APSIPA Transactions on Signal and Information Processing, 2022, 11, e39

Overview Paper

An Application-Oriented Taxonomy on Spoofing, Disguise and Countermeasures in Speaker Recognition

Lantian Li[1,2], Xingliang Cheng[1] and Thomas Fang Zheng[1*]

[1] *Center for Speech and Language Technologies, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China*

[2] *School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China*

ABSTRACT

Speaker recognition aims to recognize the identity of the speaking person. After decades of research, current speaker recognition systems have achieved rather satisfactory performance, and have been deployed in a wide range of practical applications. However, a massive amount of evidence shows that these systems are susceptible to malicious fake actions in real applications. To address this issue, the research community has been responding with dedicated countermeasures which aim to defend against fake actions. Recently, there are several reviews and surveys reported in the literature that describe the current state-of-the-art research advancements. Even so, these reviews and surveys are generally based on a canonical taxonomy to categorize spoofing attacks and corresponding countermeasures from the technology-oriented perspective. This

[*] Corresponding author: Thomas Fang Zheng, fzheng@tsinghua.edu.cn. This work was supported by the National Natural Science Foundation of China under Grants no. 61633013, and also the Tsinghua-SPD Bank Joint Research Center for Digital Finance Technologies.

Received 12 April 2022; Revised 31 May 2022
ISSN 2048-7703; DOI 10.1561/116.00000017

paper provides a new taxonomy from the application-oriented perspective and extends to two major fake forms: spoofing attack and disguise cheating. This taxonomy starts from the applications of speaker recognition technology, e.g., access control, surveillance and forensic, and then rezones two fake forms according to different application scenarios: one is spoofing attack that imitates the voice of an authorized speaker to get access to the target system; the other one is disguise cheating that makes someone unrecognizable by altering his/her voice. Furthermore, for each fake form, more delicate categories and related countermeasures are presented. Finally, this paper discusses future research directions in this area and suggests that the research community should not only focus on the technical view but also connect with application scenarios.

Keywords: Speaker recognition, spoofing, disguise, countermeasures, application.

1 Introduction

Biometrics is a measurable physiological or behavioral characteristic that can be used for automated recognition. When biometrics is used to narrate a process, i.e., biometric recognition, it refers to an automated technique of recognizing an individual based on its characteristics [5, 6, 32]. Since biometrics represents the inherent characteristics of a person and also has the attributes of distinctness, uniqueness, stability and non-reproducibility, one can "*authenticate self by self*" in anytime and anywhere. Compared to the conventional authentication approaches (such as password, IC card, and USB key), the emergence of biometric recognition technology provides enhanced security and more convenience [139] and achieves a wide range of applications, such as fingerprint recognition in immigration control, face recognition in smartphone login, speaker recognition in remote banking.

Generally, biometrics can be categorized into two main types: physiological biometrics and behavioral biometrics [88]. The former refers to the *static and distinct* characteristics that are related to an individual physical body shape like face, eye (iris and retina), fingerprint, palm and so on. The latter refers to the *dynamic and unique* characteristics that are related to individual behavioral patterns like keystroke, signature, motion, gait and so on.

Besides the two types, voiceprint biometrics is regarded as a coalition of physiological and behavioral characteristics [33, 139]. From the physiological view, voice contains speaker-distinct characteristic according to the differences of the shapes and sizes of the speech production organs (e.g., vocal tract, larynxes, lungs, nasal cavity, and others) among different speakers. From the behavioral view, each speaker owns his/her unique speaking manner, such as the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on [43]. Voiceprint recognition, commonly known as speaker recognition, is a biometric modality that uses both physiological and behavioral characteristics to recognize the identity of the speaking person. Furthermore, due to its advantages of easy use, non-intrusive, non-touching and low privacy leakage, speaker recognition has been regarded as a new generation biometric technique. With theses advantages, speaker recognition has found broad deployment in real-life applications, such as access control, surveillance and forensic [27, 65].

Despite the benefits brought by speaker recognition technology, it is vulnerable to fake actions in practice. A canonical definition of fake action refers to an adversary (attacker) who counterfeits as the target speaker to get access to a system, also denoted as spoofing attack [21, 22, 38, 62, 100]. Depending upon how spoofing attacks are presented to speaker recognition systems, there are two attack types: direct (also known as Physical Access (PA)) attacks and indirect (also known as Logical Access (LA)) attacks [70, 127]. The PA attacks occur at the system sensor (i.e., microphone and transmission level) while LA attacks involve by passing the sensor (i.e., system level). Another view to categorize spoofing attacks is based on the attacker variations which consist of four types: Speaker Imitation, Replay Attack (RA), Speech Synthesis (SS) and Voice Conversion (VC). Besides, Adversarial Attack (AA) has emerged to be a new attacker and attracted much research effort [34, 59, 126, 136].

Due to the facile of obtaining speech data via a mass of speech media sources by replay attack and the advance of speech synthesis and voice conversion technologies, countermeasures against spoofing attacks are necessary to ensure the security of systems. In recent years, several efforts have been fostered to the development of spoofing countermeasures or Presentation Attack Detection (PAD) solutions for speaker recognition. Perhaps, the most successful effort is the ASVspoof challenge series. There are a total of four ASVspoof challenges up-to-date, including ASVspoof 2015 [124], ASVspoof 2017 [44], ASVspoof 2019 [127], and ASVspoof 2021 [128]. The aim of ASVspoof challenge series is to promote the development of spoofing and countermeasures for automatic speaker verification (ASV), a major branch of speaker recognition. This challenge series provides a level playing field to facilitate the comparison of different spoofing countermeasures on standard datasets, protocols and metrics. Thanks to this challenge series, lots of novel countermeasures have been fostered against different spoofing attacks.

Recently, there are several reviews conducted on the developments of spoofing and countermeasures in speaker recognition, particularly in the ASV domain. [3, 11, 123] presented an overview of spoofing and countermeasures in ASV systems. [38, 104, 128] successively provided a review of techniques from the ASVspoof challenge perspective that includes protocols, databases and future directions. [91, 100] focused on the overview of PAD, and [79, 94] focused only on the review of replay attacks.

In short, there have been a large number of technical reviews covering this field. Most of them were from the *technology-oriented* perspective and were more suitable for practitioners who specialize in the speaker recognition community. Unlike these reviews, this paper presents a new taxonomy on fake actions from the *application-oriented* perspective. It starts from the applications of speaker recognition technology and then presents the basic concept of each kind of fake technique and its representative countermeasures. This application-oriented taxonomy will be easier for readers to understand the topic of fake actions in real application scenarios. To the best of our knowledge, this is the first work towards this taxonomy. The main contribution of this paper is three-fold:

- This paper concentrates on different fake actions in speaker recognition from different application scenarios, including access control, surveillance and forensic. There are two major fake forms: spoofing attack and disguise cheating.
- According to the production mode and evaluation subject, fake techniques in each form are further classified into different types. Representative countermeasures against each kind of fake technique are finally presented.
- This paper focuses more on macroscopic categories from the application perspective rather than in-depth understandings from the technical perspective. The readers of this paper could be experts/technologists specialized in this field or preliminary students/engineers who are interested.

The paper is organized as follows: Section 2 presents our proposed taxonomy on spoofing attack and disguise cheating in speaker recognition from the application-oriented perspective. Sections 3 and 4 briefly review the countermeasures against spoofing attack and disguise cheating, respectively. Section 5 discusses the future research direction in this area and Section 6 concludes the paper.

2 Our Taxonomy

In this section, we will present our taxonomy on spoofing attack and disguise cheating in speaker recognition according to real application scenarios. Firstly, this taxonomy starts from three widely deployed applications of speaker recognition technology, i.e., access control, surveillance and forensic. Secondly, we

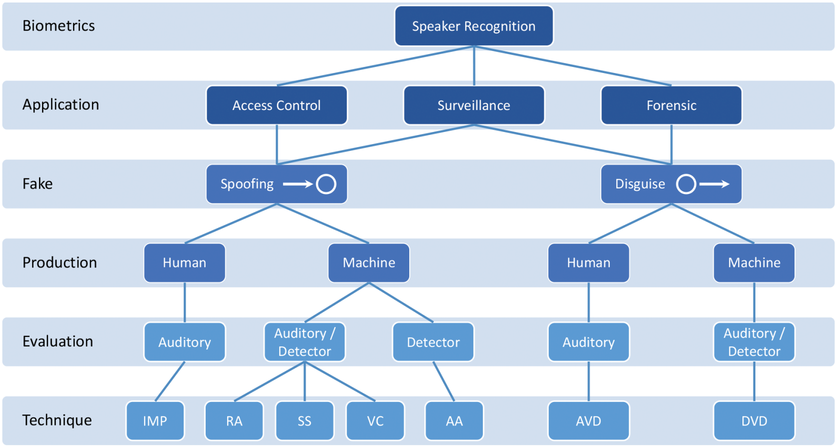

Figure 1: Our taxonomy on spoofing attack and disguise cheating in speaker recognition applications. →○ illustrates the objective of spoofing attack which aims to *get close to* the target speaker; ○→ illustrates the objective of disguise cheating which aims to *keep far away from* the target speaker. IMP: Impersonation; RA: Replay Attack; SS: Speech Synthesis; VC: Voice Conversion; AA: Adversarial Attack; AVD: Artificial Voice Disguise; DVD: Digital Voice Disguise.

rezone two forms of fake actions according to different application scenarios: spoofing attack and disguise cheating. Finally, according to the production mode (via human or machine) and evaluation subject (by auditory perception or automatic detector), fake actions in each form are further classified into different technical types. Figure 1 illustrates our taxonomy on fake actions of speaker recognition in applications.

2.1 Speaker Recognition and Its Applications

Speaker recognition has been employed in broad application areas, such as access control, surveillance and forensic due to its high accuracy rate, natural interaction, non-touching, low privacy leakage and low-cost devices. In this subsection, we briefly present the three most common applications of speaker recognition.

- Access control: it is one of the most popular biometric applications as it allows the users to identify an authorized individual based on his/her voice. It has broad deployment in security and financial services.

- Surveillance: it is mainly important for security agencies to collect important information, such as electronic eavesdropping on telephone and radio conversations. For instance, recognizing the target speakers who are of interest, and monitoring parolees at a random time to verify that they are in the restricted area.

- Forensic: it can be used to automatically compare the speech sample recorded during the crime and the suspect's voice. The comparison result can be regarded as auxiliary evidence to prove the identity of the suspect and confirm a judgment of guilt or innocence.

2.2 Fake Forms in Applications

Among these application scenarios, there are two major fake forms: spoofing attack and disguise cheating. The former (spoofing attack) aims to imitate the voice of an authorized speaker to get access to the intended system, while the latter (disguise cheating) aims to make someone unrecognizable by altering his/her voice. For each form, in terms of the production mode and evaluation subject, it can be further categorized into different technical types. It should be mentioned that we argue that spoofing attack and disguise cheating may be produced from either human or machine in terms of the production mode, and their corresponding countermeasures could be either by human auditory or by automatic detectors in terms of the evaluation subject.

2.2.1 Spoofing Attack

The objective of spoofing attack in speaker recognition applications, e.g., access control and surveillance, is to counterfeit an authorized individual in an identity authentication system to bypass and get access to the intended system. Generally, the identity authentication system is constructed by an automatic speaker verification technique which decides if an identity claim is true or false. Furthermore, according to the production mode via either human or machine and the evaluation subject by either the auditory perception of listeners or automatic detector via algorithms, spoofing attacks in these applications can be further categorized into three groups: (1) Impersonation; (2) Replay attack, Speech synthesis and Voice conversion; (3) Adversarial attack.

Group 1: Impersonation. It is produced by humans and evaluated by the auditory perception of listeners Impersonation (IMP) is defined as a human-spontaneous production mode of producing the similar voice pattern and speech behavior of the target speaker's voice [28, 50, 64]. This can be done either by professional mimics/imitators (by utilizing behavioral characteristics) or by twins (by utilizing physiological characteristics). For professional imitators, they intend to mimic the claimed speaker's prosody, accent, pronunciation, lexicon, and other high-level speaker traits. For twins, the pattern of speech signals, pitch contours, formant contours, and spectrograms for identical twin speakers are very similar. Recently, a BBC news reported that one non-identical twin could successfully bypass the speaker verification system and access the other twin's bank account [107]. Hence, the threats that impersonation posed to speaker verification systems must not be underestimated. As impersonation does not require any technical background or machines to imitate the target speaker, it is also denoted as a zero-effort human attack.

Group 2: Replay attack, Speech synthesis and Voice conversion. They are produced by machines and evaluated by either the auditory perception of listeners or automatic detector via algorithms

- Replay Attack (RA) is performed by a recording and replay process. The attacker firstly records the voice of the target speaker and then replays the recorded speech to a speaker verification system to gain access. The speech may be concatenated or clipped to obtain the desired utterance. Replay is regarded as the most common type of spoofing attacks [2, 60]. On one hand, it is very easy to conduct by anyone using recording and replay devices, such as a smartphone; compared to the other three attacks as shown below, the replay attack requires no prior knowledge of signal processing. On the other hand, with high-quality recording and playback devices, the replayed speech is highly similar to the original speech, leading to a serious practical risk to speaker verification systems.

- Speech Synthesis (SS) attack is performed by the text-to-speech (TTS) system that takes a prompted text as input and generates a speech of the target speaker. SS is now able to generate high-quality voice due to recent advances in unit selection [26, 31], statistic model [105, 106, 131] and deep learning [73, 102]. Recently, deep learning-based techniques, including multi-speaker TTS and one-shot personalized TTS based on speaker embedding [37, 46, 76, 82, 115], are able to produce very natural sounding speech both in timbre and prosody. Based on these techniques, given an utterance or a speech segment from the target speaker, SS systems can produce high-quality speech to spoof speaker verification systems.

- Voice Conversion (VC) attack is performed by the voice conversion system that converts the voice of an attacker into the voice of the target speaker while preserving the linguistic content [7, 45, 125]. Modern voice conversion methods [39–41, 68, 83, 95] are advanced from statistical modeling to deep learning based on large-scale training with non-parallel data. The basic idea is to learn a disentanglement model that can separate content and speaker information in speech signals, and then perform the conversion by picking up content information from the source speaker and speaker information from the target speaker. With the development of VC techniques, it can achieve real-time voice conversion meanwhile offer distinguished voice quality. This has become one of the most easily accessible techniques to carry out spoofing attacks against speaker verification systems.

Group 3: Adversarial attack. It is produced by machines and evaluated by automatic detectors via algorithms Adversarial Attack (AA) is performed by involving adversarial perturbations to the voice of an attacker; these adversarial perturbations are imperceptible to human auditory perception, but can easily fool the DNN-based speaker recognition system and misclassify the attacker into the target speaker. The basic idea of AA is to destroy the posterior distribution of DNN-based models by only perturbing the input samples by a very small amount [12, 129]. AA can be further categorized into two scenarios: white-box attack with the prior knowledge of the model's internals and parameters, and black-box attack without any prior knowledge of the model. Recent research [34, 47, 52, 56, 59] has shown that both the white-box and black-box attacks show a great threat to the modern DNN-based speaker recognition models.

2.2.2 Disguise Cheating

Speaker recognition in forensic [10, 86, 87] is a popular application. For instance, if there is a speech sample recorded during the crime, then the suspect's voice can be compared. The result can prove the identity of the criminal and discharge the innocent during a court case. However, the suspect (cheater) could conceal his/her real identity by *deliberate* voice disguise [23, 81, 132]. Note that there is another *non-deliberate* voice disguise aspect, which refers to the voice variation due to the speaker-related factors (such as aging, illness, and emotional stress) or the speech distortion due to the environment-related effects (such as transmission channel and background noise). We argue that this non-deliberate voice disguise is more like a robust challenge rather than fake cheating for speaker recognition. Therefore, we do not consider this aspect and only focus on deliberate voice disguise. In terms of the production mode and evaluation subject, the deliberate voice disguise can be further classified into two categories: Artificial voice disguise by humans and Digital voice disguise by machines.

Group 1: Artificial voice disguise. It is produced by humans and evaluated by the auditory perception of listeners Artificial Voice Disguise (AVD) refers to distorting the voice by changing the speaker's vocal track and pronunciation manner, such as raising the pitch and pinched nostrils. This distortion can produce the variation of both the low-level acoustic characteristics (such as fundamental frequency, formant and bandwidth) and the high-level linguistic characteristics (such as accent, dialect and prosody), which makes speaker recognition become more difficult and even impossible [66, 85]. Therefore, it is necessary to study the effect of AVD in speaker recognition.

Group 2: Digital voice disguise. It is produced by humans and evaluated by either the auditory perception of listeners or automatic detector via algorithms Digital Voice Disguise (DVD) aims to automatically alter or modify voices by algorithms in order to hide the real identity of a speaker. It is often classified into two categories: Voice Transformation (VT) and Voice Conversion (VC) [114]. VC, as mentioned in Section 2.2.1, intends to transform a source speaker's voice to sound like a target speaker's voice. Unlike VC, VT is defined as a technique of changing the voice without the intention of any target speaker, for example, scaling the pitch by frequency warping or temporal stretching. It is apparent that VC is to change one's voice in order "to be recognized as another person" while VT is to change one's voice in order "not to be recognized". Since no target speaker is required, VT is much easier to implement than VC in practice, leading to the fact that VT has been incorporated in many prevailing audio editors and becomes a non-negligible threat in forensic applications. Therefore, in this paper, we only discuss VT in DVD.

2.3 Summary

Section 2.2 presents our taxonomy on two fake forms in speaker recognition applications. For clear presentation, in this subsection, we immigrate these fake actions into a typical speaker recognition system in terms of the possible fake points, as illustrated in Figure 2.

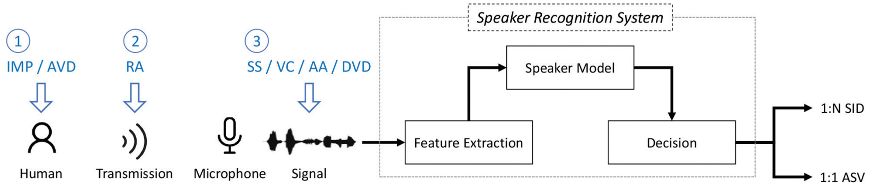

Figure 2: Possible fake points and corresponding fake techniques in speaker recognition systems (SID: speaker identification, ASV: automatic speaker verification). Human: attacker in spoofing or cheater in disguise. IMP: Impersonation; RA: Replay Attack; SS: Speech Synthesis; VC: Voice Conversion; AA: Adversarial Attack; AVD: Artificial Voice Disguise; DVD: Digital Voice Disguise.

A typical speaker recognition system generally involves three components, as shown in the block diagram:

- Feature extraction: extract features from speech signals.
- Speaker model: construct speaker models from extracted features.
- Decision: measure the similarity between the user's voice and pre-enrolled speaker models to accomplish automatic speaker verification or speaker identification tasks.

In real application, the system itself is considered to be indestructible in most cases. Therefore, fake actions are usually carried out before injecting the system. There are three possible fake points:

- Faker self: a faker impersonates the target speaker (IMP) or deliberately disguises his/her voice (AVD).
- Transmission process: a faker pre-records a voice from a target speaker and then replays it back (RA).
- Speech signal: a faker produces or modifies speech signal with the help of automatic speech processing techniques (SS, VC, AA, and DVD).

It should be mentioned that in Figure 2, there is an assumption that the fake processes of different fake techniques are completely independent. In other words, it does not consider the possibility of combining different techniques to spoof/disguise systems. For example, SS/VC/AA/DVD can also take place at the transmission side when playing generated speech samples in front of a microphone. To better illustrate the natural attributes of each fake technique, here we prefer to omit this complex situation.

3 Countermeasures against Spoofing Attacks

Recently, spoofing countermeasures have attracted a lot of interest in both research and industry communities. This is largely attributed to the effort of ASVspoof challenges [44, 124, 127, 128]. ASVspoof collects and distributes standard datasets, evaluation protocols and metrics, and facilitates competitions to explore effective spoofing countermeasures. Considering a large amount of literature such as reviews and surveys covering this area, we therefore simply present some highly-effective and widely-adopted methods. More technical descriptions in this area are reported in [38, 62, 67, 100].

3.1 Countermeasures against Impersonation Attack

Due to the lack of a standardized public dataset for impersonation attacks, there is barely any research conducted on detecting impersonation attacks in speaker verification systems. More unfortunately, research has presented completely opposite conclusions. [49, 50] showed that an impersonation attack has a chance to attack the speaker verification system even by a non-professional imitator. However, other research has shown the opposite [63]. As a whole, research on countermeasures against impersonation attack is still limited. In a recent work on detecting speech impersonation [72], genuine speech and imitated speech was collected from two celebrity speakers, Mel-frequency cepstral coefficients

(MFCC) was used as the input feature, and a convolutional neural network was used as the classifier. The results in terms of equal error rate on impersonation detection was 35.85%, indicating that there is a need to develop a robust spoofing countermeasures against impersonation.

3.2 Countermeasures against Replay Attack

Recent work on spoofing countermeasures against replay attack is mainly focused on two directions: detection via signal distortion and detection via additional factors. The former exploits the signal distortion caused by the recording and replay process. The latter uses an additional source of information to detect whether the input speech comes from a human or a machine, also denoted as liveness detection.

3.2.1 Detection via Signal Distortion

1. Feature: Many features are elaborately designed to describe the distortion between genuine speech and replayed speech. The representative magnitude-based features include constant-Q cepstral coefficients (CQCC) [103], Mel-frequency cepstral coefficients (MFCC) [71], inverted Mel-frequency cepstral coefficients [53], linear frequency cepstral coefficients [29], linear predictive cepstral coefficients [1, 116], perceptual linear predictive analysis [1], power-normalized cepstral coefficients [42]. Other phase-based features include group delay function [9], modified group delay function [16], relative phase [75], and so on.

2. Model: Regarding the modeling approach, the Gaussian mixture model (GMM) is among the most popular ones. Recently, more and more research focuses on the end-to-end learning by deep neural nets. For instance, LCNN [51], ResNet [14], Res2Net [55], DenseNet [29], SincNet [130]. The essence of all these models is a two-class model to discriminate the genuine speech and replayed speech. Except for these two-class model, recent research advocated a one-class view for replay detection, by which only genuine speech is modeled, and replay detection is formulated as out-of-distribution detection [4, 15, 17, 110].

3.2.2 Detection via Liveness Detection

A group of work focus on detecting whether the input speech comes from a human or a machine, the so-called liveness detection. To achieve this goal, an additional sensor or device is involved, such as an airflow sensor [111] to detect the airflow, throat microphone [92] to detect the throat vibration, magnetometer sensor [13] to detect the machine-produced spoofed speech, smartphone audio system [133, 134] as a sonar to detect the user's articulatory gestures, or the dissimilarities between air-conducted voices and bone-conducted vibrations.

3.3 Countermeasures against Speech Synthesis and Voice Conversion

Speech Synthesis (SS) and Voice Conversion (VC) were grouped as one subcategory due to the two attacks were similar. They often require the use of an audio processor called vocoder to produce artificial voice. As these attacks require knowledge of signal processing, assistance from professionals may be needed. Two systematic reviews on logical access detection are reported in [67, 109].

In the first stage, spoofing countermeasures against SS and VC attacks mostly depend on the attribute of particular vocoders. The human auditory system is known to be relatively insensitive to phase whereas these vocoders are typically based on a minimum-phase vocal tract model. Such differences in the phase spectra are used as a feature to detect the synthetic speech [20]. Similarly, a cosine normalization phase spectrum (CosPhase) along with modified group delay function was proposed to detect converted speech [122]. Apart from phase-based features, lots of magnitude-based features and other distinct features were also proposed [109].

In the second stage, with the advance of deep learning in SS and VC, deep neural networks (DNNs) have significantly improved the speech quality in an end-to-end way. To pay these attacks back in the same coin, research welcomed various DNN architectures to detect these attacks [9, 30, 84, 98, 137]. Such DNN-based countermeasures are preferred for two main reasons. Firstly, the data-driven DNNs can automatically learn both short-term and long-term features from the raw speech signal. Secondly, DNNs with non-linear transformations can capture fine-grained difference between genuine speech and spoofed speech. More technical reviews in this area are reported in [67]. Besides, recent studies explored that these DNN-based countermeasures tend to overfit on the training data and fail to generalize. To improve their generalizability, a series of data argumentation methods were employed, such as signal compression [19, 78], linear and non-linear convolutive noise [99] and additive noise [99].

3.4 Countermeasures against Adversarial Attack

In recent years, deep learning models have achieved state-of-the-art performance on several benchmark datasets [8, 18, 69, 89, 90]. In spite of the great success, these deep learning models are recently found to be vulnerable to Adversarial Attack (AA). The attacker potentially discovers blind spots in the model, and crafts adversarial samples that are composed of the normal speech signals and inconspicuous adversarial perturbations. These adversarial samples are easy to attack the well-trained deep speaker models, such as d-vector [108] and x-vector [96].

Lots of AA algorithms have been validated in deep speaker models, such as the fast gradient sign method [80, 118, 119], basic iterative method [93, 118], projected gradient descent [24, 35, 136] and so on.

To address concerns on AA, many spoofing countermeasures have emerged recently. These countermeasures can be categorized into two themes. The first theme aims to explore the difference between genuine speech and adversarial sample and construct detection models. [117] adopted the neural vocoder to re-synthesize speech and found that the difference between the decision scores of the original and re-synthesized speech is a good indicator to discriminate between genuine speech and adversarial sample. [61, 74] studied a randomized smoothing approach to certifying that no adversarial sample lies in a correct prediction radius without additional retraining. Besides, [54, 80] directly construct a separate detection model to classify genuine and adversarial speeches.

The second theme aims to improve the robustness of deep speaker models against adversarial perturbations. Perhaps data augmentation is the most popular approach by augmenting adversarial speech into model training [24, 34, 77]. Recently, [118, 119] proposed a self-supervised learning approach to purify the adversarial perturbations whilst also maintaining the performance of genuine speech.

3.5 Summary

The above subsections present the basic concepts of various spoofing attacks and their corresponding countermeasures. This subsection will summarize different spoofing attacks in terms of accessibility, effectiveness, and countermeasures, as shown in Table 1. Adapted from [123], *accessibility* reflects the practicality of the attack in real applications, and *effectiveness* reflects the risk of the attack against speaker recognition systems.

- For impersonation, it largely depends on the skill of the imitators, which measures the acoustic similarity between the attacker's voice and the target speaker's. In practice, both accessibility and effectiveness are relatively low.

- For replay attack, an attacker requires no specialized uttering knowledge, and only conducts with a pair of recording and replay devices. Therefore, its accessibility is very high. The replay attack is highly effective in both text-independent and text-dependent applications, while less effective in text-prompted applications.

- For speech synthesis and voice conversion, more and more advanced tools are open to the public. By learning the usage of these tools, attackers can use them to generate high-quality speech. Compared with the knowledge-free replay attack, the accessibility of SS and VC attacks could be considered as mid to high and the effectiveness is high.

- For adversarial attack, it has a great threat to speaker recognition systems (high accessibility), while it also requires lots of specialized knowledge (low to high effectiveness).

Table 1: Summary on spoofing attacks and countermeasures.

Spoofing attacks	Accessibility (practicality)	Effectiveness (risk)	Countermeasures
Impersonation (IMP)	Low	Low	– Acoustic feature analysis
Replay attack (RA)	High	Low to high	– Signal distortion – Liveness detection
Speech synthesis (SS) Voice conversion (VC)	Mid to high	High	– Vocoder attribute – Deep learning
Adversarial attack (AA)	Low to mid	High	– Adversarial detection – Robust against perturbation

4 Countermeasures against Disguise Cheating

With the increase of crimes conducted by voice disguise, a lot of research has explored the impacts of disguised voices in forensic speaker recognition. In this section, we will briefly review countermeasures against two deliberate voice disguise forms: artificial voice disguise and digital voice disguise. The latest systematic survey in this area is reported in [23].

4.1 Countermeasures against Artificial Voice Disguise

Artificial Voice Disguise (AVD) is an active behavior of humans to hide a specific identity. In this case, changes in the vocal track are performed, so that some voice characteristics such as fundamental frequency, accent, prosody, voice quality and so on, are modified. It differs from impersonation in authentication. Impersonation represents an attacker simulating ‘a specific target speaker’ while AVD represents a cheater to conceal ‘his/her real voice’.

Research in [36, 101, 132] revealed the vulnerability of traditional speaker recognition techniques against human disguise of voices. [135] reported that human disguise can largely increase the equal error rate even under more powerful deep speaker models. To address the threat of AVD in speaker recognition, [25] analyzed and compared several acoustic features and found that fundamental frequency is the most detrimental factor. Nevertheless, research on spoofing countermeasures against AVD is still limited.

4.2 Countermeasures against Digital Voice Disguise

With high disguise quality and ease of implementation by abundant tools, digital voice disguise (DVD) has been used in more and more criminal cases, and has presented threats to forensic speaker recognition. As mentioned in Section 2.2.2, DVD is generally divided into Voice Transformation (VT) and Voice Conversion (VC). Compared to efforts on VC disguise as mentioned in Section 3.3, VT disguise (VTD) has received less attention. Here we only discuss spoofing countermeasures against VTD.

Perhaps the most popular VTD technique is pitch scaling [48] and vocal tract length normalization (VTLN) [97]. Pitch scaling can raise or decrease the pitch of voices by frequency shifting or temporal stretching in a linear way, VTLN can be regarded as a non-linear pitch scaling by frequency warping.

Roughly, there are two directions against VTD. The first direction is to defend against VTD by directly detecting whether a voice is disguised or not. [58] adopted MFCC-based GMM models to identify disguised voices by pitch scaling. [112, 120, 121] utilized MFCC and its derivations as input features and construct support vector machines to classify normal speech and pitch-scaling disguised voice. [113] presented a dense convolutional network to detect pitch-scaling disguised voice from genuine speech.

The second direction is to integrate the VTD detection with speaker recognition. Specifically, the disguised voice is firstly restored and then fed into speaker recognition. This restoration of the disguised voice not only can improve the accuracy of automatic speaker recognition, but also is necessary for listening tests as interpretable evidence. [57] applied dynamic time warping algorithm to restore pitch-scaling disguised voices by estimating the degree of disguise, and tested on a vector quantization based speaker recognition system. [114] utilized the ratio of fundamental frequencies to estimate the degree of pitch-scaling disguise for voice restoration, and validated on a GMM-UBM based speaker verification system. [138] presented a more systematic study on both pitch scaling and VTLN based voice disguise and their corresponding restoration, and tested on a more advanced x-vector speaker recognition model. All these works proved the necessity and usefulness of the restoration of disguised voices.

4.3 Summary

Same with Section 3.5, this subsection presents a summary on disguise cheating along with their accessibility, effectiveness and countermeasures, as shown in Table 2.

- For artificial voice disguise, in contrast to impersonation in spoofing attacks, it is a spontaneous behavior of the target speaker. Due to the randomness of within-speaker variations, both accessibility and effectiveness are high.

- For digital voice disguise, a cheater can utilize many prevailing audio editors to alter or modify his/her voice. Similar to speech synthesis and voice conversion in spoofing attacks, its accessibility is mid to high, and the effectiveness is high.

Table 2: Summary on disguise cheating and countermeasures.

Disguise cheating	Accessibility (practicality)	Effectiveness (risk)	Countermeasure
Artificial voice disguise (AVD)	High	High	– Acoustic feature analysis
Digital voice disguise (DVD)	Mid to high	High	– Disguise detection – Restoration and re-recognition

5 Discussion

In this section, we will discuss some topics that may be possible for future research directions. Firstly, we suggest that the research community in this area should not only stare at the technical view but combine it with practical application. Therefore, the application-oriented fake actions and countermeasures in speaker recognition should be paid more attention to. Moreover, we advocate that studies on countermeasures should meet three requirements:

- *Explainability*: In spite of the impressive success of deep neural networks (DNNs) in fake detection, the understanding and explanation of the internal function of these DNNs are still limited, leading to the ‘black-box’ to a large extent. In real applications, these models are unexplainable, especially when they occur an error decision. Therefore, explainability is a primary requirement on countermeasures.

- *Robustness*: In practice, we would like countermeasures to be robust against various kinds of variations, such as codec, transmission and channel variability of speech signal, and also high-quality speech synthesizers and voice converters. Therefore, it is worth studying how to design features, classifiers or systems to be robust against variations.
- *Generalizability*: For a deployed detection system, it not only should detect *seen* fake samples, but also can handle *unseen* fake samples. For instance, unseen devices in recording & replay, emerging methods in speech generation, black-box issues in adversarial attack. Incremental learning or continuous learning seems very important. Besides, the one-class approach is a potential direction, i.e., only models the distribution of genuine speech, and rejects any speech with a low likelihood on the genuine model. Essentially, it formulates fake detection as an out-of-distribution detection problem, rather than a conventional binary classification (genuine & fake) problem.

Besides that, there are two potential research directions:

- *Liveness detection via secondary factor*: The speech signal simulated by a high-quality recording loudspeaker and playback device becomes indistinguishable from the live human voice. In this case, liveness detection via secondary factors may be a vital countermeasure. For instance, a secondary information source (from physical sensors) or modality (such as mouth-speech synchronization) can be utilized to detect fake speech.
- *Partial fake detection*: In practice, fakers can manipulate parts of the evaluation trials or integrate with genuine and fake segments. This will largely increase the detection difficulty. Therefore, a more fine-grained detector should be investigated.

6 Conclusions

This paper presents an application-oriented taxonomy on spoofing, disguise and countermeasures in speaker recognition. Firstly, three popular kinds of applications are introduced: access control, surveillance and forensic. Secondly, we categorize fake actions among different applications into two forms: spoofing attack and disguise cheating. The two fake forms have opposite objects. Simply put, the spoofing attack aims to get as close as possible to the target speaker while the disguise cheating aims to keep as far away as possible from the target speaker. Thirdly, according to the production mode and evaluation subject, fake techniques in each form are further classified into different types. Representative countermeasures against each kind of fake technique are then presented. Finally, we discuss the future research direction in this area. This paper suggests that the research community in this area should pay more attention to connecting research techniques with practical applications.

Biographies

Lantian Li received the B.S. degree from the China University of Mining and Technology, Beijing, China, in 2013, and the Ph.D. degree from the Department of Computer Science, Tsinghua University, Beijing, China, in 2018. Since 2018, he has been a Postdoctoral Fellow with the Center for Speech and Language Technology, Tsinghua University. His research interests include speaker recognition with machine learning methods.

Xingliang Cheng received the B.S. degree in Computer Science and Technology from the Harbin University of Science and Technology, China, in 2016. He is a Ph.D. candidate in the Center for Speech and Language Technologies, Tsinghua University, Beijing, China. His research interests include voice biometrics and voice anti-spoofing.

Thomas Fang Zheng received the B.S. and M.S. degrees in Computer Science and Technology and the Ph.D. degree in Computer Application Technology from Tsinghua University, Beijing, China, in 1990, 1992, and 1997, respectively. He is a Professor, Director of the Center for Speech and Language Technologies, Tsinghua University. His research and development interests include speech recognition, speaker recognition, emotion recognition and natural language processing.

References

[1] M. Adiban, H. Sameti, N. Maghsoodi, and S. Shahsavari, “SUT System Description for Anti-spoofing 2017 Challenge”, in *Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017)*, 2017, 264–75.

[2] F. Alegre, A. Janicki, and N. Evans, “Re-assessing the Threat of Replay Spoofing Attacks against Automatic Speaker Verification”, in *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2014, 1–6.

[3] Z. K. Anjum and R. K. Swamy, “Spoofing and Countermeasures for Speaker Verification: A Review”, in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, IEEE, 2017, 467–71.

[4] A. R. Avila, M. J. Alam, D. D. O'Shaughnessy, and T. H. Falk, "Blind Channel Response Estimation for Replay Attack Detection.", in *Interspeech*, 2019, 2893–7.

[5] *Biometrics Glossary (BG)*, https://www.hsdl.org/?view&did=32101 (accessed on 08/01/2008).

[6] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, *Guide to Biometrics*, Springer Science & Business Media, 2013.

[7] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial Impostor Voice Transformation Effects on False Acceptance Rates", in *Interspeech*, 2007.

[8] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, "VoxSRC 2021: The Third VoxCeleb Speaker Recognition Challenge", *arXiv preprint arXiv:2201.04583*, 2022.

[9] W. Cai, H. Wu, D. Cai, and M. Li, "The DKU Replay Detection System for the ASVspoof 2019 Challenge: On Data Augmentation, Feature Representation, Classification, and Fusion", *arXiv preprint arXiv:1907.02663*, 2019.

[10] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic Speaker Recognition", *IEEE Signal Processing Magazine*, 26(2), 2009, 95–103.

[11] A. Chadha, A. Abdullah, L. Angeline, and S. Sivanesan, "A Review on State-of-the-Art Automatic Speaker Verification System from Spoofing and Anti-spoofing Perspective", *Indian Journal of Science and Technology*, 14(40), 2021, 3026–50.

[12] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A Survey on Adversarial Attacks and Defences", *CAAI Transactions on Intelligence Technology*, 6(1), 2021, 25–45.

[13] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen, "You Can Hear But You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones", in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 2017, 183–95.

[14] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and Model Fusion for Automatic Spoofing Detection", in *Interspeech*, 2017, 102–6.

[15] X. Cheng, M. Xu, and T. F. Zheng, "Cross-Database Replay Detection in Terminal-dependent Speaker Verification", in *Interspeech*, 2021, 4274–8.

[16] X. Cheng, M. Xu, and T. F. Zheng, "Replay Detection using CQT-based Modified Group Delay Feature and ResNeWt Network in ASVspoof 2019", in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2019, 540–5.

[17] B. Chettri, T. Kinnunen, and E. Benetos, "Deep Generative Variational Autoencoding for Replay Spoof Detection in Automatic Speaker Verification", *Computer Speech & Language*, 63, 2020, 101092.

[18] J. S. Chung, A. Nagrani, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "VoxSRC 2019: The First VoxCeleb Speaker Recognition Challenge", *arXiv preprint arXiv:1912.02522*, 2019.

[19] R. K. Das, J. Yang, and H. Li, "Data Augmentation with Signal Companding for Detection of Logical Access Attacks", in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 6349–53.

[20] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of Speaker Verification Security and Detection of HMM-based Synthetic Speech", *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8), 2012, 2280–90.

[21] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the Vulnerability of Speaker Verification to Realistic Voice Spoofing", in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2015, 1–6.

[22] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and Countermeasures for Automatic Speaker Verification", in *Interspeech*, 2013, 925–9.

[23] M. Farrús, "Voice Disguise in Automatic Speaker Recognition", *ACM Computing Surveys (CSUR)*, 51(4), 2018, 1–22.

[24] A. Gomez-Alanis, J. A. Gonzalez-Lopez, and A. M. Peinado, "GANBA: Generative Adversarial Network for Biometric Anti-Spoofing", *Applied Sciences*, 12(3), 2022, 1454.

[25] R. González Hautamäki, V. Hautamäki, and T. Kinnunen, "On the Limits of Automatic Speaker Verification: Explaining Degraded Recognizer Scores through Acoustic Changes Resulting from Voice Disguise", *The Journal of the Acoustical Society of America*, 146(1), 2019, 693–704.

[26] P. Gujarathi and S. R. Patil, "Review on Unit Selection-Based Concatenation Approach in Text to Speech Synthesis System", in *Cybernetics, Cognition and Machine Learning Applications*, Springer, 2021, 191–202.

[27] R. M. Hanifa, K. Isa, and S. Mohamad, "A Review on Speaker Recognition: Technology and Challenges", *Computers & Electrical Engineering*, 90, 2021, 107005.

[28] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors Meet Imitators: On Vulnerability of Speaker Verification Systems against Voice Mimicry", in *Interspeech*, Citeseer, 2013, 930–4.

[29] L. Huang and C.-M. Pun, "Audio Replay Spoof Attack Detection by Joint Segment-based Linear Filter Bank Feature Extraction and Attention-enhanced Densenet-BiLSTM Network", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2020, 1813–25.

[30] L. Huang and C.-M. Pun, “Audio Replay Spoof Attack Detection using Segment-based Hybrid Feature and DenseNet-LSTM Network”, in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 2567–71.

[31] A. J. Hunt and A. W. Black, “Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database”, in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 1, IEEE, 1996, 373–6.

[32] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*, Springer Science & Business Media, 2007.

[33] A. K. Jain, A. Ross, and S. Pankanti, “Biometrics: A Tool for Information Security”, *IEEE Transactions on Information Forensics and Security*, 1(2), 2006, 125–43.

[34] A. Jati, C.-C. Hsu, M. Pal, R. Peri, W. AbdAlmageed, and S. Narayanan, “Adversarial Attack and Defense Strategies for Deep Speaker Recognition Systems”, *Computer Speech & Language*, 68, 2021, 101199.

[35] S. Joshi, J. Villalba, P. Żelasko, L. Moro-Velázquez, and N. Dehak, “Study of Pre-processing Defenses against Adversarial Attacks on State-of-the-art Speaker Recognition Systems”, *IEEE Transactions on Information Forensics and Security*, 16, 2021, 4811–26.

[36] S. S. Kajarekar, H. Bratt, E. Shriberg, and R. De Leon, “A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition”, in *The Speaker and Language Recognition Workshop (Odyssey)*, IEEE, 2006, 1–6.

[37] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient Neural Audio Synthesis”, in *International Conference on Machine Learning*, PMLR, 2018, 2410–9.

[38] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, “Advances in Anti-spoofing: From the Perspective of ASVspoof Challenges”, *APSIPA Transactions on Signal and Information Processing*, 9, 2020.

[39] T. Kaneko and H. Kameoka, “CycleGAN-VC: Non-Parallel Voice Conversion using Cycle-consistent Adversarial Networks”, in *2018 26th European Signal Processing Conference (EUSIPCO)*, IEEE, 2018, 2100–4.

[40] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion”, in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 6820–4.

[41] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion”, *arXiv preprint arXiv:2010.11672*, 2020.

[42] C. Kim and R. M. Stern, “Power-normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7), 2016, 1315–29.

[43] T. Kinnunen and H. Li, “An Overview of Text-independent Speaker Recognition: From Features to Supervectors”, *Speech Communication*, 52(1), 2010, 12–40.

[44] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection”, in *Interspeech*, ISCA, 2017, 2–6.

[45] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, “Vulnerability of Speaker Verification Systems against Voice Conversion Spoofing Attacks: The Case of Telephone Speech”, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2012, 4401–4.

[46] J. Kong, J. Kim, and J. Bae, “HiFi-gan: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis”, *Advances in Neural Information Processing Systems*, 33, 2020, 17022–33.

[47] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, “Fooling End-to-End Speaker Verification with Adversarial Examples”, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 1962–6.

[48] J. Laroche, “Time and Pitch Scale Modification of Audio Signals”, in *Applications of digital signal processing to audio and acoustics*, Springer, 2002, 279–309.

[49] Y. W. Lau, D. Tran, and M. Wagner, “Testing Voice Mimicry with the YOHO Speaker Verification Corpus”, in *International Conference on Knowledge-based and Intelligent Information and Engineering Systems*, Springer, 2005, 15–21.

[50] Y. W. Lau, M. Wagner, and D. Tran, “Vulnerability of Speaker Verification to Voice Mimicking”, in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, IEEE, 2004, 145–8.

[51] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, “STC Antispoofing Systems for the ASVspoof2019 Challenge”, *arXiv preprint arXiv:1904.05576*, 2019.

[52] J. Li, X. Zhang, J. Xu, S. Ma, and W. Gao, “Learning to Fool the Speaker Recognition”, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3s), 2021, 1–21.

[53] L. Li, Y. Chen, D. Wang, and T. F. Zheng, “A Study on Replay Attack and Anti-spoofing for Automatic Speaker Verification”, *arXiv preprint arXiv:1706.02101*, 2017.

[54] X. Li, N. Li, J. Zhong, X. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Investigating Robustness of Adversarial Samples Detection for Automatic Speaker Verification", *arXiv preprint arXiv:2006.06186*, 2020.

[55] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise Gated Res2Net: Towards Robust Detection of Synthetic Speech Attacks", *arXiv preprint arXiv:2107.08803*, 2021.

[56] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial Attacks on GMM i-vector Based Speaker Verification Systems", in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, 6579–83.

[57] Y.-p. Li, D.-y. Tao, and L. Lin, "Study on Electronic Disguised Voice Speaker Recognition Based on DTW Model Compensation", *Comput. Technol. Develop.*, 27(1), 2017, 93–6.

[58] Y. Li, L. Lin, and D. Tao, "Research on Identification of Electronic Disguised Voice Based on GMM Statistical Parameters", *Computer Technology and Development*, 27(1), 2017, 103–6.

[59] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical Adversarial Attacks Against Speaker Recognition Systems", in *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, 2020, 9–14.

[60] J. Lindberg and M. Blomberg, "Vulnerability in Speaker Verification-A Study of Technical Impostor Techniques", in *Sixth European Conference on Speech Communication and Technology*, 1999.

[61] T. Maho, T. Furon, and E. Le Merrer, "Randomized Smoothing under Attack: How Good is it in Pratice?", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[62] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-spoofing: Presentation Attack Detection*, Vol. 2, Springer, 2019.

[63] J. Mariéthoz and S. Bengio, "Can a Professional Imitator Fool a GMM-based Speaker Verification System?", *tech. rep.*, IDIAP, 2005.

[64] D. Markham, *Phonetic Imitation, Accent, and the Learner*, Vol. 33, Lund University, 1997.

[65] J. A. Markowitz, "Speaker Identification and Verification (SIV) Applications and Markets", in *Workshop on Speaker Biometrics and VoiceXML*, Vol. 3, 2008.

[66] H. Masthoff, "A Report on a Voice Disguise Experiment", *International Journal of Speech Language and the Law*, 3(1), 1996, 160–7.

[67] A. Mittal and M. Dua, "Automatic Speaker Verification Systems and Spoof Detection Techniques: Review and Analysis", *International Journal of Speech Technology*, 2021, 1–30.

[68] S. H. Mohammadi and A. Kain, "An Overview of Voice Conversion Systems", *Speech Communication*, 88, 2017, 65–82.

[69] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "VoxSRC 2020: The Second VoxCeleb Speaker Recognition Challenge", *arXiv preprint arXiv: 2012.06867*, 2020.

[70] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2), 2021, 252–65.

[71] C. A. H. Nava, P. L. Velázquez, E. A. R. García, S. G. De los Cobos Silva, M. Á. G. Andrade, and R. A. M. Gutiérrez, "Speech Spoofing Detection using Neural Networks", in *XXIII International Symposium of Mathematical Methods Applied to Sciences*, 2021.

[72] M. Neelima and I. Santiprabha, "Mimicry Voice Detection using Convolutional Neural Networks", in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, IEEE, 2020, 314–8.

[73] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A Review of Deep Learning based Speech Synthesis", *Applied Sciences*, 9(19), 2019, 4050.

[74] R. Olivier, B. Raj, and M. Shah, "High-Frequency Adversarial Defense for Speech and Audio", in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 2995–9.

[75] Z. Oo, L. Wang, K. Phapatanaburi, M. Liu, S. Nakagawa, M. Iwahashi, and J. Dang, "Replay Attack Detection with Auditory Filter-based Relative Phase Features", *EURASIP Journal on Audio, Speech, and Music Processing*, 2019(1), 2019, 1–11.

[76] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A Generative Model for Raw Audio", *arXiv preprint arXiv:1609.03499*, 2016.

[77] M. Pal, A. Jati, R. Peri, C.-C. Hsu, W. AbdAlmageed, and S. Narayanan, "Adversarial Defense for Deep Speaker Recognition using Hybrid Adversarial Training", in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 6164–8.

[78] M. Pal, A. Raikar, A. Panda, and S. K. Kopparapu, "Synthetic Speech Detection using Meta-learning with Prototypical Loss", *arXiv preprint arXiv:2201.09470*, 2022.

[79] H. A. Patil and M. R. Kamble, "A Survey on Replay Attack Detection for Automatic Speaker Verification (ASV) System", in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2018, 1047–53.

[80] Z. Peng, X. Li, and T. Lee, “Pairing Weak with Strong: Twin Models for Defending Against Adversarial Attack on Speaker Verification”, in *Interspeech*, 2021.

[81] P. Perrot, G. Aversano, and G. Chollet, “Voice Disguise and Automatic Detection: Review and Perspectives”, *Progress in Nonlinear Speech Processing*, 2007, 101–17.

[82] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A Flow-based Generative Network for Speech Synthesis”, in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 3617–21.

[83] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “Autovc: Zero-shot Voice Style Transfer with Only Autoencoder Loss”, in *International Conference on Machine Learning (ICML)*, PMLR, 2019, 5210–9.

[84] Y. Ren, W. Liu, D. Liu, and L. Wang, “Recalibrated Bandpass Filtering On Temporal Waveform For Audio Spoof Detection”, in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, 3907–11.

[85] R. Rodman and M. Powell, “Computer Recognition of Speakers who Disguise their Voice”, in *The International Conference on Signal Processing Applications and Technology (ICSPAT)*, Citeseer, 2000.

[86] P. Rose, *Forensic Speaker Identification*, cRc Press, 2002.

[87] P. Rose, “Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence”, *Computer Speech & Language*, 20(2-3), 2006, 159–91.

[88] T. Sabhanayagam, V. P. Venkatesan, and K. Senthamaraikannan, “A Comprehensive Survey on Various Biometric Systems”, *International Journal of Applied Engineering Research*, 13(5), 2018, 2276–97.

[89] O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, “The 2018 NIST Speaker Recognition Evaluation”, in *Interspeech*, 2019.

[90] S. O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, J. Hernandez-Cordero, *et al.*, “The 2019 NIST Speaker Recognition Evaluation CTS Challenge”, in *The Speaker and Language Recognition Workshop (Odyssey)*, Vol. 2020, 2020, 266–72.

[91] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, “Introduction to Voice Presentation Attack Detection and Recent Advances”, in *Handbook of biometric anti-spoofing*, Springer, 2019, 321–61.

[92] M. Sahidullah, D. A. L. Thomsen, R. G. Hautamäki, T. Kinnunen, Z.-H. Tan, R. Parts, and M. Pitkänen, “Robust Voice Liveness Detection and Speaker Verification using Throat Microphones”, *IEEE/ACM*

Transactions on Audio, Speech, and Language Processing, 26(1), 2017, 44–56.

[93] A. S. Shamsabadi, F. S. Teixeira, A. Abad, B. Raj, A. Cavallaro, and I. Trancoso, "FoolHD: Fooling Speaker Identification by Highly Imperceptible Adversarial Disturbances", in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 6159–63.

[94] M. Singh and D. Pati, "Countermeasures to Replay Attacks: A Review", *IETE Technical Review*, 37(6), 2020, 599–614.

[95] B. Sisman, J. Yamagishi, S. King, and H. Li, "An Overview of Voice Conversion and its Challenges: From Statistical Modeling to Deep Learning", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2020, 132–57.

[96] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN Embeddings for Speaker Recognition", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 5329–33.

[97] D. Sundermann and H. Ney, "VTLN-based Voice Conversion", in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795)*, IEEE, 2003, 556–9.

[98] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-End Spectro-temporal Graph Attention Networks for Speaker Verification Anti-spoofing and Speech Deepfake Detection", *arXiv preprint arXiv:2107.12710*, 2021.

[99] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "RawBoost: A Raw Data Boosting and Augmentation Method applied to Automatic Speaker Verification Anti-Spoofing", *arXiv preprint arXiv:2111.04433*, 2021.

[100] C. B. Tan, M. H. A. Hijazi, N. Khamis, Z. Zainol, F. Coenen, A. Gani, *et al.*, "A Survey on Presentation Attack Detection for Automatic Speaker Verification Systems: State-of-the-Art, Taxonomy, Issues and Future Direction", *Multimedia Tools and Applications*, 80(21), 2021, 32725–62.

[101] T. Tan, "The Effect of Voice Disguise on Automatic Speaker Recognition", in *2010 3rd International Congress on Image and Signal Processing (CISP)*, Vol. 8, IEEE, 2010, 3538–41.

[102] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A Survey on Neural Speech Synthesis", *arXiv preprint arXiv:2106.15561*, 2021.

[103] M. Todisco, H. Delgado, and N. Evans, "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification", *Computer Speech & Language*, 45, 2017, 516–35.

[104] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection", *arXiv preprint arXiv:1904.05441*, 2019.

[105] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis", in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 3, IEEE, 2000, 1315–8.

[106] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based Speech Synthesis System Applied to English", in *IEEE Workshop on Speech Synthesis*, IEEE Santa Monica, 2002, 227–30.

[107] *Twins Fool HSBC Voice Biometrics - BBC*, https://www.finextra.com/newsarticle/30594/twins-fool-hsbc-voice-biometrics--bbc (accessed on 05/19/2017).

[108] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep Neural Networks for Small Footprint Text-dependent Speaker Verification", in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, 4052–6.

[109] X. Wang and J. Yamagishi, "A Practical Guide to Logical Access Voice Presentation Attack Detection", *arXiv preprint arXiv:2201.03321*, 2022.

[110] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder based Replay Channel Response Estimation", *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, 16–21.

[111] Y. Wang, W. Cai, T. Gu, W. Shao, Y. Li, and Y. Yu, "Secure Your Voice: An Oral Airflow-based Continuous Liveness Detection for Voice Assistants", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4), 2019, 1–28.

[112] Y. Wang, Y. Deng, H. Wu, and J. Huang, "Blind Detection of Electronic Voice Transformation with Natural Disguise", in *International Workshop on Digital Watermarking*, Springer, 2012, 336–43.

[113] Y. Wang and Z. Su, "Detection of Voice Transformation Spoofing based on Dense Convolutional Network", in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 2587–91.

[114] Y. Wang, H. Wu, and J. Huang, "Verification of Hidden Speaker Behind Transformation Disguised Voices", *Digital Signal Processing*, 45, 2015, 84–95.

[115] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, *et al.*, "Tacotron: Towards End-to-End Speech Synthesis", *arXiv preprint arXiv:1703.10135*, 2017.

[116] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio Replay Attack Detection Using High-Frequency Features", in *Interspeech*, 2017, 27–31.

[117] H. Wu, P.-c. Hsu, J. Gao, S. Zhang, S. Huang, J. Kang, Z. Wu, H. Meng, and H.-y. Lee, "Spotting Adversarial Samples for Speaker Verification by Neural Vocoders", *arXiv preprint arXiv:2107.00309*, 2021.

[118] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-Y. Lee, "Improving the Adversarial Robustness for Speaker Verification by Self-supervised Learning", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2021, 202–17.

[119] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-y. Lee, "Adversarial Defense for Automatic Speaker Verification by Cascaded Self-supervised Learning Models", in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 6718–22.

[120] H. Wu, Y. Wang, and J. Huang, "Blind Detection of Electronic Disguised Voice", in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013, 3013–7.

[121] H. Wu, Y. Wang, and J. Huang, "Identification of Electronic Disguised Voices", *IEEE Transactions on Information Forensics and Security*, 9(3), 2014, 489–500.

[122] Z. Wu, E. S. Chng, and H. Li, "Detecting Converted Speech and Natural Speech for Anti-spoofing Attack in Speaker Recognition", in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[123] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and Countermeasures for Speaker Verification: A Survey", *Speech Communication*, 66, 2015, 130–53.

[124] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge", in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[125] Z. Wu and H. Li, "Voice Conversion versus Speaker Verification: An Overview", *APSIPA Transactions on Signal and Information Processing*, 3, 2014.

[126] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, Universal, and Robust Adversarial Attacks against Speaker Recognition Systems", in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, 1738–42.

[127] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. A. Lee, V. Vestman, and A. Nautsch, *ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures*

Challenge Evaluation Plan, URL: https://www.asvspoof.org/asvspoof2019_evaluation_plan.pdf, 2019.

[128] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, *et al.*, "ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection", *arXiv preprint arXiv:2109.00537*, 2021.

[129] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning", *IEEE transactions on neural networks and learning systems*, 30(9), 2019, 2805–24.

[130] H. Zeinali, T. Stafylakis, G. Athanasopoulou, J. Rohdin, I. Gkinis, L. Burget, J. Černocky, *et al.*, "Detecting Spoofing Attacks using VGG and SincNet: BUT-Omilia Submission to ASVspoof 2019 Challenge", *arXiv preprint arXiv:1907.12908*, 2019.

[131] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis", *Speech Communication*, 51(11), 2009, 1039–64.

[132] C. Zhang and T. Tan, "Voice Disguise and Automatic Speaker Recognition", *Forensic Science International*, 175(2-3), 2008, 118–22.

[133] L. Zhang, S. Tan, Z. Wang, Y. Ren, Z. Wang, and J. Yang, "VibLive: A Continuous Liveness Detection for Secure Voice User Interface in Iot Environment", in *Annual Computer Security Applications Conference*, 2020, 884–96.

[134] L. Zhang, S. Tan, and J. Yang, "Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication", in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, 57–71.

[135] M. Zhang, X. Kang, Y. Wang, L. Li, Z. Tang, H. Dai, and D. Wang, "Human and Machine Speaker Recognition based on Short Trivial Events", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 5009–13.

[136] W. Zhang, S. Zhao, L. Liu, J. Li, X. Cheng, T. F. Zheng, and X. Hu, "Attack on Practical Speaker Verification System using Universal Adversarial Perturbations", in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 2575–9.

[137] Z. Zhang, X. Yi, and X. Zhao, "Fake Speech Detection Using Residual Network with Transformer Encoder", in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, 2021, 13–22.

[138] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, "When Automatic Voice Disguise Meets Automatic Speaker Verification", *IEEE Transactions on Information Forensics and Security*, 16, 2020, 824–37.

[139] T. F. Zheng and L. Li, *Robustness-related Issues in Speaker Recognition*, Springer, 2017.