Gupta *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:10
https://doi.org/10.1186/s13636-024-00328-8

EURASIP Journal on Audio, Speech, and Music Processing

REVIEW

Open Access

Vulnerability issues in Automatic Speaker Verification (ASV) systems

Priyanka Gupta[1,3], Hemant A. Patil[1] and Rodrigo Capobianco Guido[2*]

Abstract

Claimed identities of speakers can be verified by means of automatic speaker verification (ASV) systems, also known as voice biometric systems. Focusing on security and robustness against spoofing attacks on ASV systems, and observing that the investigation of attacker's perspectives is capable of leading the way to prevent known and unknown threats to ASV systems, several countermeasures (CMs) have been proposed during ASVspoof 2015, 2017, 2019, and 2021 challenge campaigns that were organized during INTERSPEECH conferences. Furthermore, there is a recent initiative to organize the ASVSpoof 5 challenge with the objective of collecting the massive spoofing/deepfake attack data (i.e., phase 1), and the design of a spoofing-aware ASV system using a single classifier for both ASV and CM, to design integrated CM-ASV solutions (phase 2). To that effect, this paper presents a survey on a diversity of possible strategies and vulnerabilities explored to successfully attack an ASV system, such as target selection, unavailability of global countermeasures to reduce the attacker's chance to explore the weaknesses, state-of-the-art adversarial attacks based on machine learning, and deepfake generation. This paper also covers the *possibility* of attacks, such as hardware attacks on ASV systems. Finally, we also discuss the several technological challenges from the attacker's perspective, which can be exploited to come up with better defence mechanisms for the security of ASV systems.

Keywords Automatic speaker verification, Spoofing attacks, Attacker's perspective, Adversarial attacks, Deepfake

1 Introduction

Automatic speaker verification (ASV) systems are voice-based biometric systems used to authenticate speakers' claimed identities. They are vulnerable to various spoofing attacks, such as identical twins, impersonation, voice conversion (VC), synthetic speech (SS), and replay [1]. In order to design robust defending mechanisms, it is important to discuss the numerous techniques, that can enable spoofing attacks on ASV systems. Assessments on the security of ASV systems can be performed whenever various possible approaches and attackers' perspectives are known a priori. Hence, possible vulnerability aspects should be examined in order to make an ASV system robust against spoofing attacks.

In ASVspoof 2015 challenge, during INTERSPEECH 2015, several countermeasures (CMs) were proposed using a diversity of feature extraction techniques. They are mostly based on signal processing strategies over the standard and statistically meaningful ASVSpoof 2015 dataset [2]. In particular, most of the participant teams concentrated on signal processing-based research strategies to develop feature sets and, then, used Gaussian mixture models (GMMs) for a two-class classification problem of distinguishing spoofed from genuine speech. Furthermore, for ASVSpoof 2017 challenge during INTERSPEECH 2017, several CMs for replay

*Correspondence:
Rodrigo Capobianco Guido
guido@ieee.org
[1] Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DAIICT), Gandhinagar, India
[2] Instituto de Biociências, Letras e Ciências Exatas, Unesp - Univ Estadual Paulista (São Paulo State University), Rua Cristóvão Colombo 2265, Jd Nazareth, 15054-000 São José do Rio Preto - SP, Brazil
[3] Department of Communication and Computer Engineering, The LNM Institute of Information Technology, Jaipur, India

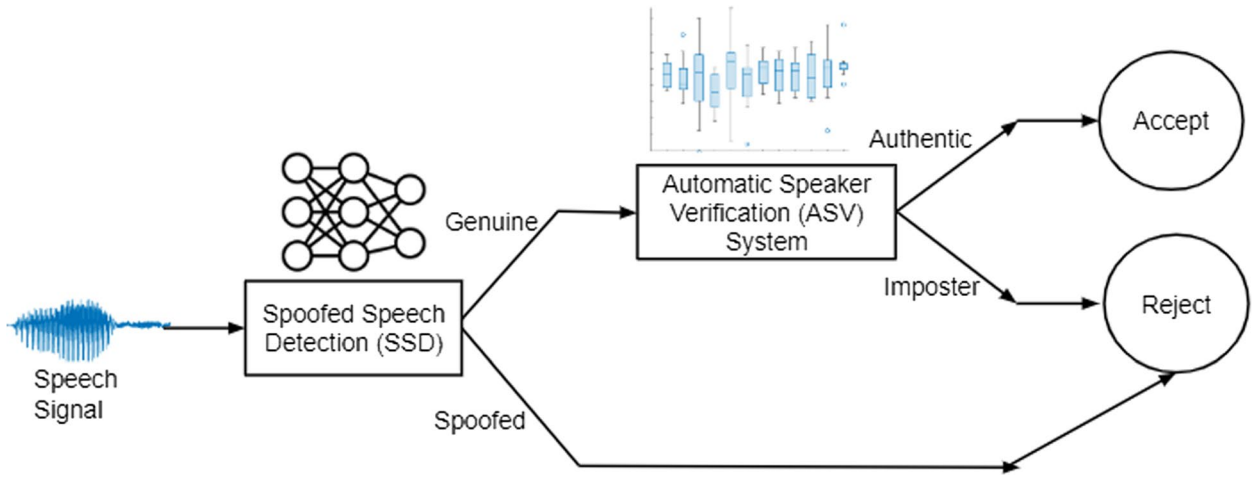


Fig. 1 SSD and ASV systems in *tandem*. After [9]

spoof detection were presented [3, 4], including the use of deep learning-based methods. Recently, ASVSpoof 2021 challenge——a satellite event of INTERSPEECH 2021——focused additionally on *deepfake* detection [5]. In addition to this in 2023, the ASVSpoof 5 challenge was organized, focusing on two main tasks——(1) data collection of spoofing/deepfake attack data, and (2) design of integrated CM-ASV system for deepfakes. The purpose of the first task is to have more real-world data as the deepfake attacks are expected to be more adversarial than in previous editions of ASVspoof challenges and to fool both ASV and CM systems. The purpose of the second task is to have an integrated solution in the form of a spoofing-aware speaker verification (SASV) system, which is an integrated CM-ASV system [6]. Furthermore, very recently, two editions of audio deepfake detection (ADD) challenges, namely, ADD 2022 [7] and ADD 2023 [8], were organized, indicating the vibrant and synergistic activities in this field.

In order to detect spoofed speech, a spoof speech detection (SSD) system is considered in *tandem* with ASV system, making it a two-class problem, as shown in Fig. 1, where the SSD system identifies an attack and, subsequently, denies the corresponding spoofed speech to enter into the ASV system. Nevertheless, due to the advancement of deep neural network (DNN) architectures, ASV systems remain vulnerable to powerful attacks, such as voice conversion and *deepfake* generation based on adversarial training and generative adversarial networks (GANs) [10]. Consequently, the security of ASV systems can be compromised by using these approaches as individual methods of attack or a possible combination of these in the near future. Therefore, this study explores the various vulnerabilities and attacking approaches to an ASV system. It should be noted that the study in [11] reports the attacker's perspective mainly on non-proactive attacks, such as VC and SS, and proactive attacks, which are mainly adversarial attacks. Unlike [11], the discussion in this work is not limited to well-known adversarial attacks and spoofing attacks, such as replay, VC, and SS only. Contrary to this, our study investigates attacking strategies and vulnerabilities, such as target selection, deepfake generation, enrolled users with malicious intent, and complementary attacks, in addition to the technical challenges faced by an attacker, while mounting an attack. The approach of target selection can be used by an attacker to select the most vulnerable speaker to imitate, in order to be authorized by an ASV system. It is based on the hypothesis that a pool of speakers contains speakers with varying vulnerability levels and varying effects on the performance of the ASV system [12]. The approach of deepfake generation is in line with the current trends, especially with fast-paced research in *generative AI* [13]. Another attacking approach discussed in this work is the scenario when there are enrolled users with malicious intent [14, 15]. This attacking scenario/perspective is important to note because in such a case the attacker is not an outside entity, whereas usually most of the attacking strategies and prevention techniques assume the attacker to be an external entity, with no knowledge or access to the ASV system. Furthermore, we also discuss the effect and the role of publicly available corpora for anti-spoofing, as well as publicly available audio content on the Internet through websites, such as YouTube. In complement, this paper also gently discusses *possible* hardware attacks. Furthermore, this study also presents

the experimental findings and observations w.r.t. various attacking techniques in the literature.

The remainder of this paper is organized as follows. Classification of attacks is presented in Section 2. Apart from the most known attacks discussed in that section, various other vulnerabilities of ASV systems are described in Section 3. Section 4 presents the various technological challenges faced by the attacker, while mounting a successful attack. Finally, we conclude our paper in Section 5 along with potential future research directions.

2 Classification of attacks on an ASV system

Notably, there are two main types of attacks, namely, direct and indirect, as shown in Fig. 2. Direct attacks are those implemented and carried out without understanding the internal architecture of the ASV system design. As a result, in a direct attack, the attacker does not breach or fool any internal subsystem in the target ASV system. Instead, attacks on the microphone and transmission levels are carried out. To that effect, a successful direct attack does not need any prior knowledge of the ASV system in question. This is the reason why such an event is also known as *black box attack* [16]. Thus, this kind of attack poses a significant threat to the security of the ASV system due to its ease of execution. Types of direct attacks are spoofing attacks, hardware attacks, and adversarial attacks, as shown in Fig. 2.

Contrary to this, indirect attacks are those occurring in system-levels, being feasible whenever the attacker has access to the internal subsystems of the target ASV system. If the attacker has complete knowledge and access to all the subsystems, the attack is termed as a *white box* attack. It represents an ideal scenario for attackers, which is not practically realistic. However, despite their unrealistic nature, these attacks should not be ignored since they represent the worst-case possibility for the security of ASV systems. The robustness of an ASV system should be evaluated against such a worst-case scenario so that the ASV systems, and their associated countermeasures, are fully prepared to prevent most of the possible attacks.

A more realistic case of indirect attacks is that in which the attacker has partial knowledge of the target ASV system. Such indirect attacks are termed as *grey box attacks.* Most of the indirect attacks are grey box attacks due to their realistic nature. An attacker can perform more serious damage to the ASV system security by implementing a grey box attack as compared to a black box attack because more power, i.e., knowledge on the grey-box target ASV system exists. We now briefly comment on each of the attacks shown in Fig. 2, i.e., spoofing attacks, hardware attacks, and attacks on corpora. Specifically, adversarial attacks are discussed in much greater detail in the next section.

2.1 Spoofing attacks

Spoofing attacks fall under the category of direct attacks and are the most discussed attacks in the literature. Spoofing attacks generated from text-to-speech (TTS)

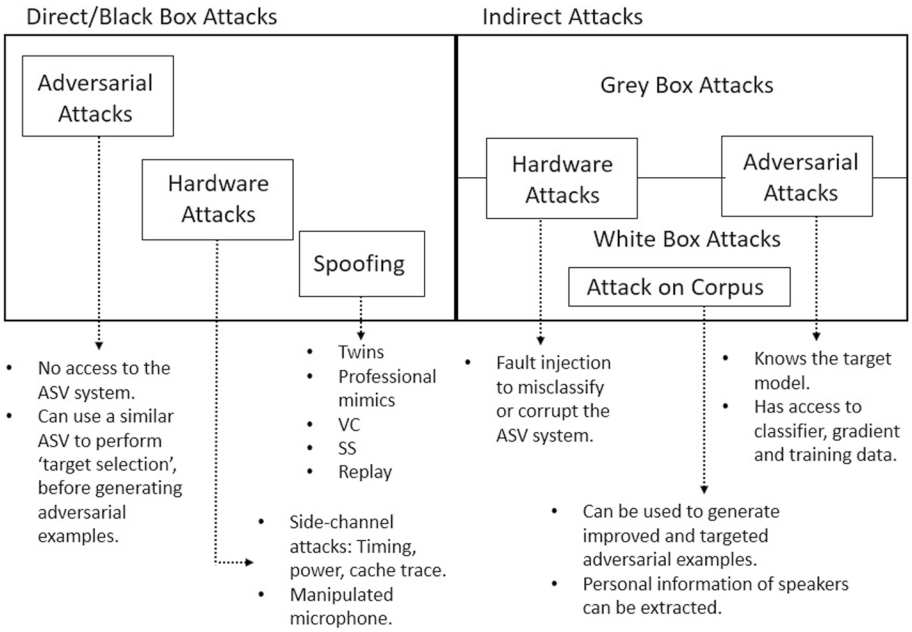


Fig. 2 Classifying various attacks on an ASV system. After [16]

and voice conversion (VC) techniques are called *logical access (LA)* attacks. In opposition, spoofing attacks, which are generated in real physical space are called *physical access (PA)* attacks. The most common type of PA attack is a replay attack. Furthermore, a recent type of attack, known as *deepfake* is also a direct attack, which involves generating spoofing utterances using TTS and VC algorithms, similar to LA. Currently, *deepfake* attacks are known to be the most successful types of attacks. However, the ease of mounting and executing an attack also plays a role from an attacker's perspective. To that effect, replay attacks are the easiest to mount, most difficult to detect, and do not even require the attacker to be technically knowledgeable!

2.2 Hardware attacks

Due to flaws in hardware implementations of security algorithms, an attacker can find the possibility of mounting a hardware attack. These attacks can be direct as well as indirect. In case of a direct hardware attack, the attacker can keep track of outputs from the hardware, such as power, timing, and cache traits, to get enough information about the ASV system in order to attack it. Such attacks are called *side-channel attacks*. Simple power analysis (SPA) and differential power analysis (DPA), for instance, are classic examples of such type of attacks [17, 18]. Differently, in the case of grey-box and white box attacks, where the attacker has partial or complete access to the victim hardware, the hardware attacks are performed by deliberately mounting faults in the electrical circuitry to alter the behavior of the circuit used. An example of fault injection attack is performed by injecting parametric *Trojan* [19]. With the help of parametric Trojan, the electrical characteristics of the logic gates used in the circuit is altered. However, hardware attacks are usually mounted on systems, which use cryptographic algorithms for their security. In this regard, to the best of the authors' knowledge and belief, a hardware attack on an ASV system is yet to be uncovered, and hence, it is an open research problem!

2.3 Attack on corpora

Attacks over *unprotected* corpora are categorized as white box attacks. Attacks over unprotected corpora do not necessarily lead to attack on an ASV system, however, can be used to determine personal information about speakers. The ISO/IEC International Standard 24745 on Biometric Information Protection [20] enforces that, for full privacy protection, biometric references should be irreversible and *unlinkable* [21–23]. An unprotected speech corpus, i.e., a biometric reference, enables searching for a speaker's information on the Internet [24, 25]. Likewise, the study in [26] deals with matching users' speech to celebrities' speech data available on YouTube. Thus, due to publicly available speaker data collected from YouTube, also called as *"found data"*, an attacker can look for a celebrity's voice, which resembles the most to a particular user's voice, using an approach called as *target selection*, as described ahead.

3 Vulnerabilities of ASV systems: approaches and techniques of attacks

In this section, we present various attacking approaches on the ASV systems, mainly including target selection-based and adversarial attacks. Furthermore, a detailed analysis of various attacks found in the literature is coherently shown in Table 1. We also dedicate space for a discussion on vulnerability scenarios, such as malicious enrolled users, and the lack of robust universal countermeasures leading to the attacker benefiting from the weakness of the SSD system.

3.1 Target selection attacks

To intuitively understand the attacker's approach of *target selection*, we assume log-likelihood ratio (LLR) as being the similarity score. Usually, it is compared to a predefined threshold, which then defines the false acceptance rate (FAR) and false rejection rate (FRR). Additionally, the LLR is computed in terms of probabilistic linear discriminant analysis (PLDA) score for state-of-the-art x-vector-based approach for ASV [44]. Target selection attacks can be performed in one of the following two ways:

1. By selecting the most vulnerable speaker, from the speaker classification step as shown in Table 2, referred to as *"lamb"* in [12], from the set of enrolled speakers. Lambs are the speakers who are easiest to mimic w.r.t. a specific attacker. Thus, the speaker with the highest LLR score w.r.t. that attacker is selected as being the lamb.
2. By selecting the most skillful attacker, referred to as *"wolf"* in [12], w.r.t. a pre-defined victim speaker. Thus, an attacker with the highest LLR score w.r.t. the fixed pre-defined victim is selected as being the wolf.

In order to increase the chances of a successful attack, an attacker selects the most vulnerable target by using the attacker's own ASV, as shown in Fig. 3, consequently increasing the FAR [12]. Then, the attacker succeeds with good and appropriate target selection. It is worth mentioning that, while such an attack may not always show an increase in FAR, this approach can still

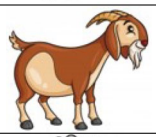
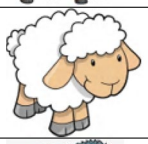
Table 1 Selected attacking techniques in the ASV literature

Basis of attack	Corpus used	Observations
Choosing the closest target based on FAR using GMM [27]	YOHO	• If the number of sessions are more in which the attacker has listened to the target voice, higher verification error rate is obtained • The highest FAR achieved was 35% by imitator2
Choosing the closest target using attacker's ASV on the basis of EER [28]	VoxCeleb1 and VoxCeleb2	• Transferability is observed from the attacker's ASV to the attacked ASV in the order of the closest, median, and the farthest speakers • Contrary to the intuition, if the target speaker's voice is already similar to the impersonator's voice, the verification error score is lowered! However, in case of the targets that are not close to the attacker, impersonation increases the verification error, thereby improving the attack
Training feedback controlled voice conversion system, with feedback coming from the black box target ASV. The VC method used is phonetic posteriorgram (PPG)-based [29]	Subset of ASVSpoof 2019 LA dataset	• Higher EER indicates better attack. Overall EER achieved using PPG-VC with feedback attack method was 30.73%, whereas without feedback it was 29.25% • Male speakers were observed to be more vulnerable due to PPG-based VC attack with EER of 32.90% and 31.60% for the cases of with and without feedback, respectively • Female speakers on the other hand were comparatively less vulnerable due to reduced EERs of 26.67% and 25%, for the cases of with and without feedback, respectively
Crafting adversarial examples at the acoustic feature-level, i.e., MFCC and log power magnitude spectrum (LPMS). To generate perturbation, fast gradient sign method (FGSM) is used to solve the optimization problem [30]	VoxCeleb1	• In black box setting, for perturbation $\epsilon = 20$, EER of 74.62% was achieved • In white box setting, LPMS i-vector-based system was found to be more vulnerable than MFCC *i*-vector. For $\epsilon = 1$, FAR and EER obtained by LPMS *i*-vector were 99.99% and 99.95%, respectively
Crafting adversarial examples using FGSM and local distribution smoothness (LDS) method [31]	TIMIT	• EER is improved by (i) +18.89% and (ii) +5.54% for *the original test set* using the regularized model • Furthermore, EER is improved on the *adversarial example test set* by (i) +30.11%, and (ii) +22.12%
Developing an audio-agnostic universal generating sound distortions by estimating perturbation. Robustness is improved by the room impulse response (RIR) [32]	Multi-speaker corpus from Voice Cloning Toolkit (VCTK)	• 90% attack success rate is achieved on both x-vector and d-vector-based ASVs • Attack time is sped up by 100 times. Both were achieved on white box scenarios
Crafting adversarial examples using biometrics transformation network configuration (ABTN), which jointly processes the loss best of the PAD and ASV systems to generate black box and white box adversarial examples [33]	ASVSpoof 2019	• ABTN outperforms adversarial attacks, obtaining 10.28% and 10.14% higher EER joint w. r. t. the PGD ($\epsilon = 1.0$) in the LA and PA test sets, respectively.
Voice conversion using weighted frequency warping (WFW) [34]	TIMIT and CMU-ARCTIC	• The WFW-based attack failed on speaker identification systems as the source voice and its corresponding speaker could be identified in numerous cases
Text-to-speech (TTS) system, which contains a speaker encoder network, a sequence-to-sequence synthesis network, and an auto-regressive WaveNet-based vocoder network, which converts the Mel spectrogram into time-domain signal [35]	VCTK and LibriSpeech	• It is demonstrated that synthesized speech is reasonable natural sounding speech, similar to real even on unseen speakers

Table 1 (continued)

Basis of attack	Corpus used	Observations
		• Human-level naturalness is not achieved despite the use of a WaveNet vocoder
An autoencoder-based voice conversion system [36]	VCTK	• Generalizes well to unseen speakers
		• Speaker characteristics are disentangled from the linguistic content by the encoder bottleneck
		• Like [35], it also uses WaveNet vocoder
SV2TTS [27]	Customized Data	Azure, and WeChat can accept at least 1 synthesized attack utterance
ASV is trained under unconstrained recording and speaking conditions [37]	Collected Impersonation Dataset (CID)	• Attacks using deepfake speech are more likely to be successful than the other attacking techniques, including speech synthesis and impersonation by professionals
		• It is established that the fine structures in the speech present due to the human speech production mechanism can capture the discriminating acoustic cues between natural and machine-generated speech, such as deepfake speech
DolphinAttack: inaudible voice commands on ultrasonic carriers [38]	–	• Even though inaudible, DolphinAttack voice commands can successfully activate the audio hardware of devices, such as Siri, Alexa, and GoogleNow
		• The attack leads to various vulnerabilities, such as visiting a malicious website, spying, injection of fake information, and denial of service
Targeted adversarial attack called as *FAKEBOB* under black box setting [50]	LibriSpeech and VoxCeleb	• Success rate of 99% is achieved on both open source and commercial systems
		• It is concluded that it is difficult to differentiate the speakers of the original voices from those of the generated adversarial voices
Two attacking setups: different speaker attack setup, and conversion attack setup [39]	MOBIO and Voxforge	• Statistically significant difference with p-value = 0 (for males) and p-value = 0.0015 (for female) is observed between the mean FAR of the two attacking methods on ASV system
		• Conversion attack is significantly more successful than the different speaker attack
SIRENATTACK: based on particle swarm optimization (PSO) and fooling gradient method [40]	Common Voice dataset and VCTK	• The attack threat is evaluated on the DeepSpeech model, in black box and white box scenarios
		• In particular, on ASV systems, average success rate from 91.65% to 99.45% is achieved in black box scenario, on various models
Professional Swedish impersonator (male) [41]	–	• Low correlation between human perception and speaker verification system is observed
		• The human listeners perceive prosodic features in addition to the other speech characteristics. However, machine-based ASV systems do not take prosodic features into account
Voice identity morphing [42]	LibriSpeech	• Morph attack success rate of over 80% on two popular speaker recognition systems (ECAPA-TDNN and x-vector) is observed
Voice synthesis and deepfake attacks [43]	Customized Data	• More than 30% of the deepfake attacks were successful, and that there was at least one successful attack for more than half of the participants

Table 2 Classification of speakers for target selection to attack ASV system. After [12, 16]

Types of Speakers in an ASV	Symbolic Notation	Vulnerability to ASV
Well-behaved Speakers: **Sheep**		Not a vulnerability (Low FRR)
Difficult to Recognize Speakers: **Goats**		Increased FRR
Easy to Mimic (Easy to Attack): **Lambs**		Increased FAR
Successful at Imitating Other Speakers: **Wolves**		Increased FAR

be useful in determining how secure a closed-domain targeted ASV system is [45].

Notably, target selection is different from a speaker identification perspective. In the latter, a claimed identity is compared with all the speaker models and, then, the speaker model with the maximum closeness to the claimed identity is chosen. Contrary to this, in the former, as shown in Fig. 3, there is no single speaker claiming his/her identity and, hence, the ASV system has to be run in an iterative manner in order to include all the speakers. Moreover, the chosen target is responsible for maximum FAR, out of all the enrolled speakers.

3.2 Adversarial attacks

Adversarial attacks aim to intentionally misclassify input data to a machine learning (ML) model based on a minor signal perturbation, which forces the ML model to generate an incorrect output. Usually, the perturbation is so modest that it is not even perceivable by humans. The speech signal with the intentionally added perturbation is called as *adversarial example*. An adversarial example w.r.t. to an original speech signal x can be represented as:

$$\bar{x} = x + \delta, \tag{1}$$

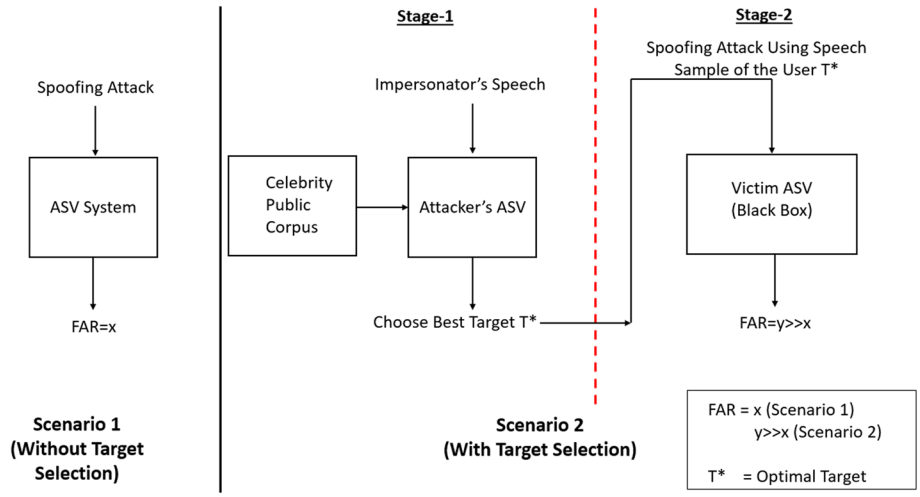


Fig. 3 Target selection: by using the attacker's ASV to attack the victim's ASV. After [16]

where δ is so small that $\bar{x}$ is perceptually the same as x. Nevertheless, δ is large enough to cause misclassification. This is in agreement with the finding that there may exist speech feature parameters that are acoustically relevant for ASV, e.g., fine structure features derived from glottal flow derivative waveform, but perceptually insignificant [46, 47]. Assuming that the ASV system to be attacked is a black box, from the attackers' perspective, to the best of our knowledge, the work reported in [48] was the first to propose adversarial attacks against machine learning (ML) methods, where the attacker has no access to a large training dataset. The attack is performed by training an attacker's model based on the labels assigned by the existing *victim* ML model; however, the attacks presented in this work are not confined to ASV systems and pertain to more general adversarial attacks in machine learning.

Notably, in paper [49], the approach of target selection is combined with adversarial attack, wherein an adversarial attack is used to optimize master voices (MV), originally referred to as *"wolves"* in [12], where the search for MVs is performed by using a dictionary attack, i.e., one-by-one. Furthermore, in paper [50], adversarial attacks were evaluated on various scenarios including transferability of attacks, practicability of over-the-air attacks by replay, and human-imperceptibility to demonstrate the imperceptibility of adversarial samples.

3.3 Deepfake attacks

Deepfakes correspond to false (or fake) data in both audio and visual domains, which are generated using deep learning algorithms. *Deepfakes* become each time closer to the real data as the iterative process used to generate them. This has led to serious misuse of the deepfake technology [13, 51]. In speech, DNN models, such as *Wavenets* [52], are capable of generating artificial speech signals from speaker embeddings, providing state-of-the-art performance, when evaluated by human listeners. Another model, known as *Waveglow*, combines *Wavenets* and *Glow*, i.e., generative flow model [53]. It is capable of generating speech from multi-speaker datasets. Another interesting generative adversarial network (GAN)-based model known as speech enhancement generative adversarial network (SEGAN) uses input speech signals enhanced by a convolutional autoencoder [54] to perform noise-robust speech enhancement task [55]. Additionally, in [56], voice cloning based on speaker adaptation and speaker encoding is shown to be possible by training models using just a few samples. Another strategy by the attacker is that he/she hides some small fake segment of audio in the genuine audio. This poses a serious threat since it is difficult to distinguish that small fake segment of audio from the whole speech utterance [57]. The experimental analysis in [57] shows that such partially fake audio is much more challenging to detect as compared to fully spoofed audio.

Each biometric sample or template in a biometric system is usually linked to a single identity. Recent studies, however, have shown that it is feasible to create "morph" biometric samples that can accurately match many identities.

3.4 Enrolled users with malicious intents

For all the types of attacks, the intuitive assumption is that the attacker will be an external entity. Thus, the realistic scenario of an enrolled speaker with malicious intent has been ignored during the design of current ASV systems. To that effect, *twins fraud* is a classic example of such a problem in the biometrics literature [14], where both the co-twins, in principle, are enrolled speakers in the ASV system. In that case, if one of the co-twins happens to have malicious intent, more specifically, malicious intent towards his/her other co-twin speaker, he/she will have more power, i.e., higher similarities in features, to fool the ASV system as compared to someone who is not enrolled. To that effect, Fig. 4 shows two utterances and their corresponds spectrograms, each corresponding to a co-twin speakers in a pair of twins. The utterances are taken from the twins corpus reported in [58], where the twins are a pair of 25 years old males at the time of recording. It can be observed from Fig. 4 that the overall pattern of spectral energy densities for twins are very much similar, if not identical, and moreover, spectral features, such as mel frequency cepstral coefficients (MFCCs) are predominantly used still today for ASV systems, thereby making twins fraud a serious technological challenge for ASV. The significance of the study of twins fraud was originally reported several decades ago in [59]; however, attention to this problem has not been sufficient in the ASV literature primarily due to practical issue w.r.t. unavailability of statistically meaningful twins corpora. This is why there is no anti-spoofing ASVSpoof challenge addressing CMs for twins spoof in the literature up-to-date. Furthermore, the situation for the design of CMs for mimicry is not different. This is due to the fact that mimicry attack is highly subjective, depending on the relative skillfullness of the potential attacker. Nevertheless, recent real case examples of twins fraud involved the *HSBC bank fraud*, where a BBC journalist and his non-identical twin spoofed the HSBC bank's voice authentication system [15]. To that effect, designing a robust countermeasure for such a case is a challenge since twins' physiological characteristics, such as size and shape of the vocal tract system [59], are practically indistinguishable. Furthermore, the countermeasure can also prevent genuine and zero-effort imposters from verification, thereby increasing FRR.

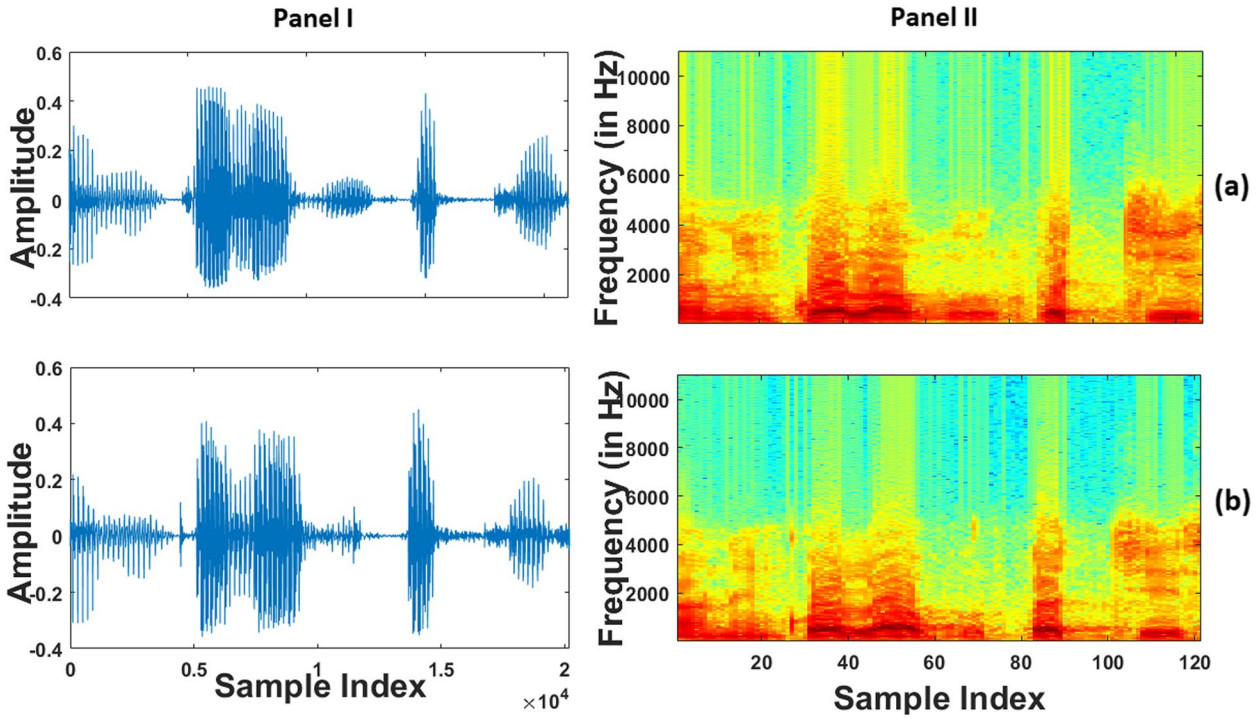


Fig. 4 Panel I shows the time-domain speech signal, and panel II shows the corresponding spectrogram for 25 years old male twin-pair where (**a**) 1st co-twin and (**b**) 2nd co-twin. Speech taken from the dataset in [58]

3.5 Complementary attack: utilizing the weakness of SSD systems

As of now, SSD systems are designed considering only a single type of attack. Therefore, we are far from designing a *versatile* SSD system, which would alleviate all the five types of presentation attacks, as well as unknown attacks. For instance, *segmental* information is responsible for twin's fraud, and prosodic information is found to be significant for skillful mimicry attacks, whereas reverberation, transmission channel, and acoustic environment-related information are useful for replay attacks [60]. Thus, SSD systems will always have a limitation on the types of attacks they can anti-spoof. Although the attacker is an independent entity, being free to come up with a new attack, which can be an amalgamation of the various kinds of spoofing attacks, the SSD will not be able to anti-spoof it in real-life practical settings, unfortunately. This means that we are yet to design an universal attack-proof mechanism for ASV system. Hence, the current SSD systems still give a great margin to the attacker to mount an unknown attack on ASV.

4 Technological challenges faced by the attacker

In this section, we present various issues the attacker faces in order to attack any given ASV system.

4.1 Number of trials on victim ASV access

In realistic scenarios, an effective ASV system should have an upper limit to the number of trials that can be allowed for a particular speaker. Nevertheless, an assumption for target selection attacks is that the attacker can have in principle, an infinite number of trials, since the attacker uses his/her own ASV to attack, to effectively practice the mimicry, which is impossible in practical scenarios of ASV system development.

4.2 Corpora for attacker's perspective

The attacker can proceed with the target selection attacks only when the corpus used for ASV is public, such as VoxCeleb. This is because target selection should be performed over the same corpora as that of the victim ASV. If this is not the case, then the probability of a good LLR score will decrease drastically, as the probability of the existence of a speaker, who is also the most vulnerable in two different datasets is almost negligible.

Not only this but various corpora are available in the literature w.r.t. anti-spoofing research, such as the ASVSpoof 2015, 2017, 2019, and 2021 datasets; however, these standard datasets are limited to a fixed number of configurations of data collection setup and recording conditions. Moreover, datasets are prepared with certain underlying assumptions. Such assumptions

keep us far away from developing anti-spoofing systems suitable for real-world applications. For instance, the generation of spoof utterances in ASVSpoof 2015 dataset is limited to ten algorithms of VC and SS. Similarly, the replay spoofing utterances in ASVSpoof 2017, 2019, and 2021 datasets are limited to a fixed number of recording configurations. This makes the attacker to mount complementary attacks by utilizing the weakness of the underlying SSD system because till now the corpora for anti-spoofing are limited to a specific attack only. Therefore, we are far away from designing a versatile SSD system that would alleviate all five types of presentation attacks as well as unknown attacks. Additionally, these publicly available corpora are in principle, available to the attacker as well. To that effect, attacks over unprotected corpora can be used to determine personal information about speakers using techniques, such as *target selection*, which enables an attacker to select the most vulnerable speaker from a corpus [16, 45]. Figure 5 shows a Venn diagram w.r.t. the publicly available corpora for developing anti-spoofing defenses against various spoofing attacks. Datasets, such as the ASVSpoof 2015, 2019 LA, and 2021 LA, share two common spoofing attacks, namely, voice conversion and speech synthesis. However, these datasets are not structured w.r.t. other spoofing attacks like replay, deepfake, and twins attacks. Likewise, datasets, such as BTAS, ReMASC, VSDC, POCO, ASVSpoof 2017, 2019 PA, and 2021 PA are focused only on replay attack conditions. These datasets lack the environmental and recording conditions for other spoofing attacks, such as voice conversion, speech synthesis, and deepfakes. Nevertheless, it should be noted that there exists no dataset that aims at developing CMs for *all* the spoofing attacks. This situation is denoted by "?" in Fig. 5. Therefore, there is still a long way to come up with generalized CMs, that are suitable for real-world SSD deployment.

4.3 Transmission channel

As per the recent anti-spoofing literature, transmission channel conditions are known to play an important role in the performance of the SSD systems. Therefore, anti-spoofing over a phone channel was chosen as the topic of the recent ASVSpoof 2021 challenge [5]. Thus, the transmission channel also forms one of the technological challenges in the attacker's perspective as well.

4.4 Perturbation minuteness in adversarial attacks

While attacking by adversarial ML approach, the boon for the attacker can even become disadvantageous. The small perturbation might not be captured over the air, causing the attack to be unsuccessful, specially in case of voice assistant systems [61]. Consequently, over the air, the performance of perturbed signals should also be considered, while evaluating the chances of a successful attack by adversarial ML methods. Furthermore, the perturbation should be such that it bypasses any smoothing technique used in the ASV system [62].

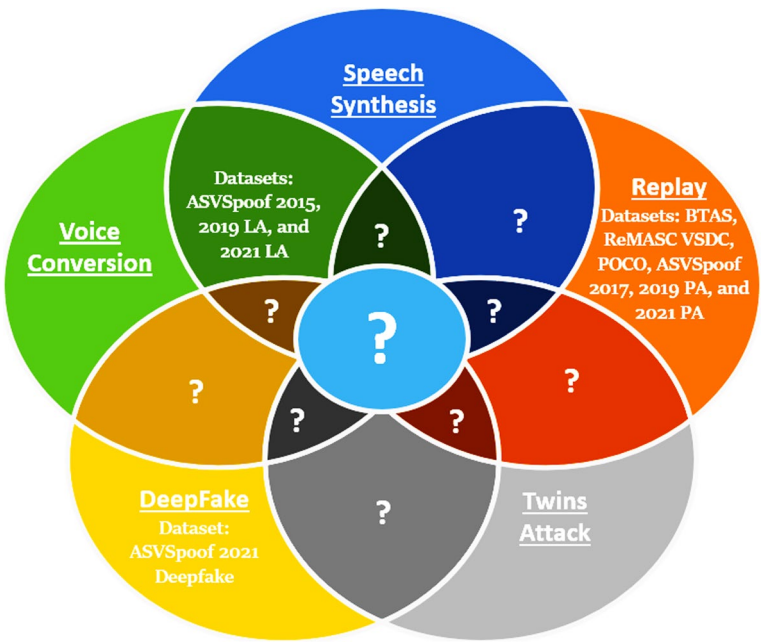


Fig. 5 Publicly available corpora for anti-spoofing research, and the associated known attacks. Here, "?" indicates gap area to develop anti-spoofing corpora from attacker's perspective

4.5 Voice privacy systems

Voice privacy (VP) aims to hide a speaker's identity while retaining the speech's linguistic content and naturalness [23, 63]. If the users publish data without anonymization, the attacker gains illegal access to it and can further use speakers' information to attack the ASV system. If a speech signal undergoes a considerably good algorithm for VP, it will be almost impossible for an attacker to perform target selection due to the absence of mapping between the speech data and actual speaker identity [64, 65]. If VP is used, then the most vulnerable target T^* cannot be chosen correctly. Consequently, the approach of *optimal* target selection will not be useful, and the attacker will be left with only a few attack strategies.

In Fig. 6, when the published data does not undergo voice privacy, the attacker is likely to have a successful attack. However, when the published data is anonymized using voice privacy techniques, the attacker does not have access to the actual information about the speakers, and hence, an attack using anonymized data is most likely to fail, and the attacker will not be granted authorization by the ASV system.

4.6 Voice liveness detection (VLD)

The countermeasure solutions developed in the ASVSpoof challenges are specific to particular attacks. Given the attacks on the ASV systems can be known or unknown attacks, VLD systems aim to detect only the live speech signal and reject all the other *non-live* speech, which are generated from known and unknown attacks [67, 68]. VLD is an emerging research area in which *pop noise* has been used actively as a discriminative acoustic cue to detect live speech [69–71]. Pop noise is generated by live speakers due to the breathing effect captured by the microphone if the speaker is in close proximity to the microphone. VLD systems enhance the security of the ASV system. Given that VLD systems aid in enhancing the robustness against attacks on ASV, it has also become a technological challenge for attackers. In particular, VLD systems are highly efficient against replay attacks. Replay attack requires only a recording device to capture a genuine user's voice from a distance. The attacker can then replay the recorded speech to spoof the ASV system; however, as shown in Fig. 7, due to the distance of the recording device from the speaker, liveness cues, such as pop noise, are faintly captured or even absent in some cases. Moreover, even in the case of artificially synthesized signals, a playback device/loudspeaker is needed to mount the attack, which in turn diminishes the strength of pop noise, which is strongly present in live speech.

Moreover, till now, VLD is performed w.r.t. replay attacks only; however, the scope of VLD in other spoofing techniques, such as VC and SS, remains to be explored.

4.7 *Deepfake* detectors

Advances in *deepfake* generation techniques have made fake data each time more accessible. Thus, *deepfake* detection has gathered immense interest, especially in images and videos [73, 74]. Nevertheless, given the interest of this paper, we focus our discussion on speech *deepfake* detectors, which have not been considered as much as image and video *deepfake* detectors. In [75], higher-order power spectrum correlations are considered in the frequency domain. Bi-spectral characteristics, such

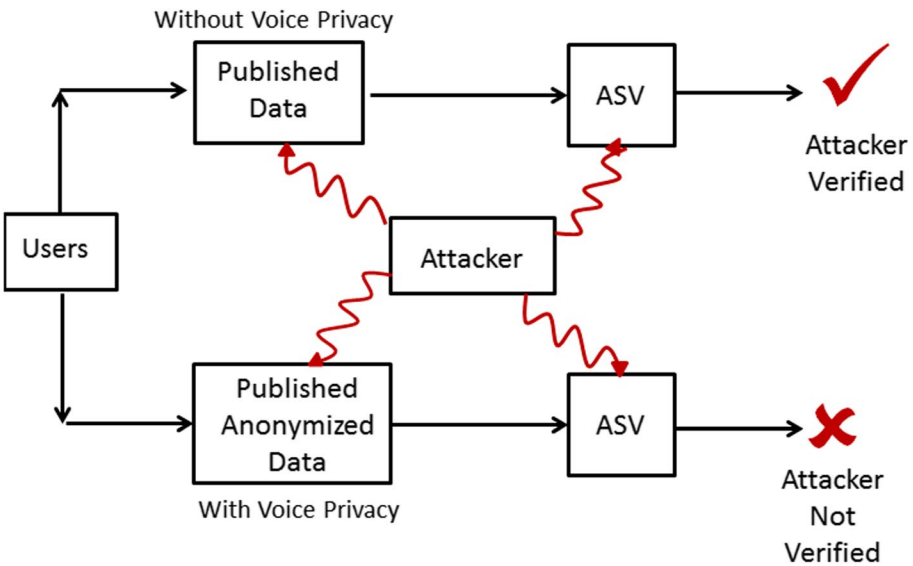


Fig. 6 Game between an attacker and VP system. After [66]

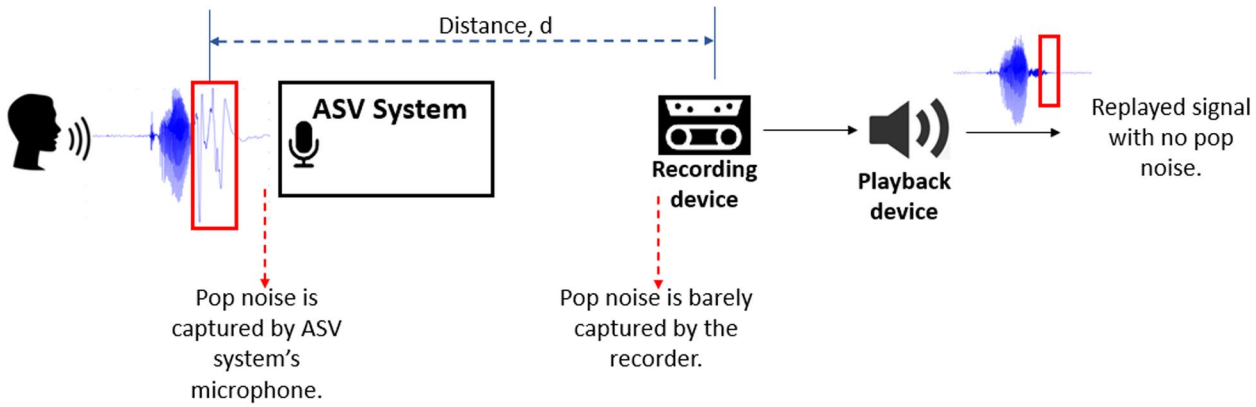


Fig. 7 A typical experimental setup for VLD task. The rectangular box (in red) before the ASV system contains pop noise, whereas pop noise is absent in replay speech. After [72]

as bi-coherent magnitude and phase spectra, were used to observe third-order correlations. Differences were observed in the bi-coherent magnitude and phase spectra between natural and synthetic speech. In [76], semantically rich information was extracted by using latent representation. Particularly, *XcepTemporal* convolutional recurrent neural network was introduced for *deepfake* detection by stacking multiple convolution modules. Recently, Whisper (which is a state-of-the-art ASR system [77]) features were used for the detection of deepfake on the ASVspoof 2021 DF dataset [78]. Furthermore, with the ADD 2023-the Second Audio Deepfake Detection Challenge, research towards deepfake detection has paced, and the best-performing system so far used Wav2vec2.0 architecture [79].

5 Summary and conclusions

The main objective of this study was to introspect attackers' perspectives to understand possible vulnerabilities of ASV systems in such a way that countermeasures for SSD systems can be designed effectively. In addition, while designing SSD systems for ASV, attacks can be used as benchmarks for testing the security of those systems. A new test, i.e., "attacker's test," can be performed with each update to the ASV system.

In addition, with the advancement in adversarial machine learning, over-the-air performance, i.e., noise introduction over various channels, should also be evaluated for increased chances of successfully attacking the system. Furthermore, privacy-preservation by VP systems should also be a topic of future interest. To that effect, the classification of speakers, as various categories in Doddington's menagerie [12], on the basis of their vulnerability even after voice privacy remains an open research question from the attacker's perspective. If an anonymization system is used, then the attacker's attempts towards the target selection approach will fail. Moreover, the attacker's perspective is different for a voice privacy system than for an ASV system, in a way that in a voice privacy system, the attacker can attempt to de-anonymize the output. This perspective remains an open research problem. Contrary to this, cryptography algorithms have their own limitations, i.e., deployment and increased computational complexity [23]. Therefore, if their deployment is simplified, and the computational complexity is dealt with optimally in implementation, more secure systems can be designed in the near future. Furthermore, given the various vulnerabilities associated with the secure design of ASV systems, such as the lack of generalized CMs to anti-spoof all the five known spoofing attacks on ASV, with issues of generalizability of CMs and VLD systems, we are far from designing robust and secure ASV systems.

Acknowledgements
The authors would like to thank the organizers of INTERSPEECH 2020 special session on attacker's perspective, ASVSpoof challenge organizers, and voice privacy challenge organizers for releasing standard and statistically meaningful speech corpora without which the synergistic research activities taken up by the community across the world would not have been possible. They also thank the authorities of DA-IICT Gandhinagar, India, for their support in carrying out this research work.

Authors' contributions
All the authors contributed equally.

Funding
R.C.Guido gratefully acknowledges the grants provided by the Brazilian agencies "National Council for Scientific and Technological Development (CNPq)" - Brazil, and "The State of São Paulo Research Foundation (FAPESP)" - Brazil, respectively through the processes 303854/2022-7 and 2021/12407-4, in support of this research.

Availability of data and materials
It does not apply.

Declarations

Competing interests

The authors declare no competing interests.

Received: 19 April 2023 Accepted: 9 January 2024
Published online: 10 February 2024

References

1. A.T. Patil, R. Acharya, H.A. Patil, R.C. Guido, Improving the potential of enhanced teager energy cepstral coefficients (ETECC) for replay attack detection. Comput. Speech Lang. (72), 101281 (2022)
2. W. Zhizheng, et. al., in *INTERSPEECH*, ASVSpoof 2015: The first automatic speaker verification spoofing and countermeasures challenge (Dresden, 2015), pp. 2037–2041
3. R. Font, J.M. Espín, M.J. Cano, in *INTERSPEECH*, Experimental analysis of features for replay attack detection-results on the ASVSpoof 2017 challenge (Stockholm, 2017), pp. 7–11
4. P. Gupta, P.K. Chodingala, H.A. Patil, Replay spoof detection using energy separation based instantaneous frequency estimation from quadrature and in-phase components. Comput. Speech Lang. **77**, 101423 (2023)
5. J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K.A. Lee, T. Kinnunen, N. Evans, et al., in *ASVSpoof Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection (Satellite Event, 2021), https://www.asvspoof.org/index2021.html. Accessed 20 Mar 2023
6. ASVSpoof5 evaluation plan. https://www.asvspoof.org. Accessed 27 Nov 2023
7. J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, et al., in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Add 2022: the first audio deep synthesis detection challenge (IEEE, 2022), pp. 9216–9220
8. Add 2023: The second audio deepfake detection challenge (2023), http://addchallenge.cn/add2023. Accessed 20 Mar 2023
9. T. Kinnunen, H. Delgado, N. Evans, K.A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi et al., Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2195–2210 (2020)
10. T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, E. Khoury, in *In Proceedings of the Odyssey Speaker and Language Recognition Workshop*, Generalization of audio deepfake detection (Tokyo, 2020), pp. 1-5
11. R.K. Das, X. Tian, T. Kinnunen, H. Li, in *INTERSPEECH 2020*, The attacker's perspective on automatic speaker verification: an overview (Shanghai, 2020), pp. 4213–4217
12. G. Doddington, W. Liggett, A. Martin, M. Przybocki, D. Reynolds, Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 Speaker Recognition Evaluation (NIST, Gaithersburg, 1998), Tech. rep
13. Why deepfakes are the greatest threat to the idea of truth. https://timesofindia.indiatimes.com/india/why-deepfakes-are-the-greatest-threat-to-the-idea-of-truth/articleshow/78075687.cms. Accessed 2 Feb 2022
14. A.K. Jain, S. Prabhakar, S. Pankanti, On the similarity of identical twin fingerprints. Pattern Recognit. **35**(11), 2653–2663 (2002)
15. HSBC reports high trust levels in biometric tech as twins spoof its voice ID system. Biom. Technol. Today. 2017(6), 12 (2017)
16. P. Gupta, H.A. Patil, in *Voice Biometrics: Technology, Trust and Security, Carmen Gracia-Mateo and Gerad Chollet* eds. Voice biometrics: Attackers perspective (IET, UK, 2021), pp. 39-65
17. P. Kocher, J. Jaffe, B. Jun, P. Rohatgi, Introduction to differential power analysis. J. Cryptographic Eng. **1**(1), 5–27 (2011)
18. P. Kocher, J. Jaffe, B. Jun, in *Annual International Cryptology Conference*, Differential power analysis (Springer, Santa Barbara, 1999), pp.388–397
19. R. Kumar, P. Jovanovic, W. Burleson, I. Polian, in *In IEEE, Workshop on Fault Diagnosis and Tolerance in Cryptography*, Parametric trojans for fault-injection attacks on cryptographic hardware (Busan, 2014), pp. 18–28
20. Document ISO/IEC, Information technology- security techniques-biometric information protection. ISO/IEC JTCI SC27 Secur. Tech. **24745**, 2011 (2011)
21. M. Gomez-Barrero, J. Galbally, C. Rathgeb, C. Busch, General framework to evaluate unlinkability in biometric template protection systems. IEEE Trans. Inf. Forensic Secur. **13**(6), 1406–1420 (2017)
22. B.M.L. Srivastava, A. Bellet, M. Tommasi, E. Vincent, Privacy-preserving adversarial representation learning in ASR: Reality or illusion? (2019), arXiv preprint arXiv:1911.04913. Accessed 9 Aug 2020
23. A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M.A. Hmani, A. Mtibaa et al., Preserving privacy in speaker and speech characterisation. Comput. Speech Lang. **58**, 441–480 (2019)
24. Y.W. Lau, M. Wagner, D. Tran, in *International Symposium on Intelligent Multimedia, Video, and Speech Processing*, Vulnerability of speaker verification to voice mimicking (Hong Kong, 2004), pp. 145–148
25. J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, T. Kinnunen, Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data (2018), arXiv preprint arXiv:1803.00860. Accessed 10 Aug 2020
26. V. Vestman, B. Soomro, A. Kanervisto, V. Hautamäki, T. Kinnunen, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Who do i sound like? Showcasing speaker recognition technology by YouTube voice search (Brighton, 2019), pp. 5781–5785
27. Yee Wah Lau, M. Wagner, D. Tran, in *Proc. of International Symposium on Intelligent Multimedia, Video and Speech Processing*, Vulnerability of speaker verification to voice mimicking (Hong Kong, 2004), pp. 145–148
28. T. Kinnunen, R.G. Hautamäki, V. Vestman, M. Sahidullah, in *ICASSP*, Can we use speaker recognition technology to attack itself? Enhancing mimicry attacks using automatic target speaker selection (Brighton, 2019), pp. 6146–6150
29. X. Tian, R.K. Das, H. Li, in *Odyssey 2020 The Speaker and Language Recognition Workshop*, Black-box attacks on automatic speaker verification using feedback-controlled voice conversion (Tokyo, 2020)
30. X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, H. Meng, in *ICASSP*, Adversarial attacks on GMM i-vector based speaker verification systems (Barcelona, 2020), pp. 6579–6583
31. Q. Wang, P. Guo, S. Sun, L. Xie, J.H. Hansen, in *INTERSPEECH*, Adversarial regularization for end-to-end robust speaker verification (Graz, 2019), pp. 4010–4014
32. Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, B. Yuan, Real-time, robust and adaptive universal adversarial attacks against speaker recognition systems. J. Signal Proc. Syst. **93**, 1–14 (2021)
33. A. Gomez-Alanis, J.A. Gonzalez, A.M. Peinado, in *Proc. INTERSPEECH 2021*, Adversarial transformation of spoofing attacks for voice biometrics (Valladoid, 2021), pp. 255–259, http://dx.doi.org/10.21437/IberSPEECH.2021-54. Accessed 2 Apr 2021
34. M. Pal, G. Saha, On robustness of speech based biometric systems against voice conversion attack. Appl. Soft Comput. **30**, 214–228 (2015)
35. Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Adv. Neural Inf. Process. Syst. **31**, 4485-4495 (2018)
36. K. Qian, Y. Zhang, S. Chang, X. Yang, M. Hasegawa-Johnson, in *International Conference on Machine Learning (ICML)*, Autovc: Zero-shot voice style transfer with only autoencoder loss (Long Beach, 2019), pp. 5210–5219
37. Y. Gao, J. Lian, B. Raj, R. Singh, in *IEEE Spoken Language Technology Workshop (SLT)*, Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems (Virtual Conference, 2021), pp. 544–551
38. G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, W. Xu, in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dolphinattack: Inaudible voice commands (Dallas, 2017), pp. 103–117
39. D. Mukhopadhyay, M. Shirvanian, N. Saxena, in *European Symposium on Research in Computer Security*, All your voices are belong to us: Stealing voices to fool humans and machines (Springer, Vienna, 2015), pp.599–621
40. T. Du, S. Ji, J. Li, Q. Gu, T. Wang, R. Beyah, in *Proceedings of the15thACM Asia Conference on Computer and Communications Security*, Sirenattack: Generating adversarial audio for end-to-end acoustic systems (Taiwan, 2020), pp. 357–369

41. E. Zetterholm, M. Blomberg, D. Elenius, A comparison between human perception and a speaker verification system score of a voice imitation. Evaluation 119(116.4), 116–4 (2004)
42. S.K. Pani, A. Chowdhury, M. Sandler, A. Ross, Voice morphing: Two identities in one voice (2023), arXiv preprint arXiv:2309.02404. Accessed 25 Nov 2023
43. D. Bilika, N. Michopoulou, E. Alepis, C. Patsakis, Hello me, meet the real me: Audio deepfake attacks on voice assistants (2023), arXiv preprint arXiv:2302.10328. Accessed 25 Nov 2023
44. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, X-vectors: Robust DNN embeddings for speaker recognition (IEEE, 2018), pp. 5329–5333
45. V. Vestman, T. Kinnunen, R.G. Hautamäki, M. Sahidullah, Voice mimicry attacks assisted by automatic speaker verification. Comput. Speech Lang. **59**, 36–54 (2020)
46. M.D. Plumpe, T.F. Quatieri, D.A. Reynolds, Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Process. **7**(5), 569–586 (1999)
47. T.F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1st edn. (Pearson Education India, 2006)
48. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z.B. Celik, A. Swami, in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, Practical black-box attacks against machine learning (Abu Dhabi, 2017), pp. 506–519
49. M. Marras, P. Korus, N. Memon, G. Fenu, in *INTERSPEECH*, Adversarial optimization for dictionary attacks on speaker verification (Graz, 2019), pp. 2913–2917
50. G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, Y. Liu, in *2021 IEEE Symposium on Security and Privacy (SP), virtual*, Who is real Bob? adversarial attacks on speaker recognition systems (IEEE, 2021), pp. 694–711
51. Fraudsters Cloned Company Director's Voice In \$35 Million Bank Heist, Police Find. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=76e31e847559. Accessed 2 Feb 2022
52. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, in *9th ISCA Speech Synthesis Workshop (SSW)*, Wavenet: A generative model for raw audio (CA, 2016), p. 125
53. R. Prenger, R. Valle, B. Catanzaro, in *ICASSP*, Waveglow: A flow-based generative network for speech synthesis (Brighton, 2019), pp. 3617–3621
54. S. Pascual, A. Bonafonte, J. Serrà, in *INTERSPEECH*, Segan: Speech enhancement generative adversarial network (Stockholm, 2017), pp. 3642–3646
55. N. Adiga, Y. Pantazis, V. Tsiaras, Y. Stylianou, in *INTERSPEECH*, Speech enhancement for noise-robust speech synthesis using Wasserstein GAN (Graz, 2019), pp. 1821–1825
56. S. Ö. Arık, J. Chen, K. Peng, W. Ping, Y. Zhou, in *Proceedings of the32nd International Conference on Neural Information Processing Systems (NIPS)*, Neural voice cloning with a few samples (Montreal, 2018), pp. 10040–10050
57. J. Yi, Y. Bai, J. Tao, Z. Tian, C. Wang, T. Wang, R. Fu, Half-truth: A partially fake audio detection dataset (2021), arXiv preprint arXiv:2104.03617. Accessed 27 Nov 2023
58. H.A. Patil, Speaker recognition in Indian languages: A feature based approach (Ph. D. Thesis, Dept. of Electrical Engineering, Indian Institute of Technology (IIT), Kharagpur, 2005)
59. A.E. Rosenberg, Automatic speaker verification: A review. Proc. IEEE. **64**(4), 475–487 (1976)
60. R. Prasad, B. Yegnanarayana, in *INTERSPEECH*, Acoustic segmentation of speech using zero time liftering (ZTL) (Lyon, 2013), pp. 2292–2296
61. Y. Gong, C. Poellabauer, An overview of vulnerabilities of voice controlled systems (2018), arXiv preprint arXiv:1803.09156. Accessed 21 Apr 2020
62. W. Xu, D. Evans, Y. Qi, Feature squeezing: Detecting adversarial examples in deep neural networks, Proceedings Network and Distributed System Security Symposium (2017), arXiv arXiv:1704.01155. Accessed 14 May 2020
63. The Voice Privacy 2020 Challenge Evaluation Plan, https://www.voiceprivacychallenge.org. Accessed 18 Feb 2020
64. J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X. Li, Speech sanitizer: Speech content desensitization and voice anonymization. IEEE Trans. Dependable Secure Comput. **18**(6), 1 (2019)
65. P. Gupta, G.P. Prajapati, S. Singh, M.R. Kamble, H.A. Patil, in *APSIPA-ASC*, Design of voice privacy system using linear prediction (Auckland, 2020), pp. 543–549
66. P. Gupta, S. Singh, G.P. Prajapati, H.A. Patil, Voice privacy in biometrics (Springer International Publishing, Cham, 2023), pp. 1–29. https://doi.org/10.1007/978-3-031-15816-2_1
67. K. Akimoto, S.P. Liew, S. Mishima, R. Mizushima, K.A. Lee, in *INTERSPEECH*, POCO: A voice spoofing and liveness detection corpus based on pop noise (Shanghai, 2020), pp. 1081–1085
68. P. Gupta, H.A. Patil, Morse wavelet transform-based features for voice liveness detection. Comput. Speech Lang. **84**, 101571 (2024)
69. S. Mochizuki, S. Shiota, H. Kiya, Voice livness detection based on pop-noise detector with phoneme information for speaker verification. J. Acoust. Soc. Am. **140**(4), 3060 (2016)
70. P. Gupta, S. Gupta, H. A. Patil, in *International Conference on Pattern Recognition and Machine Intelligence*, Voice liveness detection using bump wavelet with CNN (LNCS, Springer, 2021)
71. P. Gupta, H. Patil, in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Effect of speaker-microphone proximity on pop noise: Continuous wavelet transform-based approach (2022), pp. 110–114. https://doi.org/10.1109/ISCSLP57327.2022.10038157
72. P. Gupta, P.K. Chodingala, H.A. Patil, in *European Signal Processing Conference (EUSIPCO)*, Morlet wavelet-based voice liveness detection using convolutional neural network (Belgrade, 2022), pp. 100–104
73. H.H. Nguyen, J. Yamagishi, I. Echizen, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Capsule-forensics: Using capsule networks to detect forged images and videos (Brighton, 2019), pp. 2307–2311
74. H.-S. Chen, M. Rouhsedaghat, H. Ghani, S. Hu, S. You, C.-C.J. Kuo, Defakehop: A light-weight high-performance deepfake detector. arXiv e-prints (2021) arXiv–2103. Accessed 26 Feb 2022
75. E.A. AlBadawy, S. Lyu, H. Farid, in *CVPR Workshops*, Detecting ai-synthesized speech using bispectral analysis (Long Beach, California, 2019)
76. A. Chintha, B. Thai, S.J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, R. Ptucha, Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE J. Sel. Top. Signal Process. **14**(5), 1024–1037 (2020). https://doi.org/10.1109/JSTSP.2020.2999185
77. A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, in *International Conference on Machine Learning*, Robust speech recognition via large-scale weak supervision (PMLR, 2023), pp. 28492–28518
78. P. Kawa, M. Plata, M. Czuba, P. Szymański, P. Syga, Improved deepfake detection using whisper features (2023), arXiv preprint arXiv:2306.01428. Accessed 25 Nov 2023
79. J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C.Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, et al., Add 2023: the second audio deepfake detection challenge (2023), arXiv preprint arXiv:2305.13774. Accessed 25 Nov 2023

Publisher's Note