electronics

Review

Adversarial Attack and Defense Strategies of Speaker Recognition Systems: A Survey

Hao Tan [1,2], Le Wang [1,2], Huan Zhang [1,2], Junjian Zhang [1], Muhammad Shafiq [1,*] and Zhaoquan Gu [1,2,*]

[1] Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China; th198@e.gzhu.edu.cn (H.T.); wangle@gzhu.edu.cn (L.W.); zhhuan@e.gzhu.edu.cn (H.Z.); 2112106069@e.gzhu.edu.cn (J.Z.)
[2] Department of New Networks, Peng Cheng Laboratory, Shenzhen 518055, China
* Correspondence: srsshafiq@gmail.com (M.S.); zqgu@gzhu.edu.cn (Z.G.)

Abstract: Speaker recognition is a task that identifies the speaker from multiple audios. Recently, advances in deep learning have considerably boosted the development of speech signal processing techniques. Speaker or speech recognition has been widely adopted in such applications as smart locks, smart vehicle-mounted systems, and financial services. However, deep neural network-based speaker recognition systems (SRSs) are susceptible to adversarial attacks, which fool the system to make wrong decisions by small perturbations, and this has drawn the attention of researchers to the security of SRSs. Unfortunately, there is no systematic review work in this domain. In this work, we conduct a comprehensive survey to fill this gap, which includes the development of SRSs, adversarial attacks and defenses against SRSs. Specifically, we first introduce the mainstream frameworks of SRSs and some commonly used datasets. Then, from the perspectives of adversarial example generation and evaluation, we introduce different attack tasks, the prior knowledge of attacks, perturbation objects, perturbation constraints, and attack effect evaluation indicators. Next, we focus on some effective defense strategies, including adversarial training, attack detection, and input refactoring against existing attacks, and analyze their strengths and weaknesses in terms of fidelity and robustness. Finally, we discuss the challenges posed by audio adversarial examples in SRSs and some valuable research topics in the future.

Keywords: speaker recognition; adversarial examples; adversarial attacks; defense methods

Citation: Tan, H.; Wang, L.; Zhang, H.; Zhang, J.; Shafiq, M.; Gu, Z. Adversarial Attack and Defense Strategies of Speaker Recognition Systems: A Survey. *Electronics* **2022**, *11*, 2183. https://doi.org/10.3390/electronics11142183

Academic Editors: Celestine Iwendi and Thippa Reddy Gadekallu

Received: 5 June 2022
Accepted: 5 July 2022
Published: 12 July 2022

1. Introduction

The rapid development of deep learning techniques has considerably advanced research progress on healthcare [1–3], IoT devices [4–6] and biometric authentication techniques [7]. Recently, face recognition as a mainstream authentication technology has reached an accuracy up to 99% and achieved commercial success. Meanwhile, end-to-end SRSs have witnessed improved performance. Recent studies [8–10] show that the equal error rate (EER) of speaker recognition models has been reduced to 0.77%, which means an unprecedentedly high recognition accuracy. Unlike other authentication methods based on biometric features, such as human faces or fingerprints, SRSs can identify the speaker by speech features unique to him/her, even in the absence of the speaker. Consequently, SRSs have seen wide adoption in such fields as remote access control, bank service, and criminal investigations.

Along with the prevalence and increasing influence of the speaker recognition technology, its security has drawn broad attention. Though SRSs have reached a high recognition accuracy, their security remains a big concern since a minor perturbation on the audio input may result in reduced recognition accuracy. Existing attacks on SRSs can be classified into three types: traditional attack, backdoor attack, and adversarial attack. There are mainly four sub-types of traditional attacks: mimicry attacks [11] in which the attacker mimics the target speaker, voice conversion (VC) attacks [12,13], text-to-speech (TTS) attacks [14–16],

and another simple but strong attack—replay attack (RA) [17–21], which fully copies and replays the speech signals of the target speaker. There have been many defense solutions to these traditional attacks that maximally perceive, mimic and copy the target speaker's voice features. Notably, the community-led initiative ASVspoof Challenge series [22,23] provided an ideal platform for the development of defense solutions to audio adversarial attacks. Backdoor attacks first emerged in image processing tasks [24–26], in which a specific trigger is introduced to the dataset to train a recognition model with high-security threats, such as a face recognition model. In the latest work, Zhai et al. [27] introduced a backdoor attack strategy against SRSs. By introducing triggers to different types of datasets clustered by K-means, they found that the speaker recognition models trained by these contaminated datasets would produce wrong recognition results once the trigger was input to the testing process. There are no effective defense solutions against these backdoor attacks at present.

Adversarial attacks aim to lead the SR models to wrong decisions by introducing imperceptible perturbations to clean audio samples. In these years, deep learning based methods have been proved to be vulnerable to adversarial attacks. Szegedy et al. [28] first found in image classification tasks that introducing unnoticeable perturbations to the training samples would make the model produce a wrong output with high confidence. Many subsequent works have focused on adversarial attacks against computer vision systems [29–32]. Later, studies on these attacks were extended to such fields as natural language processing [33,34], audio processing [35] and video classification [36]. In audio processing, most works [37–41] are on speech recognition (SR), whereas few probe into SRS or adversarial examples of SR models.

Adversarial attacks against SRSs introduce imperceptible noises into the genuine audio data such that the system misrecognizes the speaker (untargeted attacks) or recognize the subject as a designated speaker (targeted attacks). Rohan et al. [42] summarized the existing works on non-active attacks and adversarial attacks against SRSs as well as their respective defense measures, but the classification standards could not be used to clearly classify all existing attacks and defenses. Hadi et al. [43] classified the adversarial attack and defense methods of speech recognition models and SRSs based on the attack target, attack type, adversarial knowledge, and adversarial capabilities; however, their research focused on speech recognition systems, not SRSs. Chen et al. [44] reviewed the existing works on adversarial attack and defense of voice processing systems (VPSes), and classified existing adversarial attacks against SRSs from four perspectives, i.e., adversarial knowledge, adversarial target, adversarial perturbation range, and physical attack; however, they provided no further analysis on each type of attack, and analysis of existing works on adversarial defenses needed to be extended.

In this work, an overview of all previous works on adversarial attacks and defenses of SRSs is presented, and the generation methods of adversarial examples against SRSs and corresponding defense strategies are discussed and summarized. Compared with existing works in this field, more innovative and detailed classification methods for adversarial attack and defense strategies of SRSs are proposed here, and the latest works on adversarial examples against SR methods are introduced. The major contributions of the this work are as follows:

- An overview of the existing works on SRS is presented to introduce the latest advances in adversarial examples against SRSs, and from the perspectives of example generation and example evaluation, we classify the adversarial examples by such indicators as the attack task, the perturbation target, the perturbation constraint strategies, and attack effect evaluation.
- We review and classify the existing attack and defense methods from three aspects: adversarial training, attack detection, and input refactoring, and measure the effectiveness of these methods by fidelity and robustness.

The rest of the work is organized as follows: Section 2 introduces advanced SRSs and regular SR datasets; Section 3 presents the methods for classification of adversarial attacks; Section 4 introduces the defense strategies against existing SRS adversarial examples and

our classification methods; in Section 5, we discuss the shortcomings and challenges in the field of SRS adversarial attack and defense strategies; and Section 6 concludes our work and pinpoints the valuable research directions.

2. Background

In this section, we review the development of SRSs, and introduce the working principles, functional modules, and recognition tasks of SRSs.

2.1. Overview of SRS

Speaker verification systems consist of two modules: front-end embedding and back-end scoring. For any given audio, the embedding module represents the acoustic features of the audio by fixed-length high-dimensional feature vectors, and these vectors are then input to the back-end scoring module for similarity calculation to obtain the corresponding speaker labels for this segment of audio.

The earliest SR models, such as the dynamic time warping (DTW) model [45], recognize the speakers based on the speech signals by template matching. Later, some Gaussian mixture models (GMMs) [46,47], like the Gaussian mixture model-universal background model (GMM-UBM) and the Gaussian mixture model-support vector machine (GMM-SVM) model were developed, which represent the original audio signals by the trained model to recognize the speaker. Then, identify vector models (i-vector) [48] that recognize speaker voice features became the mainstream methods because they rely on data of smaller lengths. As deep learning technology advances, deep speaker vectors come to play a dominating role: deep neural networks are trained to extract speech features and represent the speech features as d-vectors [49] or x-vectors [8]. Bai et al. [50] made a detailed summary of works on DNN-based SRSs.

Figure 1 presents the general framework of traditional SR and deep learning based SRSs, which comprises three stages: training, enrolling, and verification.

- Training: over ten thousand audio clips from large amounts of speakers are used to train the speaker embedding module and obtain human voice feature distributions, regardless of single speakers;
- Enrolling: the enrolled speaker utterance is mapped onto a unique labeled speaker embedding through the speaker embedding module, and this high-dimensional feature vector is this speaker's unique identity;
- Verification: the model scores the utterance of an unknown speaker by extracting high-dimensional feature vectors from the embedding module. The scoring module assesses the similarity between the recorded embedding and the speaker embedding, and the score and decision module is based on to judge whether the speaker is legitimate.

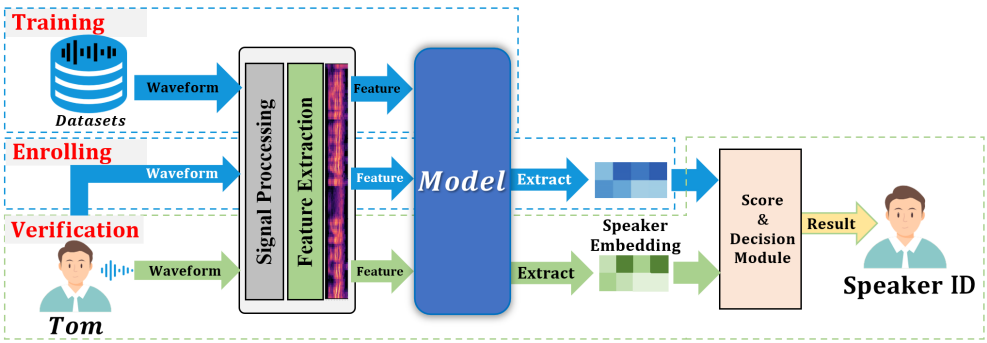

Figure 1. General framework of the speaker recognition systems.

At the training stage, the feature extraction module converts the original speech signals into acoustic waveforms with primary signal features. Regular feature extraction algorithms include Mel frequency analysis, filter-bank, Mel-scale frequency cepstral coefficients

(MFCC) [51], and perceptual linear predictive (PLP) [52]. The speaker embedding network can be modeled by models such as LSTM, ResNet, time-delay neural network (TDNN), etc. There are two types of back-end scoring models: probabilistic linear discriminant analysis (PLDA) [53] and cosine similarity [54]: the former works well in most cases, but requires training based on utterances [55]; the latter provides an alternative to PLDA but dispenses with the need for training.

2.2. SR Task

SR tasks can be divided into text-dependent and text-independent tasks by whether the audio clips are recorded by specific texts at the enrolling and verification stages. In text-dependent tasks, speech examples of specific texts are produced in both the training and testing stages, and though the model training consumes little time, the text is specific, and hence the model is short of universality. Text-independent tasks do not depend on the content of the audio, and the verification module recognizes the speaker by converting the audio content into the speaker's high-dimensional speaker feature vectors, which is convenient but consumes considerable quantities of training resources. In the present work, we consider only the adversarial attack and defense of text-independent SRSs (in fact, most works in this regard focus on text-independent SRSs). In text-independent SRSs, SR tasks can be divided by the task target into close-set speaker identification (CSI) tasks, open-set speaker identification (OSI) tasks, and speaker verification (SV) tasks.

2.2.1. CSI Task

Close-set speaker identification (CSI) tasks [56,57] can be regarded as a multi-classification problem, which identifies a specific speaker from the corpus of a set of enrolled speakers, i.e., the system always identifies an input audio as a specific label in the training dataset. Chen et al. [58] divided CSI tasks into two sub-tasks: CSI with enrollment (CSI-E) and CSI with no enrollment (CSI-NE). CSI-E strictly follows the process described above. In contrast, CSI-NE has no enrollment, and the speaker embedding module can be used directly to recognize the speaker. Thus, ideally speaking, in CSI-NE tasks, the identified speaker will take part in the training stage, whereas in CSI-E tasks, the identified speaker has already been enrolled in the enrolling stage but does not necessarily take part in the training stage. Equation (1) describes the general process of CSI tasks:

$$I = \arg\max_i \{f(x_1^e, x^t; \theta), f(x_2^e, x^t; \theta), ..., f(x_N^e, x^t; \theta)\} \quad (1)$$

where I denotes the speaker label, θ is the parameter of the embedding model, and N is the number of registered speakers. $f(\cdot)$ denotes the similarity score calculated between the registered vector x^e and the test vector x^t, and the model outputs the speaker ID with the highest score.

2.2.2. OSI Task

Different from CSI tasks, in open-set speaker identification (OSI) tasks [8], the model obtains a threshold by the PLDA or cosine similarity algorithm, and recognizes the test utterance as an enrolled speaker by comparing the calculated similarity score and the preset threshold. OSI tasks can also identify unknown speakers. That is, a speaker that is not in the original training dataset can also be enrolled in the OSI system to generate specific feature vectors, and in the verification process, the model converts the to-be-identified speaker into high-dimensional vector embeddings, and uses the back-end scoring module to produce a similarity score: if the maximum score is below the preset threshold, then the speaker is identified as an unenrolled speaker and hence is denied access. Similarly, the process of OSI tasks can be summarized by an equation, as shown in Equation (2):

$$I = \arg\max_i \{f(x_1^e, x^t; \theta), f(x_2^e, x^t; \theta), ..., f(x_N^e, x^t; \theta)\}$$
$$while \quad f(x_i^e, x^t; \theta) > \tau \tag{2}$$

where τ is a pre-received threshold in the model, the test audio will be accepted and correctly recognized by the system when and only when the maximum score exceeds the threshold τ in OSI, otherwise the model will directly filter out the audio.

2.2.3. SV Task

Both CSI and OSI tasks can be termed collectively as a 1:N identification task (i.e., discriminating input audio among a collection of N-registered speakers), and they require a large number of different speakers' speech for model training. In contrast, the SV system aims to verify whether an input voice (virtual speaker) is pronounced according to his/her pre-recorded words, which is a 1:1 identification task that models the vocal characteristics of only one speaker, and then verifies whether the input voice is produced by a unique registered speaker according to a predefined threshold, and if the score does not exceed the threshold, the input voice is considered an impostor and is rejected.

$$f(x^e, x^t; \theta) = \begin{cases} \text{Accept}, & S > \tau \\ \text{Reject}, & S \leq \tau \end{cases} \tag{3}$$

where $f(\cdot)$ represents the calculation of the similarity score S between the registered vector x^e and the test vector x^t, and θ is the parameter of the embedding model. The score is accepted if it is greater than a threshold value and rejected otherwise.

2.3. Victim Models

Existing speaker attacks are mainly against SR models built on deep neural networks (DNNs), such as SincNet, d-vector, and x-vector, rather than the template matching-based DTW models and the statistical distribution-based GMM, GMM-UBM, and GMM-SVM models.

As Table 1 shows, the i-vector SR model proposed by Kanagasundaram [59] shifts the high-dimensional speaker features into a lower-dimension full-factor subspace, models global differences in data in low dimensions, and combines systems, such as GMM-MMI, to enhance the recognition capability of the model in this low-dimensional subspace, and improves the identification capacity of the model in this low-dimension space by GMM-MMI and other systems, which reduces the computing complexity and training time. However, as the i-vector model maps the data into the full-factor subspace, the system is susceptible to noises. Therefore, Variani et al. [49] proposed to use DNN for the feature extraction of speaker audio and took the output of the last hidden layer as the speaker's features and took the average of all the speaker's features as the speaker's vocal embedding vector, a model called d-vector. The d-vector model has better performance compared to the i-vector model both in clean and noisy environments. David Snyder proposed the x-vector model [8], which uses the TDNN structure for feature extraction, and compared to the d-vector, which simply averages the speaker features as the voice pattern model, the x-vector aggregates the speaker features and inputs them into the DNN again to obtain the final voice pattern model. The r-vector model proposed by Hossein et al. [60] applies ResNet, which further reduces the EER compared to the x-vector model. Mirco Ravanelli [61] argues that acoustic features extracted by traditional i-vector methods and deep learning methods using signal processing techniques (e.g., MFCC, and FBank) would lead to a loss of acoustic features in the original audio, for which he proposed the SincNet model, which uses a data-driven approach to learn filter parameters directly, allowing the model to learn narrowband speaker characteristics, such as pitch and resonance peaks, well from the original data. In recent studies, Brecht et al. [9] proposed ECAPA-TDNN, a new TDNN-based vocal feature extractor; ECAPA-TDNN further develops on the original x-vector architecture, focusing more on the channels as well as the propagation and aggregation of features,

resulting in a 19% improvement in the EER performance of the system compared to the x-vector model. The deep speaker [62] proposed by Baidu adopts an end-to-end strategy to aggregate feature extraction and speaker recognition into the network structure, which can improve the performance of the fixed speaker list.

Table 1. Common victim SR models.

Strategy	Model	Dataset	Task	Metrics	Performance
Statistics	GMM-UBM	NIST SRE	OSI/SV	EER	1.81%
	i-vector	NIST 2008	OSI/SV	EER	6.3%
Embedding	AudioNet	LibriSpeech	CSI	ACC	99.7%
	VGGvox	Voxceleb1	CSI	ACC	92.1%
	d-Vector	Google data	OSI/SV	EER	4.54%
	x-vector	VoxCeleb	OSI/SV	EER	4.16%
	r-Vector	VoxCeleb	OSI/SV	EER	1.49%
	SincNet	LibriSpeech	OSI/SV	EER	0.96%
	ECAPA-TDNN	VoxCeleb2	OSI/SV	EER	0.87%
End to End	ResCNN	MTurk	OSI/SV	EER	2.83%
	GRU	MTurk	OSI/SV	EER	2.78%

2.4. Datasets

Depending on different tasks and target models, researchers choose publicly available mainstream datasets to evaluate their attack performance. Some mainstream datasets are presented here: TIMIT [63], NTIMIT [64], Aishell [65,66], LibriSpeech [67], Voxceleb1/2 [62,68], YOHO [69], and CSTR VCTK [70], and their details are shown in Table 2 below.

Table 2. Generic datasets for speaker recognition.

Datasets	Sample Rate	Data Size	Spk Num	Language	Text Dependency	Condition
TIMIT	16 kHz	6300 sentences	630	English	TI	Clean
NTIMIT	8 kHz	6300 sentences	630	English	TI	Telephone line
Aishell	16 kHz	178 h	400	Chinese	TI	No noise
LibriSpeech	16 kHz	153,516 utterances	>9000	English	TI	/
VoxCeleb1	16 kHz	1,128,246 utterances	1251	English	TI	Multi-media
VoxCeleb2	-	100 w sentences	6112	Multilingual	TI	Multi-media
YOHO	8 kHz	5500 phrases	138	English	TD	Office
CSTR VCTK	48 kHz	1000 sentences	30	English	TD	Wild

- TIMIT: The standard dataset in the field of speech recognition is a relatively small dataset that enables the training and testing of models in a short period of time, and its database is manually annotated down to the phoneme, with speakers from all parts of the United States, and provides detailed speaker information, such as ethnicity, education, and even height.
- NTIMIT: The dataset that puts the audio data in TIMIT on a different telephone line for transmission and then reception is a dataset created to implement voice recognition in the telephone network.

- Aishell: Aishell-1 is the first large data volume Chinese dataset, with 178 h of speech, 400 speakers, 340 people in the training set, 20 people in the test set, and 40 people in the validation set, each of whom speaks about 300 sentences. Aishell-2 expands the data volume to 1000 h of speech, with 1991 speakers, each of whom speaks 500 sentences. The words spoken by each person may be repeated.
- LibriSpeech: The dataset is a large corpus containing approximately 1000 h of English speech. The data come from the audiobook recordings read by different readers of the LibriVox project, organized according to the sections of the audiobooks. It is segmented and correctly aligned.
- Voxceleb1,Voxceleb2: Two speaker recognition datasets without intersection, both of which are obtained from open source video sites captured by a set of fully automated programs based on computer vision technology development. They differ in size, with VoxCeleb2 compensating for the lack of ethnic diversity in VoxCeleb1 by being five times larger than VoxCeleb1 in terms of data size.
- YOHO: A speech dataset collected in an office environment that is text dependent, where the speaker speaks in a restricted textual combination.
- CSTR VCTK: A dataset including noisy and non-noisy speech with a sampling rate of 48 kHz and in which the speaker is accented.

3. Adversarial Attack

Adversarial attacks against SRSs use small perturbations that are imperceptible to human ears to mislead the system to wrong decisions. In this section, we will describe the adversarial attacks in deep networks and different methods of the attacks against SRSs in detail.

3.1. Overview of Adversarial Attack

Adversarial examples are aggressive data that make DNN models confused. As revealed in recent works, DNNs in different fields, such as image recognition, object detection, sentiment recognition and speech recognition, are susceptible to adversarial attacks. There are also adversarial attacks against SRSs. Figure 2 shows the general framework of adversarial attacks against SRSs.

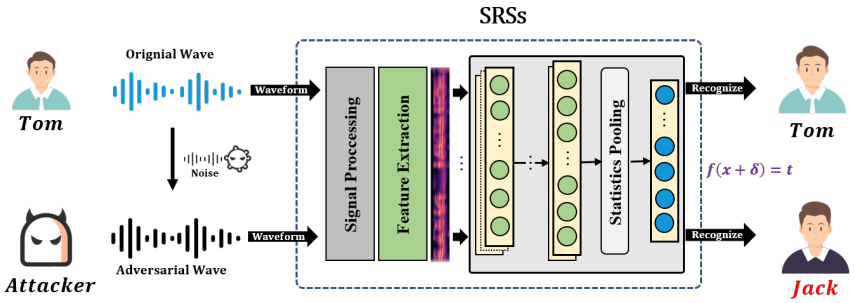

Figure 2. General framework of adversarial attacks against SRSs.

Formally, an adversarial audio can be defined as follows:

$$x' = x + \delta, \quad s.t. \; ||\delta||_p < \epsilon \tag{4}$$

where x is the original audio, δ is the perturbation introduced to the audio (as shown in Figure 3a), and x' is the adversarial audio, which can make the SRSs misjudge the original speaker y as y' (y' can be any random speaker ID other than the original speaker or a speaker ID specified by the attacker). Figure 3b illustrates a didactic example of inserting a perturbation δ into a legitimate audio x on a 2D space.

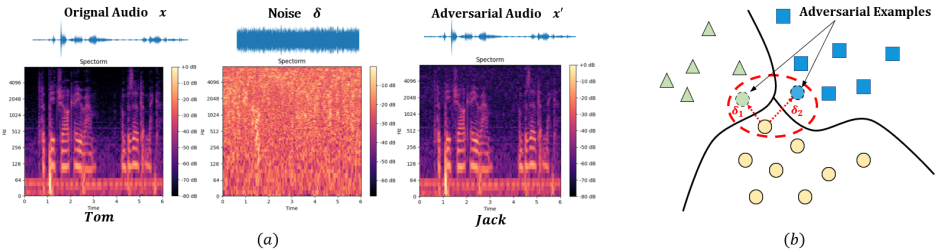

Figure 3. A sample of adversarial example generated from inserting a perturbation δ into a legitimate audio x on a 2D space. Figure (**a**) shows the waveform and mel-spectrogram of the original audio, noise and adversarial audio. Figure (**b**) denotes the presentation of dyadic samples on the data distribution.

To make the attack effective, we propose the following three constraints:

- x' must be within a proper range such that the waveform can be recovered into an audio;
- δ must be as small as possible;
- SRSs will identify x' as the special target specified by the attacker beforehand (it can also be any other random target, but this is not meaningful).

Due to the strong similarity between image processing and audio processing tasks in deep models, existing works mainly transfer the advanced attack methods in image domain to SR. Table 3 shows the specifics of attack methods, and we systematically classify current adversarial attacks against SRs in terms of attack targets, attack strategies, etc.

Table 3. Related works on adversarial attacks against existing SRSs.

Methods	Target	Capability	Knowledge	Generate Strategy	Perturbation Object	Metrics	OTA	Victim Model	Corpus
SEC4SR [58]	Both	Individual	White Black	Gradient Sign Optimization Evolutionary	Mel-Spec	ASR/SNR PESQ	Digital Physical	AudioNet GMM i-vecor x-vector	LibriSpeech
Kreuk [71]	Untarget	Individual	White Black	Gradient Sign	Mel-Spec MFCC	ACC	Digital	End-to-end	YOHO NTIMIT
Abdullah [72]	Target	Individual	Black	Audio process	MFCC	ASR	Digital	Microsoft Azure	-
Li [73]	Target	Individual	White Black	Gradient Sign	LPMS MFCC	EER	Digital	i-Vector x-vector	VoxCeleb1
Villalba [74]	Both	Individual	White Black	Optimization	MFCC	EER/SNR minDCF PESQ	Digital	ResNet34 ThinResNet34 TDNN	Voxceleb1&2
Jati [75]	Both	Individual	White	Gradient Sign	Mel-Spec STFT	ASR	Digital	1D-CNN TDNN	LibriSpeech
Joshi [76]	Both	Universal	White	Gradient Sign Optimization	Fbank	ACC	Digital	ResNet34 Transformer x-vector	LibriSpeech VoxCeleb
Two-step [77]	Target	Universal	White	Optimization	-	ASR/WER CER/SNR	Digital Physical	VGG Thin-ResNet-34 Fast-ResNet	LibriSpeech Voxceleb2
Liu [78]	Target	Individual	White	Gradient Sign	Mel-Spec	EER min-tDCF	Digital	LCNN-Big LCNN-Small SeNet	ASVspoof 2019
MI-FGSM [79]	Target	Individual	Black	Gradient Sign	log-power magnitude spectrum	ASR	Digital	LCNN/AFNet SENet50 ResNet34	ASVspoof 2019
Quasi [80]	Both	Individual	White	Optimization	MFCC	EER	Digital	GMM i-vector	Voxceleb1
FakeBob [81]	Both	Individual	Black	Evolutionary	PLP MFCC	ACC SNR	Digital Physical	GMM-UBM i-vector x-vector	LibriSpeech
Li [82]	Both	Individual	White	Optimization	Waveform	ASR/SNR PESQ	Digital	SincNet	TIMIT

Table 3. *Cont.*

Methods	Target	Capability	Knowledge	Generate Strategy	Perturbation Object	Metrics	OTA	Victim Model	Corpus
GE2E [83]	Target	Individual	White	Optimization	Feats	SR/SNR MNR	Digital	d-vector	TIMIT
Dictionary [84]	Target	Individual	White	Dictionary	Mel-Spec	SR	Digital	VGGvox	VoxCeleb2
VMask [85]	Target	Individual	Grey Black	Gradient Sign Optimization	Mel-Spec	WER/SER SNR	Digital Physical	VGGVox	LibriSpeech
Abdullah [86]	Untarget	Individual	Black	Feature Process	MFCC	ASR	Digital	End-to-end	TIMIT LibriSpeech
Siren [87]	Target	Individual	White Black	Evolutionary	MFCC	ASR SNR	Digital	End-to-end	VCTK IEMOCAP
AdvPulse [88]	Target	Individual	White	Optimization	MFCC	ASR	Physical	x-vector	VCTK
Xie [89]	Both	Universal	White	Gradient Sign	MFCC	ASR	Simulated	x-vector	VCTK
AS2T [90]	Both	Individual	White Black	Gradient Sign Optimization	Waveform	ASR/SNR PESQ/L_2	Digital Physical	Open source SRSs	LibriSpeech
Occam [91]	Target	Individual	Black	Optimization	-	ASR/SNR	API	Commercial SRSs	LibriSpeech
Li [92]	Both	Individual	White	Gradient Sign Optimization	Waveform	ACC ASR	Digital Physical	x-vector	VCTK
UAPG [93]	Target	Universal	White	Optimization	MFCC	FR/SR	Digital	x-vector	VCTK
Xie [94]	Both	Universal	White	Gradient Sign	MFCC	ASR	Simulated	x-vector	VCTK
NRI-FGSM [95]	Target	Individual	Black	Gradient Sign	Waveform	ASR/SNR PESQ/L_2	Digital	x-vector ECAPA	LibriSpeech
FoolHD [96]	Target	Individual	White	Optimization	MFCC	ASR/JND PESQ	Digital	-	Voxceleb
Inaudible [97]	Target	Individual	White	Gradient Sign	Waveform	ASR	Digital	x-vector	Aishell-1
UAPs [98]	Both	Universal	White	Gradient Sign	Waveform	SER/PTR SNR PESQ	Digital	End-to-end	TIMIT LibriSpeech

Specifically, as shown in Figure 4, existing works can be divided into two dimension-sadversarial example generation and example evaluation. Firstly, for example generation, the attack strategies designed by attackers differ for identification tasks with different target models (e.g., CSI, OSI, and SV) (detailed in Section 3.2), and also differ under different attack targets (targeted or untargeted) (detailed in Section 3.2). To the best of our knowledge, the a priori knowledge of the internal structure of the victim model is the key factor to consider when an attacker launches attacks. We classify the attacks into white-box, grey-box and black-box, according to whether the attacker has the internal information of the SRSs, including model structure, parameters, loss function and model gradient, etc. In general, due to the ability to grasp all the information of the victim model, it is easy to launch a successful white-box attack, but this is hardly the case in real-world scenarios. In contrast, its grey-box and black-box counterparts that are more challenging to launch are better aligned with real-world situations (detailed in Section 3.3). In addition, an adversarial perturbation with strong aggressiveness is necessary, so in the adversarial example generation stage, we need to discuss whether the adversarial example generated by attackers can be generalized. By the generalization capacity of the adversarial example, we divide the attacks into individual and universal. In individual attacks, the attacker must generate perturbations specific to each genuine sample, which reduces the attack efficiency (all attack methods described in Section 3.3 are individual attacks). We discuss the more practical attacks: over-the-air and commercial SRSs in Section 3.4. In universal attacks, however, the attacker only needs to introduce a universal adversarial perturbation (UAP) obtained by pretraining into the clean samples to fool the SRSs. Compared with individual attacks, universal attacks have considerably improved the attack efficiency, but the cost for generating a UAP can be enormous (detailed in Section 3.5). Adversarial attacks against SR models are different from those against image- or text-processing models: as SR models can be trained by original audio signals or frequency features, perturbations against SR

models can be divided into the time domain and frequency domain. As the name suggests, time-domain perturbations are perturbations introduced to the sampling value of the original audio, whereas the frequency-domain perturbations are perturbations introduced to acoustic features, such as MFCC (detailed in Section 3.6). A proper perturbation created as per the model architecture can considerably improve the attack success rate.

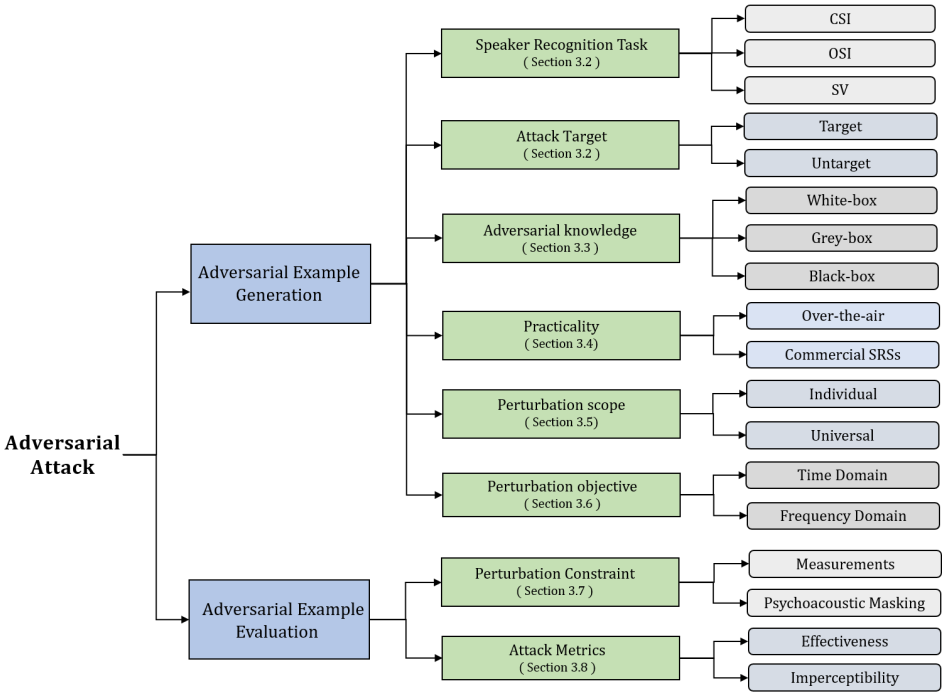

Figure 4. Taxonomy of adversarial audios in SRSs.

In terms of adversarial example evaluation, it is important to limit the size of the perturbation as small as possible since perturbations in audio are easier perceived, compared to images and texts. Methods to introduce perturbations can be divided by the perturbation constraint into perturbation regularization and psychoacoustic masking (detailed in Section 3.7). In Section 3.8, we elaborate on the evaluation metrics for adversarial attacks.

3.2. Adversarial Task

It is essential to specify which kind of speaker recognition tasks the target model own before attacking, and designing special attack algorithms for different task models can usually improve the success rate of the attack.

- CSI: As mentioned in Section 2.2.1, close-set speaker recognition is a simple classification task and involves no thresholds. Thus, how to make the confidence coefficient of the decoded adversarial audio skew toward the target label is the key to attacks against the CSI models. To transfer adversarial attack algorithms from the field of image processing is a good choice.
- OSI: Different from CSI models, the OSI model uses the back-end scoring module to obtain a decision threshold, which is relied on to make the final judgments. If the internal structure and threshold are known, the perturbation can be scaled up to increase the attack intensity; if the internal parameters of the model are unavailable, how to identify the model decision threshold is a challenge in OSI attack tasks.
- SV: In adversarial attacks against SV models, we need only simulate the voice features of a single speaker to make the model score bigger than the threshold. Attacks against SV models are easier than attacks against OSI systems; however, for models whose internal parameters are unknown, the internal structure and decision threshold of the model should be considered.

After identifying the type of task, we need then identify whether it is a targeted attack or an untargeted attack. In untargeted attacks, the attacker only needs to fool the SRSs to generate a wrong result. In targeted attacks, the attacker needs not only to fool the SRSs toward a wrong decision, but to make it generate a specific identification result. The next section provides a systematic introduction to different types of adversarial attacks.

3.3. Adversarial Knowledge

The attacker's mastery of the a priori knowledge of the attack target has a significant impact on the efficiency and success rate of his attack. Therefore, the attack scenarios can be classified as white-box, grey-box and black-box according to the different degrees of the attacker's mastery of the a priori knowledge of the SRSs. Additionally, as per whether the generated perturbation is applicable to different target audios, we can divide the attacks into individual attacks and universal attacks. In this section, we discuss the individual attack under the three attack scenarios as it is the mainstream attack, and universal attacks are discussed in Section 3.5.

3.3.1. White-Box

If the attacker knows all the information of the model, including the model architecture, parameters, loss function, activation function, input and output data, or even the embedded defense strategies, the attack is termed a white-box attack. By the basis of the attack, the attacks can be divided into gradient-sign-based attacks and optimization-based attacks.

(1) Gradient-based attack.

As shown in Figure 5, the deep neural network is trained by the gradient descent method to minimize the target loss function, and attackers against this type of network model needs only maximize the loss function along the direction of the gradient ascent. This is the underlying principle of gradient sign-based attacks, which generate adversarial perturbations based on the gradient information of the model and perform iteration through the gradient of the input to maximize the loss function and reduce the model's recognition accuracy. Common gradient-sign-based algorithms include fast gradient sign methods (FGSM), random fast gradient sign method (R-FGSM), iterative fast gradient sign methods (I-FGSM), project gradient descent (PGD), and momentum iterative fast gradient sign methods (MI-FGSM).

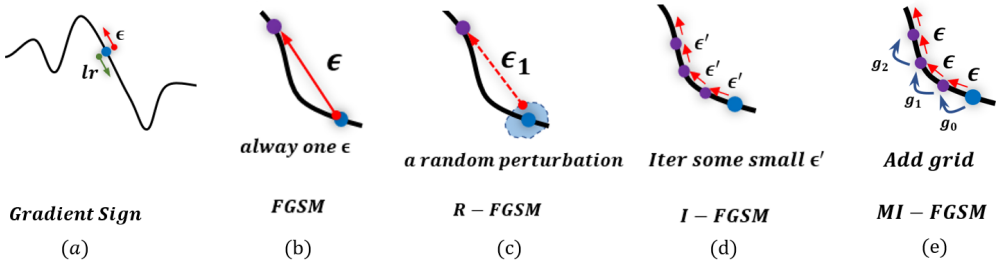

Figure 5. Gradient-based adversarial example generation algorithms. Figure (**a**) denotes the the gradient direction of the model training and adversarial updating. Figure (**b**) shows a single-step attack, and Figure (**c**) precedes the single-step attack with a random perturbation. Figure (**d**) shows a typical multi-step iterative attack, taking one small step in each step, and in figure (**e**), momentum is accumulated at each iteration step.

As Figure 5 shows, the gradient-based methods employ the internal gradient of the model, and move L steps toward the sign of the largest losses. Figure 5b,c aims to identify the proper moving steps to maximize the impacts on the target model. Among them,

FGSM [29] aims at single-step attack methods that take one step ϵ in the direction of maximum deviation, with a fixed value of ϵ,

$$x' = x + \epsilon \cdot sign(\nabla_x L(f(x), y)) \tag{5}$$

where the function $f(x)$ is the encoded network of the speaker feature vectors; $L(\cdot)$ is the loss function, which is usually the cross-entropy loss. y is the real label of clean samples. Perturbations are ensured to be undetectable by restricting $||x' - x||_\infty \leq \epsilon$. In other words, a larger ϵ means a more effective attack (reducing the model's recognition efficiency), but it also means that the perturbation is more likely to be detected.

Moving a fixed length of step each time seems to fail to work in face of gradient masking defense strategies. Tramèr et al. [99] proposed a method to introduce a random perturbation before each step:

$$\hat{x} = x + \alpha \cdot sign(N(0, I)) \tag{6}$$

$$x' = x + (\epsilon - \alpha) \cdot sign(\nabla_x L(f(x), y)) \tag{7}$$

where $0 < \alpha < \epsilon$. This simple method increases the robustness of attacks, but it is still a single-step attack. Wang et al. [100] put forward the iterative fast gradient sign method (I-FGSM) or the basic iterative method (BIM), which uses a small iteration step length σ along the gradient direction:

$$x'_{i+1} = x + clip_\epsilon(x'_i + \sigma \cdot sign(\nabla_x L(f(x), y) - x) \tag{8}$$

where $x'_0 = x$, i is the iterative step of optimization, the function $clip(\cdot)$ ensures that the $L\infty$ norm of the perturbation stays below ϵ after each time of optimization. Experiments prove that this attack is more aggressive in practical applications. In addition, the project gradient descent (PGD) method [101] generalizes for the L_P norm of the BIM, which is usually L_1, L_2 and L_∞,

$$x'_{i+1} = x + P_{p,\varepsilon}(x'_i + \sigma \cdot sign(\nabla_x L(f(x), y) - x), \tag{9}$$

where $P_{p,\varepsilon}$ is the projection operator of L_P. The PGD attack can also consider several options for taking a random initialization for the perturbation θ and using the one that produces the largest loss.

Since the I-FGSM multi-step iteration does not consider the effect of the current gradient on the perturbation, Dong et al. [102] used the sum of the upper previous gradients as momentum and improved the I-FGSM method by adding this momentum term to the optimization process. The method is called the momentum iterative fast gradient sign method (MI-FGSM), as described in Equation (10):

$$g_{i+1} = \mu \cdot g_i + \frac{\nabla_x L(f(x), y)}{||\nabla_x L(f(x), y)||_1} \tag{10}$$

$$x'_{i+1} = clip_\varepsilon\{x'_i + \sigma \cdot sign(g_{i+1})\} \tag{11}$$

where g_i gathers the gradient of the previous i iterations with the momentum decay factor $\mu = 0$. From Equation (10), we can see that the MI-FGSM will turn into I-FGSM when $\mu = 0$.

Kreuk et al. [71] introduced perturbations onto the MFCC features by the FGSM, and reconstructed the speaker features into acoustic waveforms, which achieved an ASR of 90% against an end-to-end DNN-based SAV system. They also analyzed the attack performance on different datasets and under different features (to the best of our knowledge, different speaker features can affect the model's accuracy). This is the first work to demonstrate the existence of adversarial attacks for SRSs, but their work does not consider the size of the incorporated perturbations, which may make the noise too loud to be perceived by the human ear. To achieve attacks on advanced SRSs, Li et al. [73] successfully attacked the

GMM i-vector model (a Kaldi-based system) by FGSM. To explore the transferability of the adversarial examples, they transferred the attack against the x-vector model by adversarial examples generated by the GMM i-vector model under different speaker features. Their experiments showed that the transfer attack worked better under the same model with different features, but had weaker performance when both the model and feature vary.

Both the basic iterative method (BIM) [100] and the projected gradient method (PGD) [101] can be considered multi-step iterative FGSM algorithms. Specifically, the BIM (also known as multi-step iterative fast gradient notation I-FGSM) optimizes one tiny step at a time in the gradient direction, and the optimized perturbation parametrization will be limited to a small range during the iterative process.The PGD method, on the other hand, optimizes the BIM perturbation parametrization limit by projecting the optimized perturbation to an arbitrary L_p parametrization (BIM is equivalent to PGD-L_∞), while PGD adds a random perturbation in the initialization stage to make the loss maximized (in BIM, the initial value of the perturbation is 0). Although these two methods generate more aggressive adversarial examples than FGSM, they require more training overhead. To clarify the effectiveness of the above methods, refs. [74–78] investigated various algorithms, such as FGSM, R-FGSM, BIM and PGD for several x-vector systems with different structures, and showed through experimental data that the multi-step iterative I-FGSM method has better attack performance. At the same time, as the number of iterations increases, the generated adversarial examples are more aggressive and can often achieve an attack success rate of 100%. Nonetheless, increased iterations also mean more computing resources and a lower quality of the generated audio.

The momentum iterative fast gradient notation (MI-FGSM) achieves accelerated counteracting perturbation generation by accumulating the acceleration vector in the direction of the increasing gradient of the loss function during the iterative process, and this method is able to avoid local maxima. Zhang et al. [79] integrated MI-FGSM with the iterative ensemble method (IEM) and reached an ASR 100% in attacks against white-box models. They also conducted transfer experiments of adversarial examples, in which the adversarial examples generated by a known model could successfully attack other black-box models, with a cross-model ASR as high as 84%. Their experiments also revealed that larger adversarial perturbations meant stronger transferability of the attacks, and the ϵ of the perturbation should be at least larger than 2.5 to achieve successful attacks. In fact, the ASR of black-box attacks across models varies considerably from 24% to 84%, which may be related to the structure of the victim models; in addition, the perturbations generated by their method are very likely to be perceived by human ears (to be detailed in Section 4.4).

(2) Optimization-based attack

While the gradient-sign-based method configures the attack algorithm based on the internal gradient of the model, the optimization-based algorithm does not require knowledge of the model gradient, and the attack can be performed by obtaining the logits of the model output. As shown in Figure 6, the core idea is to define the adversarial sample generation process as an optimization problem that minimizes the perturbation while successfully misleading the model recognition as a target label to achieve an imperceptible effect to the human ear. Therefore, this approach usually has two optimization goals: higher attack success rate and smaller perturbation. We can design an optimization model or use the idea of C&W to minimize the loss of the above two objectives.

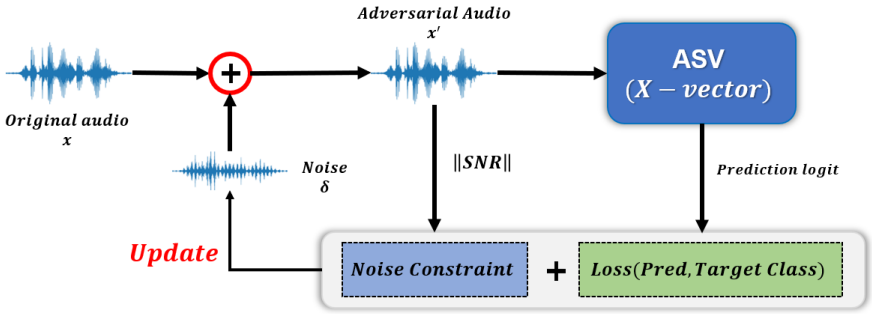

Figure 6. Diagram of optimization-based attacks.

The C&W approach performs the attack by finding the minimum perturbation δ, which deceives the neural network model while keeping the perturbation imperceptible, and δ is obtained by minimizing the loss:

$$C(\delta) = D(x, x+\delta) + c \cdot f(x+\delta) \tag{12}$$

where $D(\cdot)$ is the distance metric that constrains the size of perturbation size to achieve imperceptibility, which can usually be done with the L_2 parametrization. Different from perturbations to images, audio perturbations can be constrained from two perspectives: the original speech signals and the speaker features (detailed in Section 3.6). The function $f(\cdot)$ is defined as a standard where only when $f(x+\delta) \leq 0$ will the attack be considered successful. The SR model can be regarded as a classification model, and $f(\cdot)$ can be defined as follows:

$$f(x+\delta) = max(max\{Z(x')_{i:i\neq t}\} - Z(x')_t, -\kappa) \tag{13}$$

where $Z(x')_i$ is the logit value that the model predicts x' as class i and κ is the attack confidence hyperparameter. $f(x+\delta) \leq 0$ is an indication of a successful attack. This condition is satisfied when at least one of the logit of the attack target class is larger than the normal class by κ. We can increase the confidence in the success of the attack by setting $\kappa > 0$. The weights c balance the function $D(\cdot)$ and the function $f(\cdot)$.

In optimization-based models, to make the perturbations more covert, an attack network model G is designed to generate adversarial examples, and the SRSs prediction results and the loss of target labels are used to reversely optimize the parameters in the G network and minimize the perturbations.

Carlini et al. [30] proposed the C&W algorithm for image processing models, and applied the algorithm for the generation of adversarial examples against automatic speech recognition systems. Specifically, they optimized and modified the waveform of the original audio to generate adversarial examples, which proved the feasibility of the C&W algorithm in the field of audio processing. Some other researchers [58,74,76,77] constructed the loss function following the C&W framework, and achieved an ASR as high as 100% on targeted and untargeted attacks. A second-order proposed Newtonian attack method was proposed in [80] to solve the optimization problem based on different level approximations of Taylor expansion. The method generates a reduced relative perturbation size ρ from 9.91 to 0.24 ($\rho = \frac{||\delta||_p}{||x||_p}$, where x is the clean audio and δ is the counter perturbation) compared to the relative perturbation size in [73]. To further reduce the perturbation size, an attack network was trained in [81] to add interference to the input speech by training a lightweight attacker network. Based on this optimization idea of building a model, in the latest work [103], a FoolHD approach was proposed to generate and hide the adversarial perturbations in the original audio file using a gated convolutional autoencoder GCA, trained in the MDCT domain by a multi-objective loss function with target label probabilities and feature differences, which, compared to the attack network of [82], generates an adversarial audio PESQ value (perceptual evaluation of speech quality) was able to improve from 3.48 to 4.3,

which is a very good result. Luo et al. [83] proposed an attack based on a generalized end-to-end (GE2E) loss function for the SR model with end-to-end d-vectors. They designed a novel loss function to construct a generator that uses generalized loss to reduce the distance between the perturbed audio and the target speaker and limit the perturbation amplitude to construct a multi-factor attack strategy that generates effective adversarial examples under minor perturbations and is able to achieve an attack success rate of 82%. However, this attack is limited to the SR model using triple loss and GE2E loss.

There are other works that do not follow the idea of the C&W attack. For instance, Marras et al. [84] shifted to dictionary attacks against SRSs. Dictionary attacks can deal with large amounts of attackers and dispense with the need of knowledge about the speaker features or speech models of the target speaker. In this type of attack, perturbations are added to the master voice to maximize the similarity between the master voice's spectrogram and that of most speakers. When the spectrogram similarity degree exceeds the threshold and the master voice approaches the voices of most speakers in the crowd, the spectrogram is reversed to generate time-domain waveforms for close matching with multiple speakers in the crowd.

The above-mentioned works are perturbed for the whole audio and are static attacks. To achieve a real-time streaming attack, Li et al. [85] proposed a subsecond, synchronization-free and targeted adversarial perturbation, AdvPulse. Using an optimization-based algorithm that maximizes the expected output probability of the target class under different delay conditions, the identification of the streaming audio input is changed in a targeted and synchronization-free manner, without modifying the entire audio input, at any point of the streaming audio. Add a very short hostile interference (0.5 s) to launch a targeted counter attack. For the physical domain SRSs, the attack success rate is able to reach 89.2%.

The white-box attacks mentioned above show extent advantages when the prior knowledge of the model is available, implying that the SRSs will be completely exposed to the attacker. The white-box attack approach is an important basis for our research, but has limitations in practical attacks.

3.3.2. Grey-Box Attack

In grey-box scenarios, the adversary only knows the scores and recognition results output from the SRSs, and has no detailed information about the speaker verification system, much less whether there is a defense mechanism for the model. Existing solutions to the problem of unknown gradients in grey-box models include difference gradient estimation and natural evolution.

Specifically, in order to implement the acoustic adversarial attack without any knowledge of the model structure and parameters, Zhang et al. [85] proposed an adversarial sample generation strategy called VMask, which estimates the gradient based on the difference in similarity scores of multiple queries and uses zero-order optimization [104] to solve the gradient unknowability problem, while using psychoacoustic masking (described in detail in Section 3.7.2) to make the perturbation imperceptible. Their approach is able to achieve a successful attack in the VGGVox model with the guarantee that the adversarial audio and the original audio have exactly the same transcription. However, this attack is performed with the knowledge of the model's thresholds, which are not normally given by the SRSs. To further estimate the thresholds in SRSs, Chen et al. [81] proposed a FakeBob's attack which estimates the gradient by a natural evolutionary strategy [103] and also estimates the thresholds for SV and OSI tasks using a circular iterative algorithm. Their attack starts from the original input speech rather than from a randomly perturbed speech and uses an early stopping strategy to reduce the number of queries, i.e., stop searching once an adversarial sample is found. Similar to the C&W attack, FakeBob also provides an option to control the confidence of the adversarial examples by parameter κ.

3.3.3. Black-Box Attack

The black-box attack scenario is not like the white-box and grey-box ones, which can obtain certain information inside the model. We define the black box as the attacker only being able to obtain the decision result of the model, i.e., the decision result is "yes" and "no" in the SV system, and "speaker id" in the CSI or OSI system. The confidence of the decision is also unavailable. This is the most difficult existing attack scenario, but at the same time, it is also more in line with the realistic attack scenario. This type of attack can be divided into the data-level and model-level by the area of attack.

From data-level attacks, Abdullah et al. [86] was able to successfully attack some mainstream SRSs by performing inverse Fourier transform on MFCC features in the signal processing layer and time-domain compression on the original audio, and this method does not require a priori knowledge of the internal structure and parameters of the model. However, the adversarial audio generated by the above method is a piece of noise, and this attack is highly detectable in a real attack scenario. In their latest work, [72] proposed a spectrum estimation method based on discrete Fourier transform (DFT) and singular spectrum analysis (SSA), which introduces perturbations to single phonemes of imperceptible low-frequency audio in the original audio input (one phoneme is perturbed every eight words). This method will not reduce the quality of the original audio file, and can make successful attacks, even when there are transcoding, jitters, or packet dropouts in network transfer. As data-level attacks focus on the signal processing and feature extraction stages before the deep neural network model and do not interfere with the model, such attacks have strong transferability.

Model-level black-box attacks mainly employ substitution models and genetic algorithms [85]. As all existing SR models project a high-dimensional audio space onto a low-dimensional speaker space, we can train a local deep learning SRS to mimic the victim black-box system. The attacker can extract speaker feature vectors of the victim speaker from recordings of any random speeches from the victim speaker, and add slight noises to the audio of any other speakers to generate an attack audio that represent the victim's voice [71,73,74,78]. As the features learned from the high-dimensional space are similar, adversarial examples generated in white-box models can be used to attack black-box systems, but the ASR is below 40% across databases or across models. FakeBob proposed by Chen et al. [81] has also been applied to black-box attacks, and the major approach is to transfer adversarial examples that have successfully attacked the grey-box system. Recently, CC-CMA-ES proposed by [91] applied a cooperative co-evolution (CC) framework to the powerful covariance matrix adaptation evolution strategy (CMA-ES) to solve the large and complex problem in the strictly black-box setting. Additionally, they adopt gradient inversion method to attack

In addition to the idea of using substitution models, SirenAttack [87] used a gradient-free particle swarm optimization (PSO) method [74] to search for adversarial examples. PSO is an algorithm that mimics the food-searching behaviors of bird flocks; in PSO, the solution to each optimization problem is considered a particle in the search space, and each particle is a candidate solution, and the weighted linear combination of the inertia, individual optimum and local optimum are iteratively updated to converge into a global optimum and generate adversarial examples. Most existing attacks to black-box models are realized by sample or model transfer from white-box or grey-box attacks, which have considerable limitations and will become invalid when the model or database is changed. Though attacks based on natural optimization and genetic algorithms are effective, the noises are rather obvious, with the signal-to-noise ratio (SNR) staying between 6 and 12 dB, which is not applicable to real-world physical domain attacks. In the future, physical-domain black-box attacks will be the focus and major challenge in SRS adversarial attacks.

3.4. Practicality

The adversarial attacks mentioned above seem to show powerful attack capabilities on mainstream CSI, OSI and SV models, but do such attack capabilities really exist? In

this section, we focus on whether existing adversarial attacks are practically relevant, whether they achieve powerful attacks in over-the-air environments, whether they have good transferability, and whether they can successfully attack commercial systems.

3.4.1. Over-the-Air

Adversarial examples have shown powerful attack capabilities in the digital domain, but when applied to actual smart devices, their attack performance is greatly reduced. In the digital domain, SRS reads the voice information directly from the wav file, which is an almost non-destructive operation, whereas the attack in the physical domain is a lossy process. The audio is first converted into an analog signal by a digital signal, which is undoubtedly a lossy process. Then the analog signal will be propagated in the air, will be affected by environmental noise, reverberation and media transmission attenuation, and the audio quality will be significantly reduced. This leads to adversarial audio in over-the-air scenarios being very difficult to attack. Some works simulate over-the-air attacks by applying the room impulse response (RIR), natural noise and signal attenuation, such as [77,81,88–90]. As shown in Table 4, we consider the attack task, attack target, the distance of the audio transmission, ASR, and whether the victim model has a defense module. However, it is challenge to conduct a through evaluation of adversarial attacks over-the-air in the physical world.

Table 4. Related works on adversarial attacks for physical domain.

Ref.	Task	UT/T	OTA	Distance (m)	Methods	ASR (%)	Generation Model	Victim model	Commercial System	Defense	Time
[77]	SV	T	Real	1.7	RIR	67.7	Res34-V	Res34-V	-	RD	-
[81]	OSI/SV	T	Real	0.25–8	NES+RIR	100	GMM	-	laptop (Dell) Shinco OPPO, JBL iphone 6 plus	LS, QT AS, TDD	-
[86]	CSI/SV	T	Real	0.3	TDI, RPG, HFA, TS	100	-	-	Azure	VAD	-
[88]	SI	T	Real	1.6∼3	RIR+BPF	96.9	?	-	TKGOU Honda	-	-
[89]	CSI	T	Real	-	RIR	90.19	x-vector (Tensorflow)	x-vector (Kaldi)	-	-	0.015 s
[90]	CSI	UT	Sim. Real	-	RIR+RN	97.4	?	?	-	-	-
[91]	SI/SV	T	Real	0.15	GI	71.7	ASplRE	-	Google Assistant Siri, iFlytek Cortana Amazon Echo	-	-
[92]	SV	UT+T	Real	-	GA+RIR	50	x-vector	-	Talentedsoft Azure	-	-

UT/T: Untarget and target attack; SI: Unclear whether it is CSI or OSI; OTA: Over-the-air; Sim.: Simulate; RIR: Room impulse response; RN: Random noise; ASR: Attack success rate; Generation Model: Crafting adversarial examples from generation model; RD: Replay attack detection; LS: Local smoothing; QT: Quantization; AS: Average smoothing; TDD: Temporal dependency detection; GA: genetic algorithm; ?: Indicates mentioned in the literature but not clear.

3.4.2. Commercial SRSs

It is worth noting that most of the attacks we have discussed so far are on academic research SRSs, while SRSs used in industry consider more assumptions compared to academic SRSs, with some advanced functional modules for denoising and false audio detection deployed externally. There is currently very little work attacking actual commercial SRSs. The current attacked commercial systems contain three types: voice assistant (e.g., OPPO, and Google Assistant), smart home (e.g., JBL, and Amazon Echo) and smart car (e.g., Honda Civic Sedan).

3.5. Universal Attack

The universal attack referred to in previous work focuses on data-agnostic. In this paper, we define the universal attack as two aspects: data-agnostic and model-agnostic.

3.5.1. Data-Agnostic

The perturbation strength is also an indicator for the quality of attacks. As Table 1 shows, individual attacks in which specific perturbations are generated to each clean audio are the dominating research direction at present. Universal attacks that use universal adversarial perturbations, however, are effective on most samples, and hence are more harmful than individual attacks. Attacks based on universal adversarial perturbations (UAPs) do not require any prior knowledge of the target model in the testing stage, but are strongly invasive with a single perturbation, which lowers the bar of adversarial attacks and are likely to gain popularity among attackers. UAPs have received broad attention from researchers on speaker recognition and verification, but there are few studies in this regard. Li et al. [93] proposed a generative network to learn the mapping from the low-dimensional normal distribution to the UAPs high-dimensional subspace, synthesized 3200-dimension UAPs using 100-dimension noise inputs that conformed to standard normal distribution through multiple convolution modules, and embedded the synthesized UAPs after scaling into any input signals to fool well-trained speaker recognition models with high probability; the UAPs achieved an ASR of 97% in untargeted attacks, with a mean SNR of 49.87 and a PESQ of 3.0 (generally, the closer the PESQ value approaches 4.5, the better the quality of the audio). However, they only applied the UAPs to digital attacks, but did not analyze attacks in real-world physical scenarios.

To achieve attacks in physical scenarios, Xie et al. [94] put forward a real-time and universal attack method applicable to the physical domain. To make the UAPs fit voice inputs with different lengths, they first generated a small fixed length unit universal noise, and then built the desired length of adversarial perturbations on the top of this via repeated playback to generate the adversarial utterance. Meanwhile, the magnitude of perturbations was adaptively adjusted via spectral gating to make the attack more imperceptible. They also introduced the room impulse response (RIR) [105] loss to the primary target loss function, and simulated magnitude loss of the audio in the physical world to achieve attacks in physical environments, but only achieved an ASR of 90.19% and 90.32% in simulated indoor environments. Their method is able to not only launch effective attacks, but reduces the time cost for generation of adversarial audios. In their latest work [106], they designed a fast attack perturbation generator (FAPG) and a universal attack perturbation generator (UAPG), which can make real-time perturbations on any clean samples and lead the target model to misclassification.

3.5.2. Model-Agnostic

Model-agnostic means that adversarial examples can have strong attack performance in multiple models, regardless of whether the attacker has prior knowledge of those models, which is also called transfer attack in other works. For data-agnostic, the universal perturbation is aimed at the data level and cannot perform well in other models.

The model-agnostic adversarial attack is a commonly adopted method for black-box attacks, which generates adversarial audio from a white-box model first and then uses this audio to attack a black-box model. For example, some gradient-based white-box attack methods were utilized to attack black SRSs [73,74] to explore the transferability of the adversarial examples. However, the gradient-based attack method greedily perturbs the audios in line with the direction of the sign of the gradient at each iteration, which may easily fall into the local maximum and fail to attain global optimization, resulting in low performance in SRSs. A multi-step iterative attack method called NRI-FGSM [95], proposed to improve the attack success rate and achieve global optimization for the black-box SRS, which represents the Nesterov accelerate gradient (NAG) and root mean squared propagation (RMSProp) optimization-based iterative-fast gradient sign method. Compared

to the traditional gradient-based methods, NAG can take a step forward and meanwhile stabilize the direction of the gradient, which will correct the previously accumulated gradient and thus avoid the local maximum. The RMSProp optimization method with adaptive step size and momentum was used to optimize the step size dynamically. NMI-FGSM-tri proposed by [107] can craft strong transferable adversarial examples to achieve the attack on the black-box model. Specifically, they used ensemble ideology and NAG to enhance the transferability of adversarial examples to improve the attack performance of adversarial examples in the target system. At the same time, they found that the feature distribution of audio in different models has certain similarities, and used a few query attacks on the target model to monitor and correct the target speaker of the attack.

3.6. Perturbation Object

Besides prior knowledge of the victim model and adversarial audio generative methods, the attacker also needs to identify the perturbation object, i.e., where to add the perturbations. In the field of image processing, perturbations are added to pixels; in text processing, words are added, deleted or modified to generate adversarial examples; in audio processing, the perturbation objects are more diverse. In speaker recognition or verification tasks, perturbations can be divided by the perturbation object into two categories: time-domain perturbation and frequency-domain perturbation.

3.6.1. Time-Domain Perturbation

In attacks based on time-domain perturbations, the sampled time-domain values of the original audio signals are taken as the perturbation object, and the sampled values are minimized to fool SRSs while making the attack imperceptible. Such perturbations are easy to add; moreover, the features are not required to be inverted into audio signals in the next step, and the signal loss caused by conversion of model features into audios does not need to be considered, which makes the attacks more convenient and efficient [72,75,82,92,93,96]. The perturbations are added to the original audio and then the perturbations are minimized by an optimization algorithm to achieve effective attacks. Compared with frequency-domain perturbations, time-domain perturbations can generate adversarial audios with stronger attack performance.

To show the effect of time-domain perturbations, we here provide the waveforms of original audios and adversarial audios generated by FGSM- and BIM-based time-domain perturbations (Figure 7). As the figure shows, such perturbations mainly work in parts without semantic information, and the magnitude of the perturbation is small. It should be noted that, however, time-domain perturbations are easy to detect if the audios are converted into Mel spectrograms. Figure 8 shows the Mel spectrograms of time-series signals in Figure 6 converted by Fourier transform, and there are tangible perturbations in the frequency domain. In the mainstream SRSs, the neural network learns frequency-domain features, discrete Fourier transform is involved in the extraction of MFCC features, and this process is non-differentiable, so time-domain perturbations are applicable only to SRSs with differentiable features (unless the target SRSs is trained on time-domain features). Additionally, time-domain perturbations can also be used in attack approaches based on model optimization [82,96].

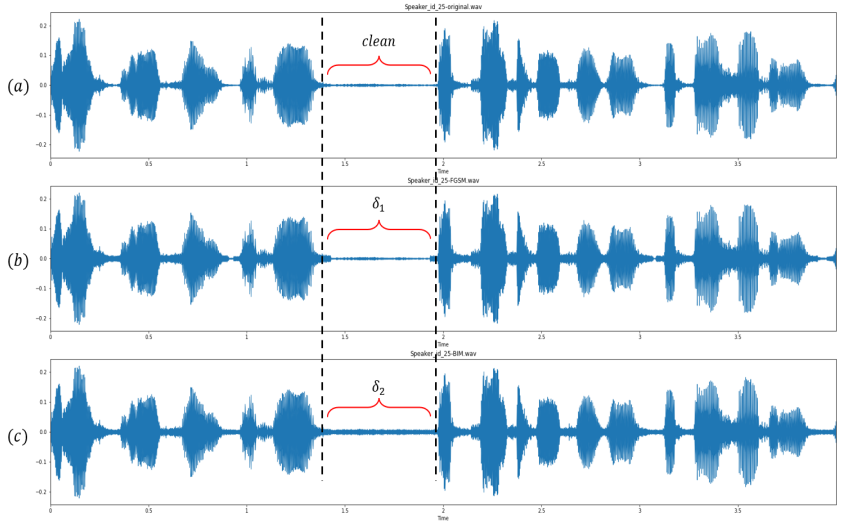

Figure 7. A example of the waveforms of original audios and adversarial audios generated by FGSM- and BIM-based time-domain perturbations. Figure (**a**) shows the waveform of the original audio, while figure (**b**,**c**) show the waveforms of the adversarial audio generated by the FGSM and BIM algorithms respectively.

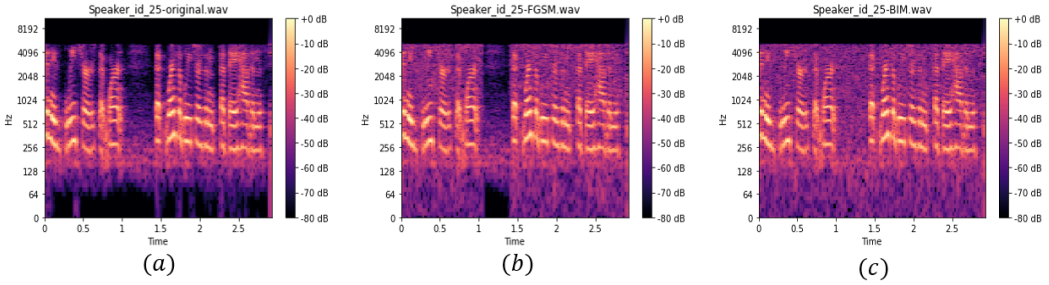

Figure 8. A example of the Mel spectrograms of audios after introduction of FGSM- and BIM-based time-domain perturbations. Figure (**a**) shows the spectrogram of the original audio, while figure (**b**,**c**) show the spectrograms of the adversarial audio generated by the FGSM and BIM algorithms respectively.

3.6.2. Frequency-Domain Perturbation

In attacks with frequency-domain perturbations, the frequency features of utterances (Mel spectrograms, MFCC, and FBank, etc.) converted by Fourier transform are taken as the perturbation object, and this type of perturbation is currently the mainstream form of perturbations. Figure 9 shows the frequency spectra of audio files with perturbations introduced to MFCC features extracted from the original audio files, which are closer to the original spectra than perturbed audios generated by adding noises directly to the original audio files. Nonetheless, refactoring of acoustic features will lead to losses in the speech waveform, which need to be considered in attacks based on frequency-domain perturbations. In some existing works [71,78,83,86], even though the adversarial examples generated with frequency-domain perturbations can make effective attacks, the attack capacity after refactoring of the time-series signals is not evaluated.

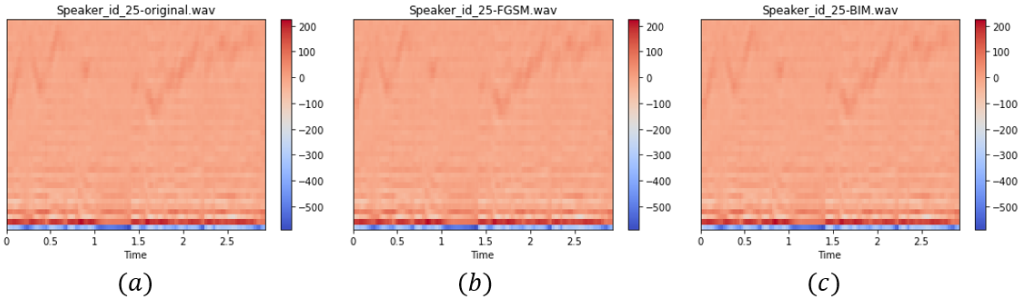

Figure 9. A example of the MFCC of audios with by FGSM- and BIM-based frequency-domain perturbations. Figure (**a**) shows the MFCC of the original audio, while figure (**b**,**c**) show the MFCC of the adversarial audio generated by the FGSM and BIM algorithms respectively.

Through experiments, we found that the acoustic features after frequency domain perturbation are highly aggressive and imperceptible. However, when we transform these features back into audios to attack SRSs, the attack weakens and there are perceptible noises in the audios.

3.7. Perturbation Constraint

Unlike visual perturbations, like perturbations to pixels in images and those to characters or sentences in texts, perturbations in voice adversarial attacks are auditory perturbations. Methods that make perturbations imperceptible include perturbation measurement and psychoacoustic masking.

3.7.1. Perturbation Measurement

Perturbation measurement is to introduce perturbation constraints to the loss function to minimize the perturbation. By the object of constraints, perturbation measurement methods can be divided into the time-domain measurement and frequency-domain measurement. The time-domain measurement is to add constraints to the original audios, and constraints used in this method include SNR, maximum signal-to-noise ratio (MNR), and root mean square (RMS), as well as the commonly used L_p-norm. The frequency-domain measurement is to perform L_2 normalization on the frequency spectrum of the audio. Shamsabadi et al. [96] employed the speech steganography technique and a gated convolutional autoencoder to generate adversarial audio examples; meanwhile, they trained the model by a multi-objective loss function, and controlled the difference between the MFCC features and the original features to produce adversarial examples. This method can achieve a high ASR while minimizing the perturbations, with a PESQ (to be detailed in Section 3.8) of 4.30. Thus, for different attack tasks, constraints can be added to different frequency-domain features to achieve effective attacks.

3.7.2. Psychoacoustic Masking

Psychoacoustics [30] provides mathematical models for statistics of sound perception of humans, and whether a sound can be perceived by human ears depends on the sound frequency, strength and noises. Psychological masking can be employed to improve the aforementioned perturbation measurement methods. In psychoacoustic models, perturbations are mainly introduced to frequency-domain features to conceal the attack. Frequency masking occurs between two sounds with similar frequencies, in which the sound with lower frequencies is covered by a simultaneous higher-frequency masker and becomes imperceptible to human ears. Frequency masking is to create a "masking threshold" in the frequency domain, and any signal below the threshold is imperceptible. Chen et al. [58] proposed that auditory masking can occur before or after the masker, which is termed time-domain masking or non-simultaneous masking. There are two types of non-simultaneous masking: (1) pre-masking, which occurs right before the masker, and (2) post-masking, which occurs after removal of the masker. The physiological mechanism

underlying non-simultaneous masking is that the auditory system needs time to process the perception of sounds, and higher-frequency sounds need more time to process than the lower-frequency sounds.

As most automatic speech recognition and speaker verification systems process frequency-domain signals, frequency masking is a dominating attack method. Inspired by imperceptible adversarial examples in white-box attacks, Wang et al. [97] constrained the perturbation under the masking threshold of the original audio, and generated targeted, inaudible adversarial examples to the original sound waveform, which achieved an ASR of 98.5%. They also applied their method to irrelevant waveforms, such as music and achieved good attack effects, but this method is still at the stage of development. Zhang et al. [85] put forward VMask, which employs psychoacoustic masking to compute the hearing threshold that indicates the masking threshold between different frequencies, and then this hearing threshold is leveraged to restrain the adversarial perturbations under the human perception threshold. VMask was proved to be effective in attacking the grey-box model VGGVox and the black-box system Microsoft Azure, with a perturbation size around 13.13 dB. Compared with FakeBob, VMask generates more imperceptible adversarial perturbations.

3.8. Attack Metrics

Generally, an adversarial example is considered good if it cannot only fool the victim model effective, but avoid being perceived imperceptible. Adversarial examples can be generated by different algorithms in different scenarios, and the quality of these examples are often measured by their attack capacity and imperceptibility.

3.8.1. Effectiveness

Attack success rate (ASR). ASR is the ratio of adversarial audio that is identified as the target speaker, assuming that the test sample is M and the number of samples that can achieve a successful attack is N:

$$ASR = \frac{N}{M} \times 100\% \tag{14}$$

In some works, ASR is also termed the prediction target rate (PTR).

False acceptance rate (FAR), false reject rate (FRR), and equal error rate (EER). Evaluation indicators for SRSs include FRR and FAR, which indicate the classification errors of the target and non-target trials, and EER is a balanced measure when the FAR equals FRR. As real-world attacks are close to non-target trials, the increase in FAR of the victim model post adversarial attacks is more valued. In target trials, adding random noises instead of adversarial perturbations can already lead the system to failure in recognizing the legitimate speaker, so an increased FRR cannot well reflect vulnerability of the SRSs to adversarial attacks. Therefore, we increase the efficiency and the efficiency coefficient to measure the vulnerability of the system to adversarial attacks.

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{FN}} \tag{15}$$

$$\text{FRR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{16}$$

where TP (true positive) is the number of correctly-classified positive samples, FN is the number of misclassified negative samples, FP is the number of misclassified positive samples and TN is the correctly classified negative samples.

MinDCF (minimum detection cost function). In most cases, EER is not required. For instance, an entrance guard system minimizes the FAR to the greatest extent, but has less strict requirements on the FRR. Therefore, different weights are assigned to FRR and FAR.

$$MinDCF = C_{miss} \cdot P_{target} \cdot \text{FRR} + C_{FalseAlarm} \cdot (1 - P_{target}) \cdot \text{FAR} \tag{17}$$

where C_{miss} and $C_{FalseAlarm}$ denote the weights of false rejection and false acceptance, respectively, i.e., the magnitude of the penalty. P_{target} and $1 - P_{target}$ denote the prior probability of occurrence of the real speaker and impostor, respectively.

3.8.2. Imperceptibility

SNR (Signal-to-noise ratio) is the ratio of the power value of the audio and the power value of the noise is used to calculate the size of distortion. The signal-to-noise ratio is calculated as follows,

$$\text{SNR} = 10 \cdot log_{10}\left(\frac{{\sigma_s}^2}{{\sigma_e}^2}\right)^2 \tag{18}$$

where $\frac{{\sigma_s}^2}{{\sigma_e}^2}$ is the mean square of the input signal/error. The larger the SNR ratio value, the less noise in the audio.

MNR (maximum signal-to-noise ratio). Decibels (dB) is a unit of acoustic measurement that calculates the acoustic characteristics of δ using the following equation:

$$\text{dB}(x) = \max_i 20 \cdot log_{10}(x_i) \tag{19}$$

As the scale of sound perceived by human ears is a relative notion, it is not practical to assess the scale of one single perturbed audio. Thus, we measure the distortion of the adversarial audio from the original audio by the decibel difference between the perturbation and the original audio input:

$$\text{dB}_x(\delta) = \text{dB}(\delta) - \text{dB}(x) \tag{20}$$

Obviously, the smaller the $\text{dB}_x(\delta)$, the closer the antagonistic audio is to the original audio, and the more difficult the added perturbation δ is to be perceived by the human.

PESQ [108] is an objective indicator of speech quality, calculated from the stable ratio of spectral density reduction to the reference signal in each time-frequency unit, which can directly and truly reflect the real situation of speech quality. After PESQ analysis, the score ranges from 0 to 5. The higher the score, the better the audio quality, which is a practical evaluation index to combat the problem of whether the audio is inaudible or not.

ABX test. In addition to objective metrics to demonstrate the imperceptibility of counteracting perturbations, the perception of perturbations by the actual human ear can also be measured by a live-action ABX test. The ABX test first provides the user with two segments of speech A and B, each of which may be either the original clean audio or the counteracting audio, and then randomly selects another segment of speech X from the set A, B, and finally asks the tester to decide whether X is A or B. Refs. [73,79,81] carried out this practical test in their work, and ABX testing is available to researchers in need via Amazon's MTurk platform [109].

4. Adversarial Defense

Boosted by the ASVSpoof Challenge series [110], most SRS defense methods are focused on detection of replay attacks, text-to-speech attacks and mimicry attacks, but studies on defense algorithms against adversarial examples are rare. In [106,111], a comprehensive overview of defense methods against adversarial attacks in computer vision is provided, but not all these methods work in speech recognition (SR) tasks. In this section, we introduce effective defense approaches against adversarial attacks to SRSs. By the perspective of defense, the existing and future defense methods can be divided into three types (Figure 10): adversarial training (detailed in Section 4.1), attack detection (detailed in Section 4.2), and input refactoring (detailed in Section 4.3). Table 5 recaps the existing defense methods.

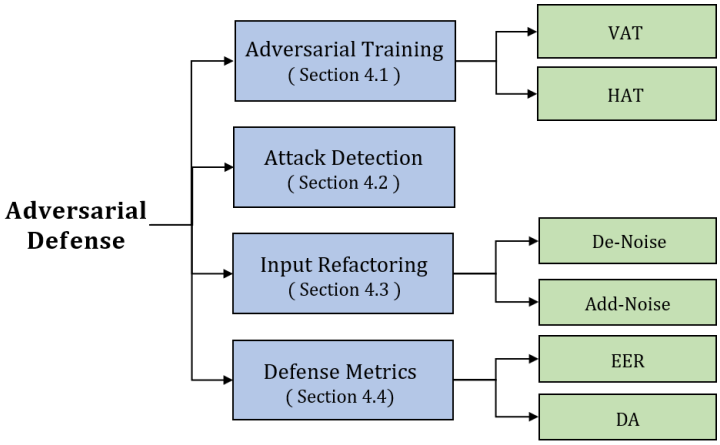

Figure 10. Taxonomy of defense methods against adversarial attacks to SRSs.

Table 5. Existing defense methods against adversarial attacks to SRSs.

Categories	Methods	Metrics (%)	Evaluation						
			Model	Dataset	Performance				
					Baseline	Attack		Defense	
Adversarial Training	Wang [112]	EER	GE2E-ASV	TIMIT	4.87	FGSM	11.89	FGSM	8.31
								LDS	9.26
	Wu [113]	ACC	VGG	ASVspoof 2019	99.99	PGD	37.06	PGD	92.40
			SENet		99.97		48.32		98.60
	Pal [114]	ACC	1D-CNN	LibriSpeech	99.55	FGSM	6.03	HTA10	90.60
						PGD-10	0.00		81.12
						CW-10	0.00		80.12
Attack Detection	Li [115]	DA	VGG-like	VoxCeleb1	-	BIM-xvec		90.65	
		EER						0.46	
	Villalba [116]	ACC	Espresso	VoxCeleb 1&2	-	CW-L2		82.9	
	Peng [117]	FAR	Twin Models	VoxCeleb1	-			4.48	
	Wu [118]	ACC	Representation	VoxCeleb 1&2	-	Vocoder (0.01 FPR)		98.92	
	Joshi [119]	EER	AdvEst	Voxceleb2	-	FGSM/BIM/CW		14.57	
Input Refactoring	Joshi [76]	ACC	ResNet34	LibriSpeech	100	BIM	0	PWG	97.2
						CW	1.3	BPDA	98.8
	Wu [113]	ACC	VGG	ASVspoof 2019	99.99	PGD	37.06	AT+Mean filter	93.76
			SENet		99.97		48.32		99.24
	Zhang [120]	EER	SE-Resnext	VCTK	1.43	FGSM	13.81	3.62	
						PGD	16.66	1.94	
	Wu [121]	ACC	LCNN	ASVspoof 2019	80–90	PGD	(5–10)	(80–90)	
			SENet		80–90		(5–10)	(80–90)	
	Wu [122]	AdvFAR	x-vector	VoxCeleb1	5.97	BIM	87.36	16.54	
		AdvFRR					51.95	16.88	
		j-FAR	r-vector		8.40		48.04	17.84	
		j-FRR					30.41	18.51	
	Wu [123]	FRR	Fast ResNet-34	VoxCeleb 1&2	ASP 2.24	BIM-10	89.38	3.6	
		FAR			2.56		91.94	16.67	

Table 5. *Cont.*

Categories	Methods	Metrics (%)	Evaluation						
			Model	Dataset	Performance				
					Baseline	Attack		Defense	
Input Refactoring	Wu [124]	EER	r-vector	VoxCeleb1	8.87	BIM	66.02	22.94	
	Oliver [125]	ACC	1D-CNN	LibriSpeech	88	PGD	7	MAD	74
						CW	9		69
	Chang [126]	ASR	i-vector	LibriSpeech	-	BIM	100	1	

4.1. Adversarial Training

Adversarial training is a method for generating adversarial perturbations based on model gradients and constraining the perturbations with normalized spheres in the embedding space, thus improving the robustness of the model, which is first proposed by Goodfellow [29]. As shown in Figure 11a, adversarial training further delineates the decision boundary of the model through a robustness-enhanced training process, we can consider it as a special adversarial data enhancement strategy. The working principle underlying adversarial training is as follows: the adversarial examples are injected into the training set as new training samples in the model training stage, so that the trained model achieves not only a higher recognition accuracy, but stronger robustness against adversarial examples. Adversarial training has proved to be an effective defense method in image processing, and shows robustness in text processing tasks [127]. Thus, for specific attacks, adversarial training may be the most effective defense method.

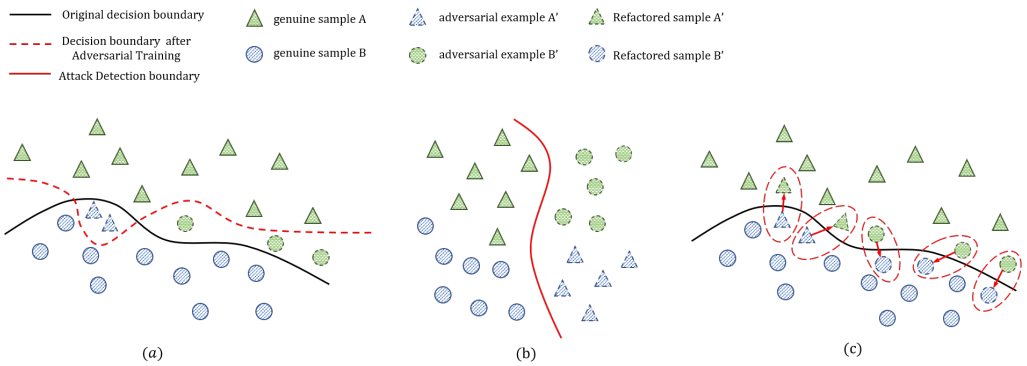

Figure 11. The two-dimensional visualization of different defense strategies. Figure (**a**) represents the adversarial training with the aim of finding robust decision bounds. Figure (**b**) represents attack detection, distinguishing clean samples from adversarial samples. Figure (**c**) represents input reconstruction, which uses denoising, noise addition and sample purification to repair the input data.

In terms of SRS defense, Wang et al. [112] first put forward a virtual adversarial training method with adversarial examples generated by the fast gradient sign method (FGSM) and the local distributional smoothing (LDS) method [128]. FGSM-Adv is a supervised adversarial training method, and it modestly reduces the EER of original attacks from 11.89% to 8.31% when applied to the GE2E model [129], which is not a satisfactory outcome. Likewise, the unsupervised virtual adversarial training scheme LDS-virtual adversarial training (LDS-VAT) merely reduces the ERR to 9.26%. Though these works proved applicability of adversarial training to SRSs, the defense effect falls short of the ideal. To increase the defense capacity of adversarial training schemes, Wu et al. [113] trained a model using adversarial examples generated by the projected gradient descent (PGD) method that has stronger attack performance than FGSM, which increased the testing accuracy of VGG from 37.06% to 98.60% (it increased the accuracy from 48.32% to

92.40% on the SENet model). They also found that equipping adversarial training with spatial smoothing based on introduction of median filters or mean filters can improve the adversarial robustness.

Though single adversarial training has been proved effective in defense against similar adversarial attacks, its defense performance drops considerably when the attack strategy varies. To enhance the defense performance of adversarial training, Pal et al. [114] put forward a new defense mechanism based on a hybrid adversarial training (HAT) setup. Specifically, they employed multi-task objectives using cross-entropy (CE), feature scattering (FS) [130], and marginal loss (ML) [30] to perform HAT. As shown in Table 5, adversarial training with a richer collection of adversarial examples has better performance than individual adversarial training, can defend some black-box attacks, and has better adversarial robustness.

To the best of our knowledge, adversarial training is not perfect, and due to iterative updates of attack methods, adversarial training does not fully defend against all attacks and always lags behind existing attacks as well. On the other hand, it was shown through research [30] that adversarial training degrades the recognition accuracy of the original model to some extent, i.e., the recognition accuracy of the adversarially trained model for clean samples tends to be slightly lower than that of the model without adversarial training. Therefore, the decrease in recognition accuracy caused by adversarial training is a meaningful research topic, and integrated adversarial training is likely to become a new research hotspot in the future.

4.2. Attack Detection

Attack detection is to add a pretrained detection module to the original model to discriminate adversarial examples from genuine samples, just as Figure 11b shows. Instead of adding adversarial examples to the training set to train the SR model, the attack detection method needs to design a strongly discriminative detector. Li et al. [115] designed a VGG-like binary classification detector, which captures the difference between the adversarial examples and genuine ones by the convolutional layer and aggregates the speech sequences by the pooling layer for decision. The binary classification detector showed good adversarial robustness on cross-model SRSs, but reached a reduced recognition accuracy in face of different attacks (the model trained on adversarial examples generated by the BIM algorithm could detect 99.83% BIM attacks, but the detection rate dropped to 48.61% on JSMA attacks). To identify different attacks, Villalba et al. [116] employed representation learning to classify adversarial attacks. They applied probabilistic linear discriminant analysis (PLDA) in the x-vector system for the detection and classification of attacks, which reached a recognition accuracy as high as 71.8% on the classification of attacks within the training set, and an error rate of merely 19.6% in detecting unknown attacks. Recently, they proposed a method to estimate adversarial perturbation, which was named AdvEst [119]. Instead of adversarial examples, they trained the representation learning network by using adversarial perturbations, and employing the time-domain denoiser to estimate the adversarial perturbations. Compared with the method proposed in [115], this method can detect some unknown attacks, though the detection accuracy remains low. Wu et al. [118] employed the neural vocoder in Parallel WaveGAN [131] to detect adversarial examples. The vocoder modifies the audios in the time domain to re-synthesize new audios, and then uses the difference between the ASV scores for the original and re-synthesized audio to discriminate genuine and adversarial examples. It achieved a detection accuracy as high as 99.76%, which outperforms the Griffin–Lim algorithm [132].

Peng et al. [117] proposed a twin-model attack detection scheme and put forward a twin-model design comprising two SV models as a defense strategy. For the two models, the TDNN x-vector is set as the less-robust premier model, and the ResNet-34 r-vector as the more robust mirror model; then, the one-class classification (OCC) classifier detects inconsistency in the verification scores output from the two models for a single sample to capture potential adversarial attacks. In their method, a simple classifier, the minimum covariance

determinant (MCD) [133], is used, and it is trained only by genuine samples to identify the decision threshold for clean samples without the need for generating adversarial examples, which improves the adversarial robustness of the defense.

The methods proposed in [115,116] are proved to be robust on varied ASV models under the same type of attack, but its detection performance drops drastically for the same ASV model under unknown attacks, which means the detector can only detect attacks already present in the training set and has very limited performance. Models proposed in [117,118] make use of the weak adversarial robustness of the model and detect potential adversarial examples by score differences; such defense methods bypass the need to know the attack methods beforehand, and hence are more robust than the aforementioned methods, but it is yet to be explored whether they can defend adaptive attacks that are have stronger attack capacity. Recent work in Chen's research [134] has shown that adversarial perturbations usually occur in the high-frequency part of the audio, and they involved a MEH-FEST detector that is able to calculate the minimum energy at high frequencies from the short-time Fourier transform of the audio and use it as a detection metric. MEH-FEST can detect FakeBob attack samples with high accuracy, and false positives and false negatives can approach 0.

4.3. Input Refactoring

Adversarial training requires prior knowledge about the attack method, and shows reduced performance in the face of unknown attacks. Attack detection is a detection module that performs binary classification of samples. From the perspective of data cleansing, as shown in Figure 11c, a pretrained preprocessing module can be employed to denoise or refactor the data, thereby cleansing the data, to reduce the probability of attack and achieve the goal of defense. Such methods that preprocess the input data on the input layer are collectively termed input refactoring, which involves denoising and noise addition.

4.3.1. Denoising

Denoising, which removes or reduces perturbations or noises in audios, is the most prevalent defense strategy in audio processing at present. There are various denoising methods: we can directly process the audio files and denoise the audios in the frequency domain, or train an effective neural network to perform frequency-domain denoising. Common denoising methods include spatial smoothing, autoencoder, and separation network.

Spatial smoothing. Also termed filtering, spatial smoothing is a classical time-domain denoising method. Smoothing filters, which smooth the central pixel with the pixels around, have been widely used in denoising images. There are different smoothing filters which have different weighting mechanism to neighboring pixels, such as SEC4SR [58], the median filter [135] used in [115], the mean filter [87], and Gaussian filters, etc. Take the mean filter, for example: a sliding window moves along the audio waveform, and the center value in the window is replaced with the mean of all pixel values in the window. Olivier et al. [125] explored a high-frequency smoothing method based on additive noise masking, and applied the Gaussian filters—preemphasis filter and ButterWorth highpass filter—to Gaussian noise, which are collectively called moving average difference (MAD) smoothing or BW smoothing. They found that compared with traditional random smoothing methods, the MAD or BW smoothing methods increased the defense accuracy from 13% to 64%, which improved the adversarial robustness of the model.

Autoencoder. The key of an autoencoder is to mask a proportion of original audio inputs and then train a decoder for audio refactoring to achieve self-supervised learning. It underlies all existing self-supervised learning-based defense strategies, though the masking strategies and refactoring decoders vary. For instance, Wu et al. [121] proposed Mockingjay [136], a decoder targeted for surface noises in the spectrogram of inputs, masked selected frames to zero, and replaced all selected frames with random frames. In [124], Wu et al. proposed transformer encoder representations from alteration (TERA), a more advanced self-supervised model than Mockingjay, and introduced the cascading mechanism

to use cascaded TERA models as a deep filter, which substantially reduced the success rate of adversarial attacks (10 cascaded TERA models increased the EER from 66.02% to 22.94%). In their subsequent work [122], they introduced the voting mechanism, and uploaded a refactored audio sample by different numbers of self-supervised models into the SRSs for scoring to identify whether the sample is malicious by the average of scores. Compared with defenses without the scoring mechanism, this method improved the EER by 6%. Joshi et al. [75] refactored the speech waveform by Parallel WaveGan [131], which achieved the best defense performance combined with random smoothing, and increased the defense accuracy of the model from 52% to 93%.

Separation network. The adversarial disturbances are separated from the adversarial examples to recover the natural clean audio. Specifically, a separate filtering module is designed and applied before the audio that is about to be input to the SRSs, through which the audio, semantic and identity information is retained, while the adversarial disturbances are removed as noise. Zhang et al. [120] proposed an adversarial separation network (AS-Net) that combines a PR-Net consisting of a compression structure and a reconstruction module and an audio quality loss to reconstruct the input audio and supervise the recovered speech generated by AS-Net with a (reality quality) RQ-Net network similar to the real clean audio. It is shown experimentally that AS-Net has stronger defensive performance on SE-ResNet-based speaker recognition models compared to adversarial training and APE-GAN.

4.3.2. Noise Addition

Different from denoising, the noise addition approach tries to interfere with the adversarial audio so that the adversarial attack reduces or loses the ability to mislead the SV system. Specifically, the reason for the success of this approach is that SRS is a slightly skewed mapping function from data space to label space, which prevents the SR model from covering some regions of the input space well. The uncovered input space can be considered a blind spot, and the adversarial attacker deliberately tries various methods to find the blind spot and make the SR model misjudge, and we can consider the adversarial sample as the blind spot of the SR model. To solve the problem of existing blind spots, data expansion is a more intuitive way to enrich the distribution of the model and thus cover the blind spots. Adding noise of tiny Gaussian noise to the time domain of the input audio for defense in [126] was able to reduce the 100% attack success rate of the FakeBob attack to 5.2%. In addition, ref. [123] provided another approach to deal with blind spots: filtering potential adversarial examples by voting. They randomly sampled samples within the Gaussian sphere of the test sample as the "neighbors" of the test sample, and then asked the "neighbors" of the test sample to vote for the correct answer. Due to the small size of these blind areas, when random sampling is performed, the samples tend to jump out of the blind spot if the sampling variance σ is large enough. The sampled "neighbors" are more likely to be in the robust region of the model rather than in the blind spot region. Then, after voting, the increased probability of being in the robust region leads to an increased probability of making a "normal" decision.

Denoising and noise addition are two contradictory strategies, but in the current study, both strategies enhance the model's ability to combat robustness in the face of unknown attacks, even if the defense performance is inferior to that of mixed adversarial training. The existing noise addition strategies are performed in the time domain, and in combination with the high-frequency filtering smoothing in Section 4.3.1, noise addition in the high-frequency part of the frequency domain seems to provide some defensive effect as well. In addition, with the recent boom in the migration of self-supervised pre-trained models from text to image domains, it seems that such self-supervised mechanisms could also shine in the study of voice recognition and its adversarial defense.

4.4. Defense Metrics

The performance of defense strategies needs to be measured by harmonized metrics, but there is no universal set of evaluation metrics for now. In most works, the EER of the

victim model is used to judge the performance of the defense strategies. To evaluate the effects of adversarial training and input refactoring, Chang et al. [122] put forward AdvFAR and AdvFRR, two indicators calculated based on EER, to assess the model's performance in defending against adversarial attacks.

$$\text{AdvFAR} = \frac{|\{S_i \geq \tau : i \in T_{ant}\}|}{|T_{ant}|} \tag{21}$$

$$\text{AdvFAR} = \frac{|\{S_i < \tau : i \in T_{at}\}|}{|T_{at}|} \tag{22}$$

where τ is the threshold of the original ASV model, T_{ant} is the set of samples of nontargeted attacks, and T_{at} represents the set of samples of targeted attacks. The joint FAR (j-FAR) and joint FRR (j-FRR) are calculated after blending the adversarial examples with clean samples.

$$j\text{-FAR} = \frac{|\{S_i \geq \tau : i \in T_{jnt}\}|}{|T_{jnt}|} \tag{23}$$

$$j\text{-FAR} = \frac{|\{S_i < \tau : i \in T_{jt}\}|}{|T_{jt}|} \tag{24}$$

where $T_{jnt} = T_{ant} \cup T_{gt}$, and $T_{jt} = T_{at} \cup T_{gt}$. T_{gt} and T_{gnt} denote the sets of trials consisting of genuine target and genuine non-target trials, respectively. To assess the defense performance of attack detection methods, Chen et al. [86] employed the indicator detected accuracy (DA):

$$\text{DA} = \frac{T_{adv}}{T_{real} + T_{adv}} \tag{25}$$

The fidelity of defense methods and the attack cost after introduction of the defense methods can also be used as defense metrics. Fidelity measures the impacts of the introduction of the defense strategies or models on the recognition accuracy of the original model. As existing defense methods are not indestructible, more advanced attacks can be designed to defy these defenses; however, the defenses will, to varied degrees, increase the computing time and complexity of the attacks. Thus, the attack cost can also be used to measure the defense performance.

5. Discussion

Previous sections have elaborated on previous works on SRS attacks and defenses. In this section, we provide more discussions and point out the challenges in this field.

5.1. General Observations of Adversarial Audio

Here, we discuss the SRS attacks and defenses from the perspectives of attack dimension, perturbation amplitude, transferability, defense capacity, physical-domain black-box scenarios.

With regard to the attack dimension, most SRS adversarial attacks are model-level attacks targeting the model parameters or decisions, which have weak transferability. Data-level attacks, however, seem unique to audio processing systems. Abdullah et al. [72,86]. have probed deep into this field and realized attacks by the modification of single phonemes, time-domain compression, and the inversion of sound features. However, the adversarial audios they generated are basically noises, which are easy to be perceived by human ears, and hence the effectiveness of such attacks is yet to be proved. Undoubtedly, there is still much room for exploration in related research.

Regarding the perturbation amplitude, most existing attacks need to add adversarial perturbations to the whole audio inputs. To put it another way, the perturbation needs to last as long as the audio input, which is infeasible in processing of streaming inputs. Additionally, the attacks are based on the assumption that the audio input and the perturbation are strictly synchronized. To ensure synchronization, the adversarial perturbation needs to be blended with the audio input beforehand and then played by the speaker during the attack.

In terms of transferability, like adversarial examples in other fields, those against SRSs are generated based on a specific dataset for a specific model. These examples can successfully attack models with a similar structure as the one that outputs the examples, but witness reduced attack capacity on models with a different structure. Thus, it is worth further research to explore how to generate adversarial examples that are transferable across models and datasets.

In terms of defenses, existing defense methods are not enough to cope with all adversarial attacks. In [84], adaptive attack techniques, such as backward pass differentiable approximation (BPDA) and expectation over transform (EOT) in image processing, are employed to attack speaker recognition models, and how to cope with such high-strength adaptive attacks is worth more research effort. Metric learning techniques, such as triplet loss adversarial (TLA) training, have been proved to be able to learn close and robust embeddings to defend adversarial attacks to image processing models, and they are found to be applicable in protecting speaker recognition systems. One natural extension can verify the adversarial robustness of metric learning defenses on SRSs.

Physical-domain black-box attacks and defenses are the focus of future research in this field of adversarial attacks and defenses of SRSs. Most attacks against SRSs are digital attacks, whereas works on over-the-air attacks are rare. One challenge of over-the-air attacks is that as the audio signals, when transmitted in the physical world, may be damaged by the air medium, the audio quality will be degraded and the attack strength will reduce. There are only a few works that have attempted to address this problem: in [77,85], the researchers simulated the distortion or reverberation of audios in a closed space by introducing the idea of room impulse response (RIR) [92], and introduced RIR to the loss function to generate robust adversarial audios. Though this method works well in room simulators, there are more other factors to consider in physical-domain attacks, such as the impacts from multiple audio sources or natural noises. Moreover, as the internet of vehicles, smart vehicle-mounted systems, smart homes and other intelligent devices equipped with speech or speaker recognition functions gain popularity, it is of more practical significance to explore potential attacks to these systems and feasible defenses.

5.2. Challenges

Adversarial attack and defense remain a challenging research topic. As we predict, the major challenges in the research on SRS adversarial attacks and defenses lie in the following aspects:

Evaluation of attack or defense performance: Most of the recent works evaluate the attack performance by the attack success rate or accuracy; very few works consider the scale and efficiency, and these works factor in only the time cost for attacks. Whether there are correlations between the scale of the dataset, the time cost for attacks and the success rate of adversarial attacks is yet to be explored. If there are correlations, how to balance the three aspects is likely to be another focus of future research. The evaluation of defense performance faces the same situation.

Physical-domain attacks: Most existing attacks against SRSs are realized digitally and have reached a high success rate. However, when the generated adversarial audios are applied to SRSs in the physical world, the success rate of the attack drops substantially or even approaches 0. Though there are works that probe into physical-domain attacks and have achieved effective attacks, the generated adversarial audios are of poor quality, and the perturbations are easy to detect by human ears, which is not rational in real-world scenarios.

Real-time attacks: Existing attack methods are mainly static attacks, which are not applicable to systems with streaming audios. Advpulse [88] is a pioneering attempt that probes into real-time attacks, but falls far short of perfection. How to add imperceptible perturbations to audios processed in real time and bypass the defense of the target systems is another topic worth future research.

Generalization of defense strategies: Most defense methods are based on prior knowledge of attacks, which can be nullified if the attacker improves the original attack to produce stronger attacks. Thus, it is urgent to find universal defenses that are resistant to attacks and do not undermine the robustness of the original model.

Lack of benchmarks and toolkits: As with the case of works on text and image processing, there are no benchmarks for adversarial attacks and defenses on speaker recognition or verification. As a result, there are no universal standards for evaluation, making it hard to evaluate the effectiveness of related works. Moreover, unlike the situation for works on text and image processing that have such adversarial attack and defense toolkits as AdvBox [137], TorchAttack [138], FoolBox [139], Text attack [140] and OpenAttack [141], there are no available toolkits for works in the audio processing field.

5.3. Future Directions

As the current development status and challenges mentioned above, we would suggest the following directions for some future developments.

Firstly, a fair evaluation framework is necessary. It is difficult to evaluate the performance of adversarial attacks and defenses between different efforts. We propose the following benchmarks for future evaluation from both the attacker's perspective and the defender's perspective. From the attacker's perspective, a comprehensive evaluation is conducted in terms of the time of adversarial generation, attack success rate, audio quality (including signal-to-noise ratio, PESQ, STOI, etc.), transferability, perturbation generalizability, over-the-air, and whether it is useful to attack commercial SRSs. Different evaluation schemes are set up according to different attack scenarios to measure the merits of two or more attack methods. From the defender's perspective, in addition to the FAR, FRR and EER of the SRS, the memory and computational overhead of the defense module need to be considered. The degree of impact on the recognition accuracy of the original system after the introduction of the defense module is considered comprehensively.

Second, the attack ability of the adversarial examples can be improved by some advanced ideas in other deep learning fields, such as model inversion, gradient inversion, contrast-learning, and meta-learning. For example, there are various methods in which meta-learning can be used, such as learning initial perturbations, learning attack algorithms, learning how to optimize the internal gradient of a model, etc. In recent years, a series of adversarial attack methods based on the ideas of meta-learning were proposed in the image domain, and these methods can contribute more to the field of speaker recognition as well.

Thirdly, for physical-domain and real-time attacks, which are more practical, we can further investigate at the data level. Since most works so far focus on the model level, as far as we know, there are only three papers that aim at the data level, which is not enough. The acoustic information is a subset of the signal, and we need to focus on some signal processing aspects to improve the effectiveness of the attack.

Fourth, in terms of defense methods, most of the current defense is considered from the model layer and lacks some data layer defense methods for anomaly detection of audio acoustic features. In addition, some advanced defense methods in the field of applied imagery, such as knowledge distillation, are model-based and Bayesian model-based.

Fifth, the open-source platform SEC4SR was used to enable thorough comparison between some existing attacks and defenses (6 attacks and 24 defenses), and multiple metrics such as SNR, PESQ, and STOI have been involved to evaluate the performance. It accounts for the most comprehensive work on ASV adversarial attacks and defenses, but the codes of the platform do not allow plug-and-play like a toolkit. A readily available toolkit for adversarial attacks and defenses can save much time for repeated coding and

advance research in this field. Thus, it is necessary to develop toolkits for adversarial attacks and defenses.

6. Conclusions

In this paper, we investigated deep neural network based adversarial attack and defense works on SRSs, reviewed almost all existing works on SRS adversarial attacks and some recent works that provide possible defense solutions. First, we introduced deep speaker recognition models and some commonly used speech or speaker datasets. Then, we introduced and classified attacks by the attack method, perturbation strategies and other aspects, and analyzed the robustness of different defense strategies. Finally, we discussed the progress of works on adversarial examples against SRSs and potential problems. We also pointed out the challenges in this field in hopes of facilitating future research. In future works, we will continue focusing adversarial attack and defense against SRSs, and explore real-time adversarial attack and defense strategies for physical-domain black-box models so as to provide robust solutions to audio signal processing systems.

Author Contributions: Conceptualization, H.T. and Z.G.; methodology, H.T.; software, H.Z.; validation, H.T., H.Z. and J.Z.; formal analysis, L.W.; investigation, H.T., H.Z. and J.Z.; resources, L.W., Z.G.; data curation, H.T. and J.Z.; writing—original draft preparation, H.T. and J.Z.; writing—review and editing, H.Z., L.W., M.S., L.W. and Z.G.; visualization, H.T.; supervision, M.S., L.W. and Z.G.; project administration, L.W.; funding acquisition, Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the Major Key Project of PCL (No. PCL2022A03), the Natural Science Foundation of China (No. 61902082), and the Guangzhou Science and technology planning project (No. 202102010507) and the Guangzhou University Graduate Student Innovation Ability Cultivation Funding Program (grant no. 2021GDJC-M34).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gadekallu, T.R.; Manoj, M.K.; Kumar, N.; Hakak, S.; Bhattacharya, S. Blockchain-Based Attack Detection on Machine Learning Algorithms for IoT-Based e-Health Applications. *IEEE Internet Things Mag.* **2021**, *4*, 30–33. [CrossRef]
2. Gu, Z.; Li, H.; Khan, S.; Deng, L.; Du, X.; Guizani, M.; Tian, Z. IEPSBP: A Cost-efficient Image Encryption Algorithm based on Parallel Chaotic System for Green IoT. *IEEE Trans. Green Commun. Netw.* **2021**, *6*, 89–106. [CrossRef]
3. Gu, Z.; Wang, L.; Chen, X.; Tang, Y.; Wang, X.; Du, X.; Guizani, M.; Tian, Z. Epidemic risk assessment by a novel communication station based method. *IEEE Trans. Netw. Sci. Eng.* **2021**, *9*, 332–344. [CrossRef] [PubMed]
4. Javed, A.R.; Ur Rehman, S.; Khan, M.U.; Alazab, M.; Reddy, T. CANintelliIDS: Detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 1456–1466. [CrossRef]
5. Shafiq, M.; Tian, Z.; Bashir, A.K.; Jolfaei, A.; Yu, X. Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustain. Cities Soc.* **2020**, *60*, 102177. [CrossRef]
6. Shafiq, M.; Tian, Z.; Bashir, A.K.; Du, X.; Guizani, M. IoT malicious traffic identification using wrapper-based feature selection mechanisms. *Comput. Secur.* **2020**, *94*, 101863. [CrossRef]
7. Farokhi, S.; Flusser, J.; Sheikh, U.U. Near, infrared. Face recognition: A literature survey. *Comput. Sci. Rev.* **2003**, *21*, 1–7. [CrossRef]
8. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. x-vectors: Robust dnn embeddings for speaker recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
9. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020), Shanghai, China, 25–29 October 2020; pp. 3830–3834.
10. Pelecanos, J.; Wang, Q.; Moreno, I.L. Dr-Vectors: Decision residual networks and an improved loss for speaker recognition. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August–3 September 2021; pp. 4603–4607.

11. Hautamäki, R.G.; Kinnunen, T.; Hautamäki, V.; Leino, T.; Laukkanen, A.M. I-vectors meet imitators: On vulnerability of speaker verification systems against voice mimicry. In Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France, 25–29 August 2013; pp. 930–934.
12. Godoy, E.; Rosec, O.; Chonavel, T. Voice Conversion Using Dynamic Frequency Warping With Amplitude Scaling, for Parallel or Nonparallel Corpora. *Trans. Speech Audio Process.* **2011**, *4*, 1313–1323. [CrossRef]
13. Wu, Z.; Virtanen, T.; Kinnunen, T.; Chng, E.; Li, H. Exemplar-based unit selection for voice conversion utilizing temporal information. In Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013), Lyon, France, 25–29 August 2013; pp. 3057–3061.
14. Ze, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 7962–7966.
15. Zen, H.; Tokuda, K.; Black, A.W. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *1*, 1039–1064. [CrossRef]
16. Jia, Y.; Zhang, Y.; Weiss, R.; Wang, Q.; Shen, J.; Ren, F.; Nguyen, P.; Pang, R.; Lopez, M.I.; Wu, Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS), Montréal, QC, Canada, 3–8 December 2018; pp. 4485–4495.
17. Wang, Z.F.; Wei, G.; He, Q.H. Channel pattern noise based playback attack detection algorithm for speaker recognition. In Proceedings of the 2011 International Conference on Machine Learning and Cybernetics (ICMLC), Guilin, China, 10–13 July 2011; pp. 1708–1713.
18. Syverson, P. A taxonomy of replay attacks [cryptographic protocols]. In Proceedings of the Computer Security Foundations Workshop. Franconia, NH, USA, 14–16 June 1994; pp. 187–191.
19. Villalba, J.; Lleida, E. Preventing replay attacks on speaker verification systems. In Proceedings of the International Carnahan Conference on Security Technology (ICCST), Barcelona, Spain, 18–21 October 2011; pp. 1–8.
20. Yoon, S.H.; Koh, M.S.; Park, J.H.; Yu, H.J. A new replay attack against automatic speaker verification systems. *IEEE Access* **2020**, *8*, 36080–36088. [CrossRef]
21. Alegre, F.; Janicki, A.; Evans, N. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. In Proceedings of the 13th International Conference of the Biometrics Special Interest Group, Darmstadt, Germany, 10–12 September 2014; pp. 1–6.
22. Wu, Z.; Yamagishi, J.; Kinnunen, T.; Hanilçi, C.; Sahidullah, M.; Sizov, A.; Evans, N.; Todisco, M.; Delgado, H. ASVspoof: The automatic speaker verification spoofing and countermeasures challenge. *IEEE J. Sel. Top. Signal Process.* **2017**, *4*, 588–604. [CrossRef]
23. Nautsch, A.; Wang, X.; Evans, N.; Kinnunen, T.H.; Vestman, V.; Todisco, M.; Delgado, H.; Sahidullah, M.; Yamagishi, J.; Lee, K.A. ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Trans. Biom. Behav. Identity Sci.* **2021**, *3*, 252–265. [CrossRef]
24. Saha, A.; Subramanya, A.; Pirsiavash, H. Hidden trigger backdoor attacks. In Proceedings of the AAAI conference on artificial intelligence, New York, NY, USA, 7–12 February 2020; pp. 11957–11965.
25. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv* **2017**, arXiv:1712.05526.
26. Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; Zhao, B.Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 19–23 May 2019; pp. 707–723.
27. Zhai, T.; Li, Y.; Zhang, Z.; Wu, B.; Jiang, Y.; Xia, S.T. Backdoor attack against speaker verification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2560–2564.
28. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
29. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
30. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
31. Dong, Y.; Fu, Q.A.; Yang, X.; Pang, T.; Su, H.; Xiao, Z.; Zhu, J. Benchmarking adversarial robustness on image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 321–331.
32. Wang, J. Adversarial examples in physical world. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), Montreal, ON, Canada, 19–27 August 2021; pp. 4925–4926.
33. Zhang, W.E.; Sheng, Q.Z.; Alhazmi, A.; Li, C. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*, 1–41. [CrossRef]
34. Zhu, B.; Gu, Z.; Qian, Y.; Lau, F.; Tian, Z. Leveraging Transferability and Improved Beam Search in Textual Adversarial Attacks. *Neurocomputing* **2022**, *500*, 135–142. [CrossRef]

35. Carlini, N.; Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 24 May 2018; pp. 1–7.
36. Xu, H.; Ma, Y.; Liu, H.C.; Deb, D.; Liu, H.; Tang, J.L.; Jain, A.K. Adversarial attacks and defenses in images, graphs and text: A review. *Int. J. Autom. Comput.* **2020**, *17*, 151–178. [CrossRef]
37. Hu, S.; Shang, X.; Qin, Z.; Li, M.; Wang, Q.; Wang, C. Adversarial examples for automatic speech recognition: Attacks and countermeasures. *IEEE Commun. Mag.* **2019**, *57*, 120–126. [CrossRef]
38. Cisse, M.; Adi, Y.; Neverova, N.; Keshet, J. Houdini: Fooling deep structured prediction models. *arXiv* **2017**, arXiv:1707.05373.
39. Mode, G.R.; Hoque, K.A. Crafting adversarial examples for deep learning based prognostics. In Proceedings of the 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 14–17 December 2020; pp. 467–472.
40. Wang, D.; Wang, R.; Dong, L.; Yan, D.; Zhang, X.; Gong, Y. Adversarial examples attack and countermeasure for speech recognition system: A survey. In Proceedings of the International Conference on Security and Privacy in Digital Economy, Singapore, 30 October 2020; pp. 443–468.
41. Carlini, N.; Mishra, P.; Vaidya, T.; Zhang, Y.; Sherr, M.; Shields, C.; Wagner, D.; Zhou, W. Hidden voice commands. In Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, USA, 10–12 August 2016; pp. 513–530.
42. Das, R.K.; Tian, X.; Kinnunen, T.; Li, H. The attacker's perspective on automatic speaker verification: An overview. Interspeech 2020. In Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020; pp. 4213–4217.
43. Abdullah, H.; Warren, K.; Bindschaedler, V.; Papernot, N.; Traynor, P. Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems. In Proceedings of the IEEE Symposium on Security and Privacy (SP 2021), San Francisco, CA, USA, 24–27 May 2021; pp. 730–747.
44. Chen, X.; Li, S.; Huang, H. Adversarial Attack and Defense on Deep Neural Network-Based Voice Processing Systems: An Overview. *Appl. Sci.* **2021**, *11*, 8450. [CrossRef]
45. Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust.* **1978**, *26*, 43–49. [CrossRef]
46. Reynolds, D.A.; Rose, R.C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 72–83. [CrossRef]
47. Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* **2000**, *10*, 19–41. [CrossRef]
48. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 788–798. [CrossRef]
49. Variani, E.; Lei, X.; McDermott, E.; Moreno, I.L.; Gonzalez-Dominguez, J. Deep neural networks for small footprint text-dependent speaker verification. In Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP 2014), Florence, Italy, 4–9 May 2014; pp. 4052–4056.
50. Bai, Z.; Zhang, X.L. Speaker recognition based on deep learning: An overview. *Neural Netw.* **2021**, *140*, 65–99. [CrossRef] [PubMed]
51. Muda, L.; Begam, M.; Elamvazuthi, I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv* **2010**, arXiv:1003.4083.
52. Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.* **1990**, *87*, 1738–1752. [CrossRef]
53. Nandwana, M.K.; Ferrer, L.; McLaren, M.; Castan, D.; Lawson, A. Analysis of Critical Metadata Factors for the Calibration of Speaker Recognition Systems. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019), Graz, Austria, 15–19 September 2019; pp. 4325–4329.
54. Dehak, N.; Dehak, R.; Glass, J.R.; Reynolds, D.A.; Kenny, P. Cosine similarity scoring without score normalization techniques. In Proceedings of the Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, 28 June–1 July 2010; p. 15.
55. Wang, D. A simulation study on optimal scores for speaker recognition. *EURASIP J. Audio Speech Music. Process.* **2020**, *1*, 1–23. [CrossRef]
56. Hansen, J.H.; Hasan, T. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Process. Mag.* **2015**, *32*, 74–99. [CrossRef]
57. Jati, A.; Georgiou, P. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1577–1589. [CrossRef]
58. Chen, G.; Zhao, Z.; Song, F.; Chen, S.; Fan, L.; Liu, Y. SEC4SR: A security analysis platform for speaker recognition. *arXiv* **2021**, arXiv:2109.01766.
59. Dehak, N.; Dehak, R.; Kenny, P.; Brümmer, N.; Ouellet, P.; Dumouchel, P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In Proceedings of the 10th Annual Conference of the International Speech Communication (INTERSPEECH 2009), Association, Brighton, UK, 6–10 September 2009; pp. 1559–1562.
60. Zeinali, H.; Wang, S.; Silnova, A.; Matějka, P.; Plchot, O. But system description to voxceleb speaker recognition challenge 2019. *arXiv* **2019**, arXiv:1910.12592.
61. Ravanelli, M.; Bengio, Y. Speaker recognition from raw waveform with sincnet. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT 2018), Athens, Greece, 18–21 December 2018; pp. 1021–1028.

62. Chung, J.S.; Nagrani, A.; Zisserman, A. Voxceleb2: Deep speaker recognition. In Proceedings of the 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 1086–1090.
63. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. *Natl. Inst. Stand. Technol. (NIST)* **1988**, *107*, 16.
64. Jankowski, C.; Kalyanswamy, A.; Basson, S.; Spitz, J. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1990), Albuquerque, NM, USA, 3–6 April 1990; pp. 109–112.
65. Bu, H.; Du, J.; Na, X.; Wu, B.; Zheng, H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In Proceedings of the 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA 2017), Seoul, Korea, 1–3 November 2017; pp. 1–5.
66. Du, J.; Na, X.; Liu, X.; Bu, H. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv* **2018**, arXiv:1808.10583.
67. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. LibriSpeech: An asr corpus based on public domain audio books. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015), South Brisbane, QSD, Australia, 19–24 April 2015; pp. 5206–5210.
68. Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.* **2020**, *60*, 101027. [CrossRef]
69. Campbell, J.P. Testing with the YOHO CD-ROM voice verification corpus. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995), Detroit, MI, USA, 8–12 May 1995; pp. 341–344.
70. Yamagishi, J.; Veaux, C.; MacDonald, K. *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)*; The Centre for Speech Technology Research (CSTR), University of Edinburgh: Edinburgh, Scotland, 2019.
71. Kreuk, F.; Adi, Y.; Cisse, M.; Keshet, J. Fooling end-to-end speaker verification with adversarial examples. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada, 15–20 April 2018; pp. 1962–1966.
72. Abdullah, H.; Rahman, M.S.; Garcia, W.; Warren, K.; Yadav, A.S.; Shrimpton, T.; Traynor, P. Hear "No Evil", See "Kenansville"*: Efficient and Transferable Black-Box Attacks on Speech Recognition and Voice Identification Systems. In Proceedings of the 42nd IEEE Symposium on Security and Privacy (SP 2021), San Francisco, CA, USA, 24–27 May 2021; pp. 712–729.
73. Li, X.; Zhong, J.; Wu, X.; Yu, J.; Li, X.; Meng, H. Adversarial attacks on GMM i-vector based speaker verification systems. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2020), Barcelona, Spain, 4–8 May 2020; pp. 6579–6583.
74. Villalba, J.; Zhang, Y.; Dehak, N. x-vectors Meet Adversarial Attacks: Benchmarking Adversarial Robustness in Speaker Verification. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020), Shanghai, China, 25–29 October 2020; pp. 4233–4237.
75. Jati, A.; Hsu, C.; Pal, M.; Peri, R.; AbdAlmageed, W.; Narayanan, S. Adversarial attack and defense strategies for deep speaker recognition systems. *Comput. Speech Lang.* **2021**, *68*, 101199. [CrossRef]
76. Joshi, S.; Villalba, J.; Zelasko, P.; Moro-Velázquez, L.; Dehak, N. Study of Pre-Processing Defenses Against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4811–4826. [CrossRef]
77. Zhang, W.; Zhao, S.; Liu, L.; Li, J.; Cheng, X.; Zheng, T.; Hu, X. Attack on Practical Speaker Verification System Using Universal Adversarial Perturbations. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2021), Toronto, ON, Canada, 6–11 June 2021; pp. 2575–2579.
78. Liu, S.; Wu, H.; Lee, H.Y.; Meng, H. Adversarial attacks on spoofing counter measures of automatic speaker verification. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2019), Singapore, 14–18 December 2019; pp. 312–319.
79. Zhang, Y.; Jiang, Z.; Villalba, J.; Dehak, N. Black-Box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020), Shanghai, China, 25–29 October 2020; pp. 4238–4242.
80. Goto, K.; Inoue, N. Quasi-Newton Adversarial Attacks on Speaker Verification Systems. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2020), Auckland, New Zealand, 7–10 December 2020; pp. 527–531.
81. Chen, G.; Chenb, S.; Fan, L.; Du, X.; Zhao, Z.; Song, F.; Liu, Y. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems. In Proceedings of the 42nd IEEE Symposium on Security and Privacy (SP 2021), San Francisco, CA, USA, 24–27 May 2021; pp. 694–711.
82. Li, J.; Zhang, X.; Xu, J.; Ma, S.; Gao, W. Learning to Fool the Speaker Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4–8 May 2020; pp. 2937–2941.
83. Luo, H.; Shen, Y.; Lin, F.; Xu, G. Spoofing Speaker Verification System by Adversarial Examples Leveraging the Generalized Speaker Difference. *Secur. Commun. Netw.* **2021**, *2021*, 6664578. [CrossRef]
84. Marras, M.; Korus, P.; Memon, N.D.; Fenu, G. Adversarial Optimization for Dictionary Attacks on Speaker Verification. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019), Graz, Austria, 15–19 September 2019; pp. 2913–2917.

85. Zhang, L.; Meng, Y.; Yu, J.; Xiang, C.; Falk, B.; Zhu, H. Voiceprint mimicry attack towards speaker verification system in smart home. In Proceedings of the 39th IEEE Conference on Computer Communications (INFOCOM 2020), Toronto, ON, Canada, 6–9 July 2020; pp. 377–386.
86. Abdullah, H.; Garcia, W.; Peeters, C.; Traynor, P.; Butler, K.R.; Wilson, J. Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems. In Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019), San Diego, CA, USA, 24–27 February 2019.
87. Du, T.; Ji, S.; Li, J.; Gu, Q.; Wang, T.; Beyah, R. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In Proceedings of the 15th ACM Asia Conference on Computer and Communications Security (ASIA CCS 2020), Taipei, Taiwan, 5–9 October 2020; pp. 357–369.
88. Li, Z.; Wu, Y.; Liu, J.; Chen, Y.; Yuan, B. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS 2020), Virtual Event, 9–13 November 2020; pp. 1121–1134.
89. Xie, Y.; Shi, C.; Li, Z.; Liu, J.; Chen, Y.; Yuan, B. 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 2020), Barcelona, Spain, 4–8 May 2020; pp. 1738–1742.
90. Chen, G.; Zhao, Z.; Song, F.; Chen, S.; Fan, L.; Liu, Y. AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems. *arXiv* **2022**, arXiv:2206.03351.
91. Zheng, B.; Jiang, P.; Wang, Q.; Li, Q.; Shen, C.; Wang, C.; Ge, Y.; Teng, Q.; Zhang, S.; Zhang, S. Black-box adversarial attacks on commercial speech platforms with minimal information. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, 15–19 November 2021; pp. 86–107.
92. Li, Z.; Shi, C.; Xie, Y.; Liu, J.; Yuan, B.; Chen, Y. Practical Adversarial Attacks Against Speaker Recognition Systems. In Proceedings of the HotMobile '20: The 21st International Workshop on Mobile Computing Systems and Applications, Austin, TX, USA, 3–4 March 2020; pp. 9–14.
93. Li, J.; Zhang, X.; Jia, C.; Xu, J.; Zhang, L.; Wang, Y.; Ma, S.; Gao, W. Universal Adversarial Perturbations Generative Network For Speaker Recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2020, London, UK, 6–10 July 2020; pp. 1–6.
94. Xie, Y.; Li, Z.; Shi, C.; Liu, J.; Chen, Y.; Yuan, B. Real-time, Robust and Adaptive Universal Adversarial Attacks Against Speaker Recognition Systems. *J. Signal Process. Syst.* **2021**, *93*, 1187–1200. [CrossRef]
95. Tan, H.; Zhang, J.; Zhang, H.; Wang, L.; Qian, Y.; Gu, Z. NRI-FGSM: An Efficient Transferable Adversarial Attack Method for Speaker Recognition System. In Proceedings of the 23st Annual Conference of the International Speech Communication Association (Interspeech 2022), Incheon, Korea, 18–22 September 2022.
96. Shamsabadi, A.S.; Teixeira, F.S.; Abad, A.; Raj, B.; Cavallaro, A.; Trancoso, I. FoolHD: Fooling Speaker Identification by Highly Imperceptible Adversarial Disturbances. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP 2021), Toronto, ON, Canada, 6–11 June 2021; pp. 6159–6163.
97. Wang, Q.; Guo, P.; Xie, L. Inaudible adversarial perturbations for targeted attack in speaker recognition. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020), Shanghai, China, 25–29 October 2020; pp. 4228–4232.
98. Xie, Y.; Li, Z.; Shi, C.; Liu, J.; Chen, Y.; Yuan, B. Enabling fast and universal audio adversarial attack using generative model. In Proceedings of the AAAI Conference on Artificial Intelligence (EAAI 2021), Virtual Event, 2-9 February 2021; pp. 14129–14137.
99. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018.
100. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. In Proceedings of the Workshop of the 5th International Conference on Learning Representations (ICLR-2017), Toulon, France, 24–26 April 2017.
101. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, V. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations (ICLR-2018), Vancouver, BC, Canada, 30 April–3 May 2018.
102. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 9185–9193.
103. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-box Adversarial Attacks with Limited Queries and Information. In Proceedings of the 35th International Conference on Machine Learning (ICML 2018), Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; pp. 2142–2151.
104. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec@CCS 2017), Dallas, TX, USA, 3 November 2017; pp. 15–26.
105. Stan, G.B.; Embrechts, J.J.; Archambeau, D. Comparison of different impulse response measurement techniques. *J. Audio Eng. Soc.* **2002**, *50*, 249–262.
106. Machado, G.R.; Silva, E.; Goldschmidt, R.R. Adversarial Machine Learning in Image Classification: A Survey Toward the Defender's Perspective. *ACM Comput. Surv. (CSUR)* **2021**, *55*, 1–38. [CrossRef]

107. Zhang, J.; Tan, H.; Deng, B.; Hu, J.; Zhu, D.; Huang, L.; Gu, Z. NMI-FGSM-Tri: An Efficient and Targeted Method for Generating Adversarial Examples for Speaker Recognition. In Proceedings of the Sixth IEEE International Conference on Data Science in Cyberspace (DSC 2022), Guilin, China, 11–13 July 2022.
108. Hu, Y.; Loizou, P.C. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Trans. Speech Audio Process.* **2008**, *16*, 229–238. [CrossRef]
109. Amazon Mechanical Turk Platform. Available online: https://www.mturk.com (accessed on 2 November 2005).
110. Delgado, H.; Evans, N.; Kinnunen, T.; Lee, K.A.; Liu, X.; Nautsch, A.; Patino, J.; Sahidullah, M.; Todisco, M.; Wang, X.; et al. ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. *arXiv* **2021**, arXiv:2109.00535.
111. Akhtar, N.; Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* **2018**, *6*, 14410–14430. [CrossRef]
112. Wang, Q.; Guo, P.; Sun, S.; Xie, L.; Hansen, J.H. Adversarial Regularization for End-to-End Robust Speaker Verification. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019), Graz, Austria, 15–19 September 2019; pp. 4010–4014.
113. Wu, H.; Liu, S.; Meng, H.; Lee, H.Y. Defense against adversarial attacks on spoofing countermeasures of ASV. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4–8 May 2020; pp. 6564–6568.
114. Pal, M.; Jati, A.; Peri, R.; Hsu, C.C.; AbdAlmageed, W.; Narayanan, S. Adversarial defense for deep speaker recognition using hybrid adversarial training. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021), Toronto, ON, Canada, 6–11 June 2021; pp. 6164–6168.
115. Li, X.; Li, N.; Zhong, J.; Wu, X.; Liu, X.; Su, D.; Yu, D.; Meng, H. Investigating robustness of adversarial samples detection for automatic speaker verification. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020), Shanghai, China, 25–29 October 2020; pp. 1540–1544.
116. Villalba, J.; Joshi, S.; Żelasko, P.; Dehak, N. Representation Learning to Classify and Detect Adversarial Attacks against Speaker and Speech Recognition Systems. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (Interspeech 2021), Brno, Czechia, 30 August–3 September 2021; pp. 4304–4308.
117. Peng, Z.; Li, X.; Lee, T. Pairing weak with strong: Twin models for defending against adversarial attack on speaker verification. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (Interspeech 2021), Brno, Czechia, 30 August–3 September 2021; pp. 4284–4288.
118. Wu, H.; Hsu, P.C.; Gao, J.; Zhang, S.; Huang, S.; Kang, J.; Wu, Z.; Meng, H.; Lee, H.Y. Spotting adversarial samples for speaker verification by neural vocoders. *arXiv* **2021**, arXiv:2107.00309.
119. Joshi, S.; Kataria, S.; Villalba, J.; Dehak, N. AdvEst: Adversarial Perturbation Estimation to Classify and Detect Adversarial Attacks against Speaker Identification. *arXiv* **2022**, arXiv:2204.03848.
120. Zhang, H.; Wang, L.; Zhang, Y.; Liu, M.; Lee, K.A.; Wei, J. Adversarial Separation Network for Speaker Recognition. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020), Shanghai, China, 25–29 October 2020; pp. 951–955.
121. Wu, H.; Liu, A.T.; Lee, H.Y. Defense for black-box attacks on anti-spoofing models by self-supervised learning. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020), Shanghai, China, 25–29 October 2020; pp. 3780–3784.
122. Wu, H.; Li, X.; Liu, A.T.; Wu, Z.; Meng, H.; Lee, H.Y. Improving the adversarial robustness for speaker verification by self-supervised learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 202–217. [CrossRef]
123. Wu, H.; Zhang, Y.; Wu, Z.; Wang, D.; Lee, H.Y. Voting for the right answer: Adversarial defense for speaker verification. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (Interspeech 2021), Brno, Czechia, 30 August–3 September 2021; pp. 4294–4298.
124. Wu, H.; Li, X.; Liu, A.T.; Wu, Z.; Meng, H.; Lee, H.Y. Adversarial defense for automatic speaker verification by cascaded self-supervised learning models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021), Toronto, ON, Canada, 6–11 June 2021; pp. 6718–6722.
125. Olivier, R.; Raj, B.; Shah, M. High-Frequency Adversarial Defense for Speech and Audio. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021), Toronto, ON, Canada, 6–11 June 2021; pp. 2995–2999.
126. Chang, L.C.; Chen, Z.; Chen, C.; Wang, G.; Bi, Z. Defending Against Adversarial Attacks in Speaker Verification Systems. In Proceedings of the IEEE International Performance, Computing, and Communications Conference (IPCCC 2021), Austin, TX, USA, 29–31 October 2021; pp. 1–8.
127. Miyato, T.; Dai, A.M.; Goodfellow, I. Adversarial training methods for semi-supervised text classification. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017.
128. Miyato, T.; Maeda, S.I.; Koyama, M.; Nakae, K.; Ishii, S. Distributional smoothing with virtual adversarial training. *arXiv* **2015**, arXiv:1507.00677.
129. Wan, L.; Wang, Q.; Papir, A.; Moreno, I.L. Generalized end-to-end loss for speaker verification. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada, 15–20 April 2018; pp. 4879–4883.

130. Zhang, H.; Wang, J. Defense against adversarial attacks using feature scattering-based adversarial training. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 1829–1839.
131. Yamamoto, R.; Song, E.; Kim, J.M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020) , Barcelona, Spain, 4–8 May 2020; pp. 6199–6203.
132. Zhu, X.; Beauregard, G.T.; Wyse, L.L. Real-time signal estimation from modified short-time Fourier transform magnitude spectra. *IEEE Trans. Speech Audio Process.* **2007**, *15*, 1645–1653. [CrossRef]
133. Hubert, M.; Debruyne, M.; Rousseeuw, P.J. Minimum covariance determinant and extensions. *Wiley Interdiscip. Rev. Comput. Stat.* **2018**, *10*, e1421. [CrossRef]
134. Chen, Z. On the Detection of Adaptive Adversarial Attacks in Speaker Verification Systems. *arXiv* **2022**, arXiv:2202.05725.
135. Yang, Z.; Li, B.; Chen, P.Y.; Song, D. Characterizing audio adversarial examples using temporal dependency. In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019.
136. Liu, A.T.; Yang, S.W.; Chi, P.H. Hsu, P.C.; Lee, H.Y. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4–8 May 2020; pp. 6419–6423.
137. AdvBox. Available online: https://github.com/Ewenwan/AdvBox (accessed on 4 September 2018).
138. TorchAttack. Available online: https://adversarial-attacks-pytorch.readthedocs.io/en/latest (accessed on 15 July 2020).
139. FoolBox. Available online: https://github.com/bethgelab/foolbox (accessed on 22 September 2020).
140. TextAttack. Available online: https://github.com/qdata/textattack (accessed on 20 November 2020).
141. OpenAttack. Available online: https://github.com/thunlp/openattack (accessed on 6 August 2021).