subsystems in the generation of the adversarial examples so that adversarial utterance can deceive both ASV and CM systems.

In this paper, we implement the tandem verification system of ASV and CMs on the ASVSpoof2019 Challenge dataset. The method of attacking the tandem system is also proposed. To the best of our knowledge, it is the first work to study adversarial attacks in the tandem system of ASV and CM. Our contribution is as follows:

(1) For the tandem system of ASV and CM, a parallel branch structure is designed to derive the joint target function.

(2) The joint adversarial gradient derived from the joint target function is utilized to generate adversarial examples.

(3) Compared with the step-by-step attack, the joint adversarial attack method proposed by us is more effective.

The remaining part of the paper is organized as follows. The related works about adversarial attacks on ASV and CM are introduced in Section 2. The models of ASV and CM and their combination are introduced in Section 3. The algorithm for generating adversarial examples in the tandem system is proposed in Section 4. The settings and results of experiments are reported in Section 5. The summary and discussion are given in Section 6.

2. Related Works

In this section, we introduce the preliminaries of adversarial attacks on ASV and CM, respectively.

2.1. Adversarial Attacks on ASV. Study [26] shows that end-to-end ASV systems are vulnerable to adversarial attacks. Adversarial examples are generated by adding a perceptually indistinguishable structured noise to the original test examples. This is the first work in the field of ASV adversarial examples. Fast Gradient Sign Method (FGSM) is utilized to carry out white-box and black-box attacks in a cross-corpora and cross-feature setting. Another recent study [22] investigates the vulnerability of the Gaussian Mixture Model (GMM) i-vector-based ASV under adversarial attacks. The transferability of adversarial examples from one ASV to another is also evaluated in this work. "FakeBob" addressed in [30] investigates the impacts of threats generated by practical black-box attacks. This study considers different cases for practical scenarios, including various ASV architectures of commercial systems, transferability of attacks, practicality of over-the-air through replay, and imperceptibility based on human perception. Further studies have also explored real-time, practical, and robust adversarial attacks. The estimated room impulse response (RIR) is integrated into the adversarial example training process [25, 28].

2.2. Adversarial Attacks on CM. Unlike the adversarial attack research that has been widely explored on the ASV systems, adversarial attacks on spoofing countermeasures have received little attention. A recent work [32] investigates the vulnerability of spoofing countermeasures for ASV under both white-box and black-box attacks with the FGSM and the Projected Gradient Descent (PGD) methods. The performance of black-box attacks across spoofing countermeasure models with different network architectures and different amount of model parameters is compared in this work. It reveals that spoofing countermeasure models are vulnerable to FGSM and PGD attacks under the scenario of white-box attack. The black-box attacks are also proved to be effective. In addition to the work in [32], the work in [31] has also proposed a black-box attack utilizing the transferability of adversarial examples.

3. Two Subsystems and Their Combination

In this section, the subsystems ASV and CM are introduced. Method of their combination is also introduced.

3.1. The Tandem System as the Attack Victim. Both ASV and CM systems belong to binary classification systems [34]. Each trial of ASV is an enrollment-test pair, where u_e is collected at the enrollment phase and u_t at the verification phase. If in pair (u_e, u_t) the identities of speakers are the same, it is known as a target trial; otherwise, it is a nontarget trial. Therefore,

$$\begin{cases} H_0^{asv}\,(\text{nontarget}):\ id(u_e) \neq id(u_t), \\ H_1^{asv}\,(\text{target}):\ id(u_e) = id(u_t), \end{cases} \tag{1}$$

where $id(u) \in \mathbb{N}$ represents the speaker identity corresponding to utterance u. ASV systems may encounter the intrusion of spoofed trials. Therefore, it is necessary to

deploy the CM system to reject spoofed utterance. The object of the CM system is to verify the authenticity of test utterance u_t. If u_t corresponds to genuine speech produced by real human speaker, the trial is referred to as a bonafide trial. If u_t corresponds to nongenuine, manipulated, or synthesized speech, the trial is referred to as a spoof trial. Therefore,

$$\begin{cases} H_0^{cm}\,(\text{spoof}):\ u_t \text{ is as poofing utterance,} \\ H_1^{cm}\,(\text{bonafide}):\ u_t \text{ is a bonafide utterance.} \end{cases} \tag{2}$$

Although both ASV and CM systems have the same object of preventing illegal access, they each have specific goals. ASV system should be able to reject zero-effort imposters (nontarget speakers), and the CM system should be able to detect spoofing speakers. The ASV and CM systems play complementary roles, and both are needed to ensure spoofing-robust ASV.

Traditional fusion systems typically involve two subsystems with the same objectives, such as the fusion of two ASV systems or two CM systems. However, ASV and CM do not have the same objective function, so the ASV-CM tandem system is different from the traditional fusion system. The real target speakers' trials should be accepted by both ASV and CM. Cascaded tandem detection framework shown in Figure 1 has shown the potential in previous work [34]. Therefore, the cascaded system shown in Figure 1 is chosen as the victim system in this paper. Obviously, it is needed to set thresholds for both CM and ASV modules. The final decision result will be obtained after comparing scores with two thresholds (i.e., thresholds of CM and ASV subsystems). Trials can be accepted only when scores are not less than both thresholds. For tandem systems, three different types of trials will be encountered: (i) target, (ii) nontarget, and (iii) spoof. There are two final decisions for the tandem system: (i) accept and (ii) reject. Only trials labeled target should be accepted, and both nontarget and spoof trials should be rejected.

3.2. The Model Details of ASV and CM Subsystems. The deep neural network structure for ASV is presented in Figure 2, which follows the architecture utilized in [35, 36]. The mutual-information maximization method is utilized for training the Siamese network. All training procedures only update the back-end of ASV (Siamese and discriminator module), while the front-end feature extraction module remains fixed. The green squares in Figure 2 are "Siamese" module, and the orange squares are "discriminator" module. The traditional cosine metric or PLDA scoring at the back-end is not utilized in this modified structure. Instead, in order to obtain the gradient, a fully connected layer is utilized to measure the similarity of speech features. If both inputs x_{enrol} and x_{test} belong to the same speaker, the score is 1; otherwise, it is 0.

The Squeeze-Excitation Network (SENet) structure for CM is presented in Figure 3 and Table 1, which follows the SENet34 architecture proposed in [13]. The system proposed in [13] is ranked 3rd and 14th places for the PA and LA scenarios, respectively. SENet adaptively recalibrates the channel feature responses by explicitly modelling the dependencies between channels, which has shown great advantages in image classification tasks [37]. The use of SENet has also achieved excellent results in the field of CM.

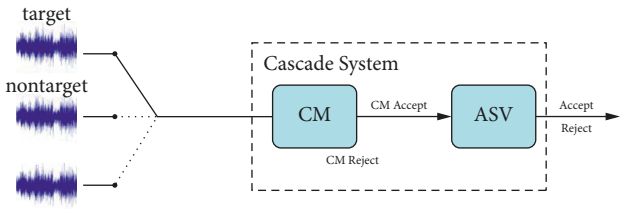


Figure 1: Integration system of ASV and CM.

4. Tandem Attack Methods

4.1. Adversarial Attack Methods. Given audio sample x, the goal of the attack is to generate a perturbed audio signal:

$$\begin{aligned} &\max_{\|\delta\|_p \le \varepsilon} L(\theta, \widetilde{x}, l) \\ &\widetilde{x} = x + \delta \; s.t. \; \|\delta\|_p \le \varepsilon, \end{aligned} \tag{3}$$

where θ is the parameters of the model that has been fixed, $\widetilde{x}$ is the perturbed audio, l is the original label of x, L is the loss function, and ε is the upper limit of perturbation. The goal of the adversarial attack is to lead the classifier to misclassify $\widetilde{x}$. If the real label of audio sample x is l, then after the adversarial attack, the classifier will identify the label of $\widetilde{x}$ as $\widetilde{l}$, and $\widetilde{l} \neq l$. In an adversarial attack, it is necessary to ensure that the perturbation is imperceptible enough so as to make it difficult for humans to distinguish between $\widetilde{x}$ and x. The value of p is generally 2 or ∞, and $p = \infty$ in this work. In order to solve the above optimization problems, Fast Gradient Sign Method (FGSM) [17] and Projected Gradient Descent (PGD) [38] are utilized in our paper.

(1) *FGSM.* FGSM is a single-step attack method with high computational efficiency. The main idea is to extract the sign of the gradient function to generate adversarial examples. Loss will increase by moving along the gradient direction. The perturbed signal generated by FGSM is as follows:

$$\widetilde{x} = x + \varepsilon \cdot \text{sign}\left(\nabla_x L(\theta, x, l)\right). \tag{4}$$

(2) *PGD.* PGD is a method of generating adversarial examples through iteration. The attack success rate of PGD is higher than FGSM, but it also consumes more computing resources. First, initialize $x_0 = x$, and then the audio after each iteration is

$$\widehat{x}_{k+1} = \text{clip}\left(x_k + \alpha \cdot \text{sign}\left(\nabla_{x_k} L(\theta, x_k, l)\right)\right), \tag{5}$$

where $0 \le k \le K$ and K is the maximum number of iterations. α is the step-size of the gradient descent update. The function clip is utilized to clip the perturbation to satisfy $0 \le \|x_K - x\|_\infty \le \varepsilon$.

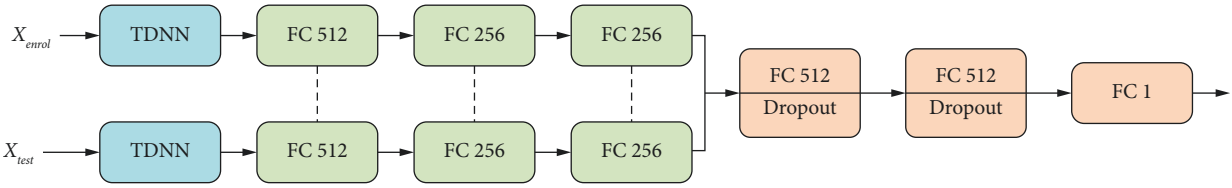


Figure 2: The neural network architecture for ASV. Dashed lines indicate shared parameters. All hidden layers use ReLU activation. Green boxes represent "Siamese" modules and orange "discriminator" modules.

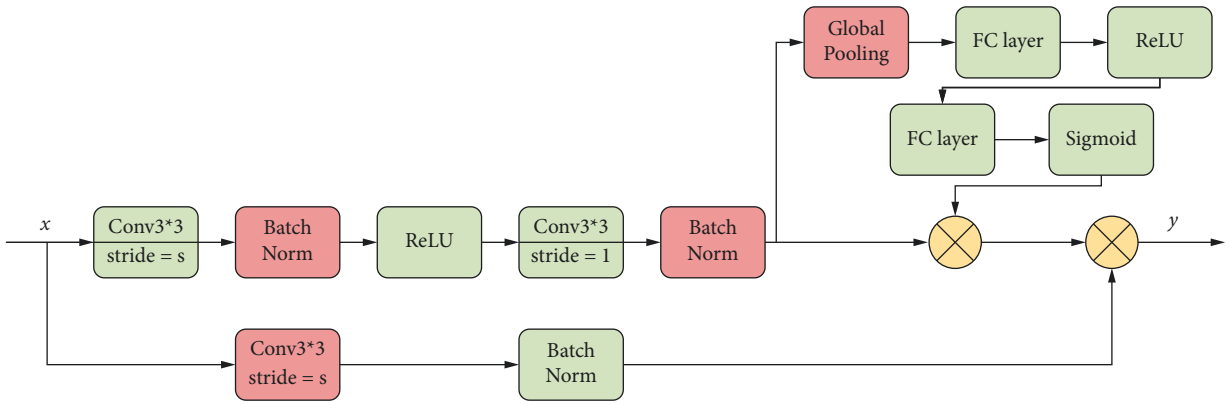


Figure 3: The network module of Squeeze-Excitation Network (SENet), where (s) is the stride control variable, (x) is the input, and (y) is the output.

Table 1: The architecture of SENet34.

Type	Filter/stride	Output
Conv	7*7/2*2	431*300*16
Batch norm	—	431*300*16
ReLU	—	431*300*16
Max pool	3*3/2*2	215*150*16
SEResNet module *3	—	215*150*16
SEResNet module *4	—	107*75*32
SEResNet module *6	—	53*37*64
SEResNet module *3	—	26*18*128
Global AvgPool	—	128
FC	—	2

4.2. Adversarial Attack in Tandem System. Previous work on adversarial attacks has been done in ASV or CM subsystems, respectively. In this paper, adversarial attacks are conducted against the CM-ASV tandem system. We propose that the joint gradient is utilized to generate adversarial examples of the tandem system so that ASV and CM systems can be deceived simultaneously. In order to derive the joint objective function utilized to generate adversarial examples, a parallel decision structure is proposed, as shown in Figure 4. In fact, there are two methods to combine CM and ASV, including a cascade one and a parallel one, whose decision principles are similar to that of victim systems in Figure 1. The tandem system accepts the input utterance only if both ASV and CM systems accept it. Therefore, our design can be utilized not only for parallel systems but also for cascaded systems. The input features for both ASV and CM in our algorithm are unified, which simplifies the computation of joint gradients regarding features. The CM and ASV

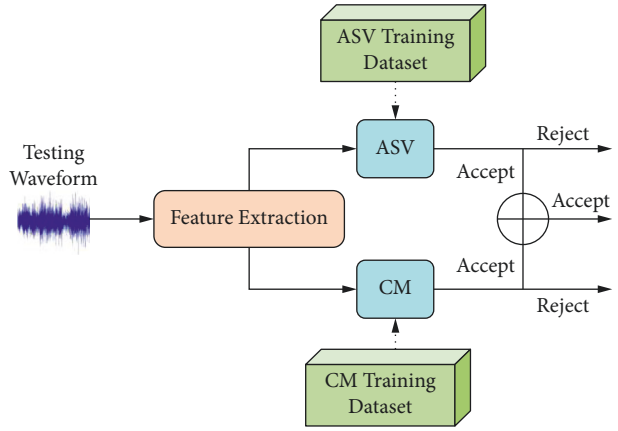


Figure 4: Parallel decision architecture. Dotted lines represent the flow of training information.

subsystems adopt the network structures introduced in Figures 2 and 3. In order to ensure the additivity of adversarial gradients generated by ASV and CM, the input features of subsystems are unified to Log Power Spectrum (LPS).

In FGSM and PGD, both the loss function and gradient information of the target system need to be obtained. The simplest method is to add perturbation to the original utterance against CM and then ASV (or swap the order). However, the perturbation added later will override the perturbation added earlier, making it impossible to deceive both systems at the same time. The analysis will be shown in Section 4. For these reasons, the joint loss function is introduced:

$$L_{\text{total}} = L_{\text{ASV}} + L_{\text{CM}}, \tag{6}$$

$$= \alpha CE\left(l_{asv}, f\left(x\right)\right) + \beta CE\left(l_{cm}, g\left(x\right)\right), \tag{7}$$

$$= \alpha\left[\sum_{i=tar} l_{asv}^{i} \cdot \log f_i\left(x\right) + \sum_{i\neq tar}\left(1 - l_{asv}^{i}\right) \cdot \log\left(1 - f_i\left(x\right)\right)\right], \tag{8}$$

$$+\beta\left[\sum_{i=\text{bon}} l_{cm}^{i} \cdot \log g_i\left(x\right) + \sum_{i\neq \text{bon}}\left(1 - l_{cm}^{i}\right) \cdot \log\left(1 - g_i\left(x\right)\right)\right], \tag{9}$$

where L_{ASV}, L_{CM}, and L_{total} are loss functions of ASV, CM, and tandem system. CE is the cross-entropy loss function. l_{asv} (target = 1; nontarget = 0) and l_{cm} (bonafide = 1; spoof = 0) are labels of ASV and CM. $f\left(x\right)$ and $g\left(x\right)$ are models of ASV and CM. α and β are weights of ASV and CM; in this paper, $\alpha = \beta = 0.5$, and $\alpha + \beta = 1$. L_{ASV} and L_{CM} are obtained by the subsystems shown in Figures 2 and 3, respectively. During the generation of adversarial examples, inputs are original utterances, labels for ASV and CM, and claimed identity. Since L_{total} is a function of x, l_{asv}, and l_{cm}, it can be represented as $F\left(x, l_{asv}, l_{cm}\right)$. The gradient of the joint loss function can be calculated as follows:

$$\begin{aligned}\frac{\partial F\left(x, l_{asv}, l_{cm}\right)}{\partial x} &= \alpha \cdot \frac{\partial F\left(x, l_{asv}, l_{cm}\right)}{\partial f\left(x\right)} \cdot \frac{\partial f\left(x\right)}{x} \\ &+ \beta \cdot \frac{\partial F\left(x, l_{asv}, l_{cm}\right)}{\partial g\left(x\right)} \cdot \frac{\partial g\left(x\right)}{x}.\end{aligned} \tag{10}$$

Substituting formulas (6)–(10) into the FGSM or PGD algorithm, the adversarial examples of the tandem system can be obtained.

5. Experiments

5.1. Datasets and Metrics. This work utilizes the ASVspoof2019 dataset, which encompasses partitions for the assessment of LA and PA scenarios. LA implies a scenario in which a remote user seeks access to a system or service protected by ASV. An example is a telephone banking service. In this scenario, attackers may connect and then send synthetic or converted voice signals directly to the ASV system while bypassing the microphone, that is, by injecting audio into the communication channel. Attacks in the LA scenario can be generated using the latest TTS and VC technologies. The best of these algorithms produces speech that is perceptually indistinguishable from bona fide speech. In the PA scenario, spoofing attacks are presented to a fixed microphone which is placed in an environment where sounds propagate and are reflected from obstacles such as floors and walls. Implementing a replay spoofing attack requires recording bonafide speech in advance and then playing those recordings back to the microphone of the ASV system with a replay device. In this paper, we only utilize the LA partition. The dataset provides spoofing samples generated by different spoofing methods, as well as labels of speaker and spoofing method [9, 10].

The ASVSpoof2019 dataset is utilized to train the CM system and evaluate the experimental results. The structure of CM has shown in Figure 3. When training the CM system, LPS are extracted according to the speaker list of ASVSpoof2019. There are 25,380 utterances in the training set for training the CM model. There are also 24,844 utterances for development and 71,237 for evaluation. The Blackman window function is utilized to extract LPS with a length of 1724 as features, with a window length of 0.0081s [14]. During the training phase, each training sample consists of a feature and a {0,1} target. If the utterance comes from an imposter, the target is 0; otherwise, it is 1. The network is updated with minibatches of size 64. The maximum iteration round is 100. The training early stops when the classification accuracy rate on the development set does not increase more than 5 iterations. During the training phase, the network is updated with parameters through the softmax and cross-entropy loss functions utilizing Adam. The training batch size is 64, and the weight decay rate is 0.001. It is worth mentioning that, in this paper, adversarial examples are added to the feature domain. Since attackers do not always have access to the feature input interface of models, it is necessary to utilize waveform to attack. During the test phase, the reconstructed adversarial waveforms are utilized to attack the tandem system. The adversarial utterances are reconstructed by combining the phase of the original spectrum with the amplitude of the adversarial spectrum, which is a standard adversarial waveform reconstructed approach when the adversarial attack algorithm is implemented on frequency domain as reported in [22, 23, 31, 32, 39].

VoxCeleb1 is utilized to pretrain the ASV. When training the ASV system, LPS are extracted according to the speaker list of VoxCeleb1. The training set contains a total of 1,211 speakers with a total of 148,624 utterances. There are also 4,874 utterances from 40 speakers to test the performance of ASV. The structure of ASV is shown in Figure 2. During the training phase, each training sample consists of two input features and a {0,1} target. If both features originate from the same speaker, the target is 1; otherwise, it is 0. The network is updated with minibatches of size 64, each containing an equal number of samples with targets 0 and 1 to avoid class imbalance in training. The network parameters are updated to minimize the cross-entropy loss between the sigmoided output of the network and the target labels utilizing Adam. The learning rate of 60 iterations is selected to be 0.001, and the weight of the two-norm regularization is set to $5e^{-5}$. After training on VoxCeleb1, the ASV model is fine-tuned on the ASVSpoof2019 dataset with the CM training list (ASVSpoof2019.LA.cm.trn.txt) [40] for another 20 iterations. The learning rate is set to 0.0001. Each utterance is reduced by TDNN to become a 512-dimensional vector [41]. During fine-tuning, TDNN modules in Figure 2 are fixed. Green "Siamese" modules and orange "discriminator" modules are updated.

A normalized version of *tandem detection cost function* (t-DCF) from [33, 34] is utilized to evaluate the performance on attacking the combined system of ASV and CM. The detection threshold (set to the EER operating point) of the

ASV system is fixed, whereas the detection threshold of the CM system is allowed to vary. Results are reported in the form of minimum normalized t-DCF values. The normalized t-DCF is defined as a function of the CM threshold, and the minimum normalized t-DCF defined in (11) is finally computed to evaluate the performance of the joint system,

$$t\text{-}DCF_{\text{norm}}^{\min} = t\text{-}DCF_{\text{norm}}\left(\text{argmin}_{\theta CM} t\text{-}DCF_{\text{norm}}\left(\theta_{CM}\right)\right). \tag{11}$$

The value of (11) ranges from 0 to 1. The closer it is to 1, the more the errors occurring in the combined system are.

When evaluating the performance of the tandem system, the same cost parameters as minimum normalized t-DCF in the ASVSpoof2019 Challenge are utilized. The threshold of ASV is fixed to its Equal Error Rate (EER) and swept over CM thresholds for minimal normalized t-DCF. t-DCF is utilized only for final evaluation and not for optimization training of ASV or CM systems. When evaluating the attack effect of adversarial examples, the False Acceptance Rate (FAR) is adopted, which is defined as the proportion of speech uttered by imposters (nontarget or spoof) but accepted by systems. If there is no special notice below, all experiments are tested on the combined protocols of the ASVSpoof2019 dataset (ASVSpoof2019.LA.asv.dev.gi.trl.txt and ASVSpoof 2019.LA.asv.eval.gi.trl.txt).

5.2. Experiments Settings. A tandem system of ASV and CM can be achieved by connecting the individually trained subsystems in the form of Figure 4. The tandem system is attacked by FGSM and PGD. In both FGSM and PGD attack settings, the maximum amplitude of perturbation ε is chosen from the set of $\{0.1, 1, 5, 10\}$. Since PGD is an iterative algorithm, the relationship between the step of PGD and the maximum amplitude of perturbation is

$$\varepsilon = K \cdot \alpha. \tag{12}$$

To achieve a valid adversarial attack, in addition to having a high attack success rate, it is also important to make adversarial examples indistinguishable from the original audios to humans. An XAB listening test is conducted to evaluate the imperceptibility of adversarial audios, which is a standard detection method to assess the detectable differences between two choices of sensory stimuli. In the XAB test, the adversarial examples generated by the PGD algorithm when $\varepsilon = 10$ are utilized. Adversarial audios are generated by combining perturbed LPS and the phase of the original utterance. Five listeners were involved in the test, each of whom was asked to listen to 50 randomly selected adversarial-original audio pairs (A and B). An utterance (X) was randomly selected from each pair, and listeners chose whether the utterance was closer to A or B.

5.3. Experiments Results

5.3.1. The Performance of Systems without Attacks. The performance of ASV and CM is evaluated under the ASVSpoof2019 protocol separately. The protocol file contains both labels of the ASV and CM. Each trial contains 4 columns, which are claimed speaker ID, utterance ID, CM label (bonafide/A01-A19), and ASV label (target/nontarget/spoof). Here, ASV-V represents the model trained on VoxCeleb1, and ASV-S represents the model fine-tuned on ASVSpoof2019. When the thresholds of subsystems are all fixed at the EER point, the FAR of ASV-V, ASV-S, and CM are 9.47%, 6.21%, and 5.43%, and the FAR and t-DCF of the tandem system are 5.67% and 0.023, respectively. When testing the t-DCF of the tandem system, the ASV threshold is fixed; adjust the CM threshold to find the minimized t-DCF, as shown in Figure 5.

5.3.2. Evaluation of White-Box Digital Attacks. In order to intuitively display the distribution of different kinds of samples, the ASV and CM subsystems were utilized to score samples with different labels. Figure 6 is the score histogram of ASV and CM systems. Figure 6(a) shows the scores of original utterances, and Figure 6(b) shows the scores of adversarial examples. For the ASV system, the adversarial object is to accept all the utterances originally labeled as nontarget. For the CM system, the adversarial object is to accept all the utterances originally labeled as spoof. Comparing Figures 6(a) and 6(b), it can be seen that the scores of some utterances labeled nontarget and spoof are less than the threshold of ASV before the adversarial attack. Also, the score distinction between bonafide and spoof examples derived from the CM system is obvious. However, after the attack, the scores of most examples are higher than the threshold of ASV, and the distinction between bonafide and spoof has become not obvious.

The performance of ASV-S, CM, and tandem systems after adversarial attack is shown in Table 2. PGD-100 means that the number of iterations is 100. FAR and t-DCF are utilized to evaluate the performance of adversarial attacks. It can be seen that the PGD method is more effective than the FGSM. At the same time, the higher the upper perturbation limit is, the more effective the attack is.

5.3.3. Visualization. Figure 7 is a t-SNE diagram of audios w/o the adversarial attack. Samples whose *Claimed IDs* are LA_0073 have been chosen to be shown in the figure. Before the attack, three kinds of samples are clearly distinguishable, and the boundaries of each type are relatively clear. The tandem system can easily distinguish the three types of samples. After the attack, the classification boundary of the adversarial examples is gradually blurred. Therefore, the FAR and t-DCF will increase. This shows that the proposed attack algorithm on the tandem system has played its due role and can make subsystems produce misclassifications at the same time. Therefore, the utterance whose original label is spoof or nontarget can be recognized as bonafide by CM and target by ASV.

5.3.4. Evaluation of Imperceptibility. The subjective XAB listening test in Section 4 results in average classification accuracy of 47.2%, which confirms the imperceptibility of

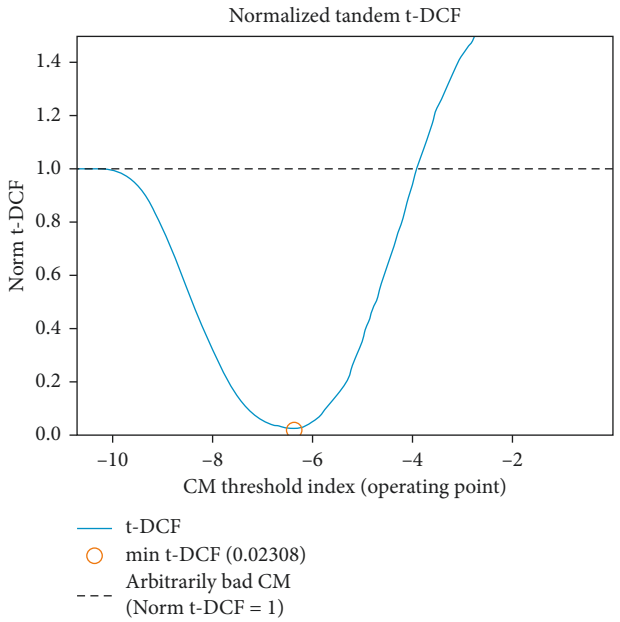
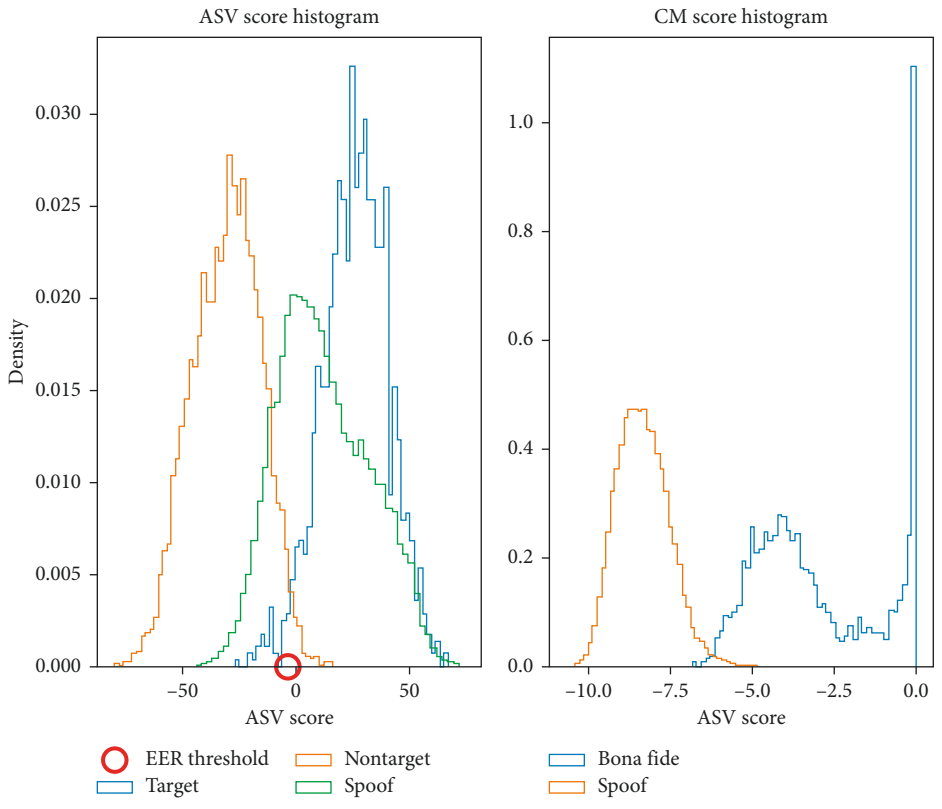


Figure 5: The curve of t-DCF varies with the threshold of the CM system.



(a)

Figure 6: Continued.

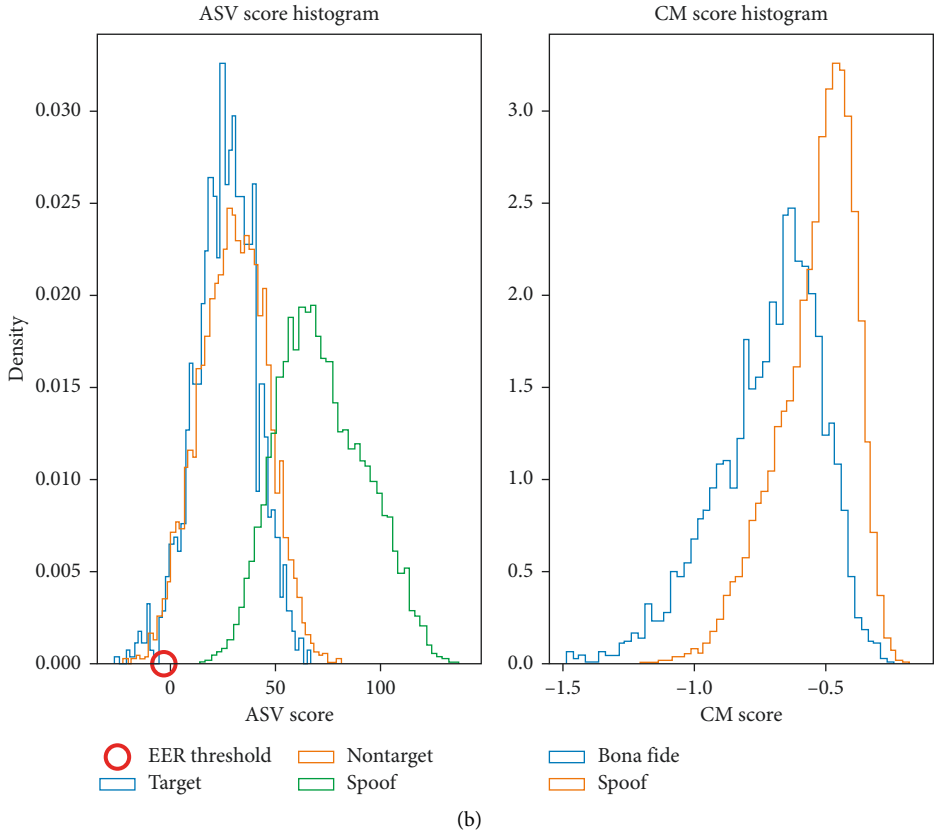


Figure 6: The score histogram of ASV and CM. (a) Score histogram of original utterances. (b) Score histogram of adversarial utterances.

Table 2: The performance of ASV-S, CM, and tandem systems.

System	$\varepsilon = 0.1$				$\varepsilon = 1$			
	FGSM		PGD-100		FGSM		PGD-100	
	FAR (%)	t-DCF	FAR (%)	t-DCF	FAR (%)	t-DCF	FAR (%)	t-DCF
ASV-S	20.1	—	23.4	—	46.4	—	82.9	—
CM	15.6	—	18.9	—	38.7	—	73.6	—
Tandem	10.2	0.213	14.3	0.369	32.5	0.675	69.9	0.852
System	$\varepsilon = 5$				$\varepsilon = 10$			
	FGSM		PGD-100		FGSM		PGD-100	
	FAR (%)	t-DCF	FAR (%)	t-DCF	FAR (%)	t-DCF	FAR (%)	t-DCF
ASV-S	64.2	—	92.4	—	91.0	—	**97.3**	—
CM	59.6	—	90.7	—	87.3	—	**98.1**	—
Tandem	55.4	0.807	87.1	0.964	81.0	0.928	**96.7**	**1.000**

adversarial examples. The spectrogram of LA_D_1000265.wav w/o adversarial attack is shown in Figure 8, where $\varepsilon = 10$. From top to bottom are the spectrogram of original utterance, perturbed utterance, and perturbation. As can be seen from the figure, the difference between the original utterance and the adversarial utterance is tiny. Most areas of the spectrogram of perturbation are very low in energy, which shows the imperceptibility of perturbation.

5.3.5. The Joint Attacks versus the Step-by-Step One. In addition, to verify that a joint attack is really effective, the performance of a step-by-step adversarial attack utilizing PGD-100 is evaluated. The step-by-step adversarial attack is divided into two situations: (1) ASV system is attacked first, and CM system is attacked again (ASV $\longrightarrow$ CM; i.e., first $\alpha = 1$ and $\beta = 0$, and then $\alpha = 0$ and $\beta = 1$). (2) CM system is attacked first, and ASV system is attacked again (CM $\longrightarrow$ ASV; i.e., first $\alpha = 0$ and $\beta = 1$, and then $\alpha = 1$ and $\beta = 0$). The results of experiments are shown in Table 3. Because the number of labels {target, nontarget, spoof} in protocols (ASVSpoof 2019.LA.asv.dev.gi.trl.txt and ASVSpoof2019.LA.asv.eval. gi.trl.txt) is not balanced, the performances of the two kinds of step-by-step attacks are different. The total number of trials is 132,127, the number of target trials is 6,854, nontarget is 39,095, and spoof is 86,178.

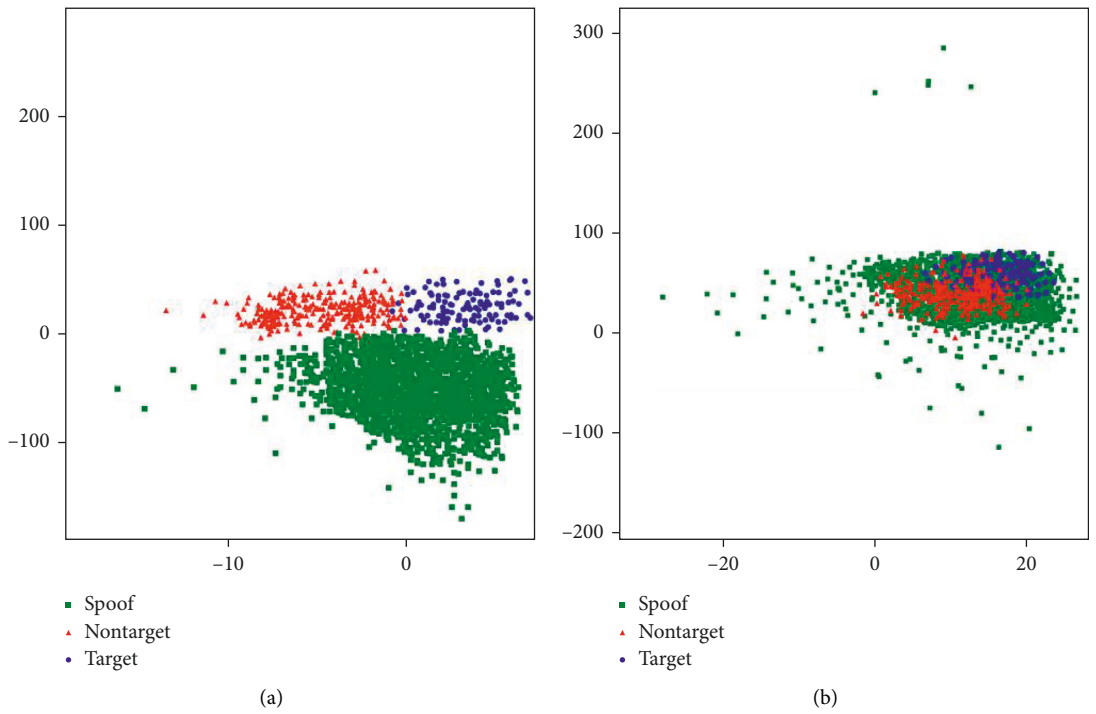


Figure 7: T-SNE figure of utterances with different labels: (a) before attack and (b) after attack.

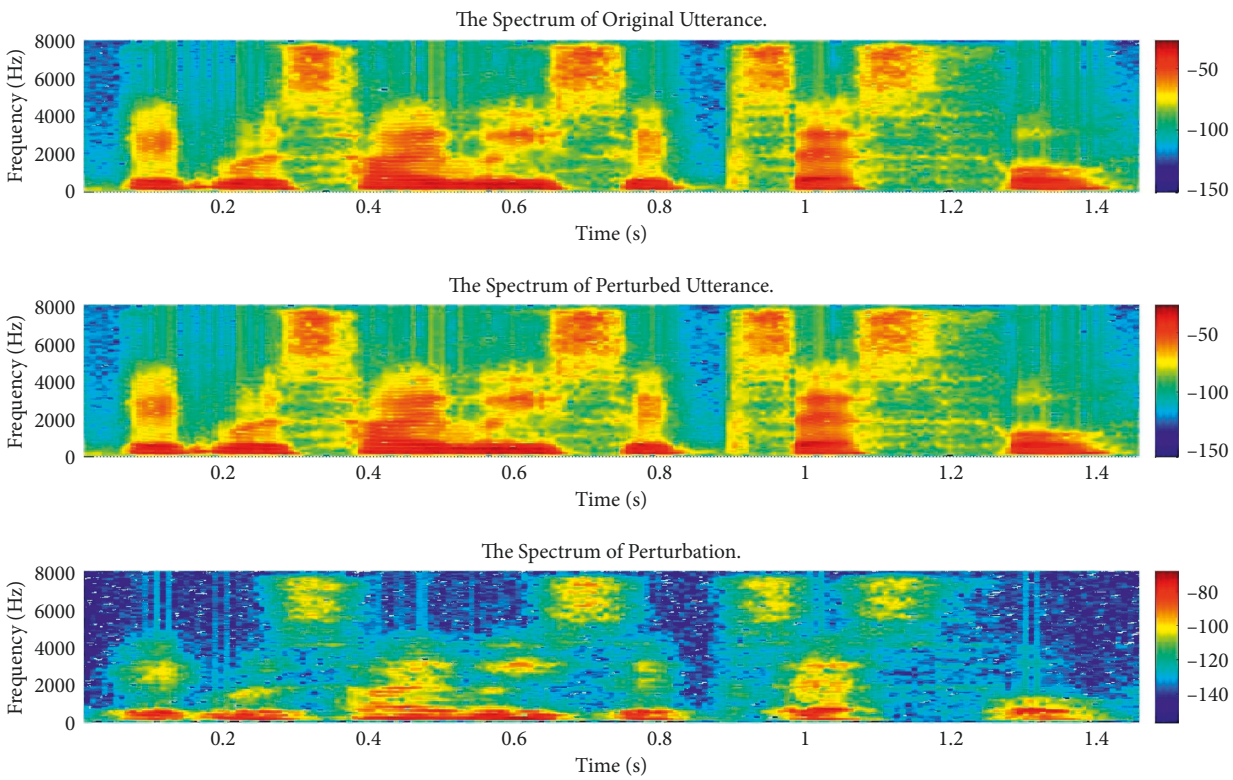


Figure 8: Spectrograms of original utterance, adversarial utterance, and perturbation.

Sample analysis in Table 4 showed that almost all trials labeled spoof were accepted in the ASV ⟶ CM attack, while trials labeled nontarget were almost rejected. The opposite phenomenon was found in the CM ⟶ ASV attack. It can be seen that the labels of trials have a strong correlation with the results of step-by-step attacks. We believe that the reason for this phenomenon may be that the objective functions of attacking CM and ASV are not exactly

Table 3: The performance of step-by-step adversarial attack.

System	ASV ⟶ CM	CM ⟶ ASV	Joint
FAR (%)	63.8	34.4	96.7
t-DCF	0.823	0.256	1.000

Table 4: Trial analysis of step-by-step attack. The number of trials labeled target, nontarget, and spoof accepted by systems is shown in the table.

System	Target	Nontarget	Spoof
Before attack	6,840	1,895	4,094
ASV ⟶ CM	6,843	1,895	78,091
CM ⟶ ASV	6,841	39,039	4,094
Total number	6,854	39,095	86,178

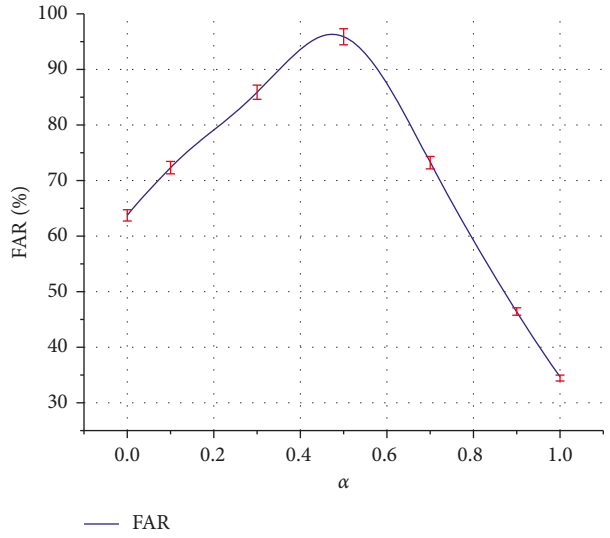


Figure 9: The FAR variation of tandem system when changing α.

the same. Since the production of adversarial samples is to add perturbation in the whole time and frequency bands, the perturbation added later will cover the perturbation added before. Therefore, an adversarial sample that is effective for one system is invalid for another. But whether it is ASV ⟶ CM or CM ⟶ ASV, the performance of a joint attack is far better than that of a step-by-step attack.

5.3.6. Parameter Sensitivity. In equations (6) to (10), α and β are introduced to adjust the weight of the loss function for ASV and CM systems, respectively. In order to explore the effect of changing α and β during joint adversarial attacks, a series of experiments are deployed. Since $\alpha + \beta = 1$, if one of the two parameters is adjusted, the other will change as well. The variation of FAR is shown in Figure 9. By studying the FAR curves of the tandem system in Figure 9, we see that when α is close to 0 (ASV subsystem has a small weight) or 1 (CM subsystem has a small weight), the effect of adversarial attacks is not satisfactory, when $\alpha = \beta = 0.5$, the best attack result can be obtained. The experiments show that the attack ability of subsystems with lower weight will be significantly reduced when the weight of ASV or CM loss function is reduced. The performance degradation when reducing α is more gradual than when reducing β. This phenomenon may be due to the uneven distribution of trials belonging to different labels. Trials labeled spoof are more numerous than nontarget. Modifying α and β shows that when attacking a tandem system, a drop in the weight of either loss function is not tolerated. It is not desirable to sacrifice the performance of one subsystem for the performance of the other. Because both subsystems play a vital role in the tandem system, it is critical to keep both subsystems performing well.

6. Conclusion

In this paper, an attack method for the tandem system of ASV and CM is proposed. PGD and FGSM are utilized to implement attacks on the tandem system. Through the proposed attack method, the tandem system can be attacked successfully. The vulnerability of the tandem system to adversarial attacks is revealed. In the future, black-box attacks against tandem systems will be explored, and adversarial defense and detection methods will also be utilized to improve the robustness and security of the tandem system.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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

References

[1] M. Algabri, H. Mathkour, M. A. Bencherif, M. Mekhtiche, and M. A. Mekhtiche, "Automatic speaker recognition for mobile forensic applications," *Mobile Information Systems*, vol. 2017, pp. 1–6, 2017.

[2] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: an overview," in *Proceedings of the 2020 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Shanghai, China, Oct. 2020.

[3] X. Wang, J. Yamagishi, M. Todisco et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, Article ID 101114, 2020.

[4] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of ASVspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, no. 1, 2020.

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

[6] Z. Wu, J. Yamagishi, T. Kinnunen et al., "ASVspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[7] Z. Wu, T. Kinnunen, N. Evans et al., "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proceedings of the 2015 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2037–2041, Dresden, Germany, Sept. 2015.

[8] T. Kinnunen, M. Sahidullah, H. Delgado et al., "The ASVspoof 2017 Challenge: assessing the limits of replay spoofing attack detection," in *Proceedings of the 2017 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 20–24, Stockholm, Sweden, Aug. 2017.

[9] M. Todisco, X. Wang, V. Vestman et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proceedings of the 2019 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1008–1012, Graz, Austria, Sept. 2019.

[10] A. Consortium, "ASVspoof2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2019, https://www.asvspoof.org/asvspoof2019/asvspoof%202019_evaluation_plan.pdf.

[11] A. Gomez-Alanis, A. M. Peinado, J. A. Gomez, and A. M. Gomez, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1985–1999, 2019.

[12] H. Zeinali, T. Stafylakis, G. Athanasopoulou et al., "Detecting spoofing attacks using VGG and SincNet: BUT-Omilia submission to ASVspoof 2019 challenge," in *Proceedings of the 2019 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1073–1077, Graz, Austria, Sept. 2019.

[13] C. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-spoofing with squeeze-excitation and residual networks," in *Proceedings of the. 2019 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1013–1017, Graz, Austria, September 2019.

[14] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC anti-spoofing systems for the ASVspoof2019 challenge," in *Proceedings of the. 2019 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1033–1037, Graz, Austria, Sept. 2019.

[15] H. Luo, Y. Shen, F. Xu, and G. Xu, "Spoofing speaker verification system by Adversarial examples Leveraging the Generalized speaker difference," *Security and Communication Networks*, vol. 2021, pp. 1–10, 2021.

[16] H. Yakura and S. Jun, "Robust audio adversarial example for a physical attack," in *Proceedings of the. 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5334–5341, Macao, China, Aug. 2019.

[17] J. Jeong, S. Kwon, M. P. Hong, J. Shon, and T. Shon, "Adversarial attack-based security vulnerability verification using deep learning library for multimedia video surveillance," *Multimedia Tools and Applications*, vol. 79, no. 23-24, Article ID 16077, 2020.

[18] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

[19] S. S. Chandrasekaran and V. Chandrasekaran, "A robust hybrid digital watermarking technique against a powerful CNN-based adversarial attack," *Multimedia Tools and Applications*, vol. 79, no. 43-44, Article ID 32769, 2020.

[20] X. Yuan, P. He, Q. Li, and X. Li, "Adversarial examples: attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[21] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," in *Proceedings of the. of DYnamic and Novel Advances in Machine Learning and Intelligent Cyber Security (DYNAMICS) Workshop*, San Juan, Puerto Rico, USA, 2018.

[22] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM i-vector based speaker verification systems," in *Proceedings of the. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6579–6583, Barcelona, Spain, May 2020.

[23] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," in *Proceedings of the. 2020 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 4228–4232, Shanghai, China, Oct. 2020.

[24] V. Jesús, Y. Zhang, and N. Dehak, "X-vectors meet adversarial attacks: benchmarking adversarial robustness in speaker verification," in *Proceedings of the. 2020 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 4233–4237, Shanghai, China, Oct. 2020.

[25] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical Adversarial Attacks Against speaker recognition systems," in *Proceedings of the. 21st International Workshop on Mobile Computing Systems and Applications (ACM Hot Mobile)*, pp. 9–14, Austin, Texas, USA, Mar. 2020.

[26] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with Adversarial examples," in *Proceedings of the. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1962–1966, Calgary, AB, Canada, Apr. 2018.

[27] A. Jati, C. C. Hsu, M. Pal, R. Peri, W. Narayanan, and S. Narayanan, "Adversarial attack and defense strategies for deep speaker recognition systems," *Computer Speech & Language*, vol. 68, no. 2021, Article ID 101199, 2021.

[28] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal and robust Adversarial Attacks Against speaker recognition systems," in *Proceedings of the. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1738–1742, Barcelona, Spain, May 2020.

[29] J. Li, X. Zhang, C. Jia et al., "Universal adversarial perturbations generative network for speaker recognition," in *Proceedings of the. International Conference on Multimedia and Expo (ICME)*, pp. 1–6, London, UK, Jul. 2020.

[30] G. Chen, S. Chen, L. Fan et al., "Who is real Bob? adversarial attacks on speaker recognition systems," in *Proceedings of the. of IEEE Symposium on Security and Privacy Workshops (SPW)*, San Francisco, CA, USA, May 2021.

[31] Y. Zhang, Z. Jiang, J. Villalba, and N. Dehak, "Black-box attacks on spoofing countermeasures using transferability of adversarial examples," in *Proceedings of the. 2020 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 4238–4242, Shanghai, China, Oct. 2020.

[32] S. Liu, H. Wu, H. Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in

Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 312–319, Singapore, Dec. 2019.

[33] T. Kinnunen, K. A. Lee, H. Delgado et al., "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proceedings of the.Odyssey 2018: The Speaker and Language Recognition Workshop*, pp. 312–319, Les Sables d'Olonne, France, June 2018.

[34] T. Kinnunen, H. Delgado, N. Evans et al., "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[35] M. Ravanelli and B. Yoshua, "Learning speaker representations with mutual information," in *Proceedings of the 2019 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1153–1157, Graz, Austria, Sept. 2019.

[36] A. Kanervisto, V. Hautamäki, T. Kinnunen, and J. Yamagishi, "An initial investigation on optimizing tandem speaker verification and countermeasure systems using Reinforcement learning," 2020, http://arXiv.org/abs/2002.03801.

[37] J. Hu, L. Shen, S. Albanie, G. Wu, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[38] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the. of International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, Apr. 2018.

[39] W. Zhang, S. Zhao, L. Liu et al., "Attack on practical speaker verification system using universal adversarial perturbations," in *Proceedings of the. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2575–2579, Toronto, ON, Canada, June 2021.

[40] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, Article ID 101027, 2020.

[41] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: robust DNN embeddings for speaker recognition," in *Proceedings of the. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, Calgary, AB, Canada, Apr. 2018.