The Attacker's Perspective on Automatic Speaker Verification: An Overview

Rohan Kumar Das[1], *Xiaohai Tian*[1], *Tomi Kinnunen*[2] *and Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]School of Computing, University of Eastern Finland, Joensuu, Finland

{rohankd, eletia, haizhou.li}@nus.edu.sg, tkinnu@cs.uef.fi

Abstract

Security of automatic speaker verification (ASV) systems is compromised by various spoofing attacks. While many types of *non-proactive* attacks (and their defenses) have been studied in the past, *attacker's* perspective on ASV, represents a far less explored direction. It can potentially help to identify the weakest parts of ASV systems and be used to develop attacker-aware systems. We present an overview on this emerging research area by focusing on potential threats of adversarial attacks on ASV, spoofing countermeasures, or both. We conclude the study with discussion on selected attacks and leveraging from such knowledge to improve defense mechanisms against adversarial attacks.

Index Terms: automatic speaker verification, attacker, spoofing, adversarial attacks

1. Introduction

Automatic speaker verification (ASV) technology is now a matured technology used in access control, forensics and surveillance applications [1, 2]. Unfortunately, unprotected ASV systems are highly vulnerable to various *spoofing attacks* [3] where an attacker (adversary) masquerades him/herself as a specific targeted user. This has motivated the study of automatic detection of spoofing attacks [4]. Such *countermeasures* have been studied as one of the important topics in system implementation, either independently of, or in conjunction with ASV.

ASVspoof challenge series [5] is a community-driven bechmarking effort to address voice spoofing attacks and their defenses. The attacks include various *voice conversion* (VC) and *text-to-speech synthesis* (TTS) techniques along with audio replay [6]. Their impact upon ASV is now far better understood than a decade ago. Nonetheless, vast majority of research in this domain focuses on *non-proactive* attacks, where the adversary takes no direct use of the attacked system. For instance, the typical objective of VC and TTS is to maximize perceptual speaker similarity and audio quality, rather to break ASV systems.

Apart from studying robust spoofing countermeasures, it is important to study the weak links of ASV to protect it from various types of attacks. In order to identify the loopholes of ASV, we need to assess the limits of spoofing attacks *from the perspective of the attacker*. For an attacker, the ideal way is to attack within the functional modules of an ASV system [4]. But this may not be feasible always as it requires access to various modules of the system. Another way to deceive a system is to craft so-called *adversarial examples* [7]. They are novel inputs crafted with some knowledge of the attacked system. Adversarial attacks have received a lot of attention across different classification tasks (especially within image processing) [8], but comparatively less in the speech field. While adversarial attacks and their defenses can be motivated from security improvement against 'hackers', another viewpoint is general robustness improvement. Modern ASV systems are robust to many perturbations, but many spoofing countermeasures lack this property.

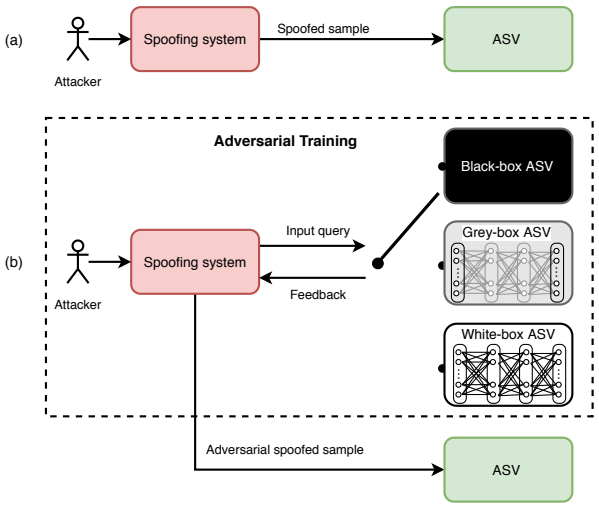

Figure 1: *Spoofing from attacker's perspective (a) non-proactive attacks (b) adversarial attacks: using black-box, grey-box and white-box ASV.*

Figure 1 illustrates both non-proactive and adversarial attacks from the perspective of the the attacker. In the latter case, the attacker leverages from information of the attacked ASV system to generate spoofed samples. The attacker can use the knowledge of either the attacked ASV or *another* similar ASV to generate adversarial samples. The former is more effective (but potentially less realistic). Adversarial attacks can be broadly divided into **black-**, **grey-** and **white-box** attacks [9]. In the first case, the attacker has access only to the system output (speaker similarity score or hard accept/reject decision) to guide crafting of new inputs [10]. The grey-box attacks are a step further, where the attacker has some information such as features of the speakers and their implementation, but not their statistical models [11]. Finally, the white-box attacks pose the greatest threat as the attackers have full knowledge of the system under attack [9]. Recent studies using adversarial attacks on various applications have demonstrated their threat to fool the system behavior [7, 12, 13].

Studies specifically focused on adversarial attacks on ASV have come up only very recently [14–18] and some of this work has revealed new, potential threats. We present an overview of these studies. Non-proactive spoofing attacks are also briefly discussed for a broader context. In general, spoofing attacks can be performed either on the core ASV system, spoofing countermeasures or both. We group the studies on these grounds. Further, we discuss the results of different such attacks and possible emerging defense strategies against these attacks.

2. Spoofing with Non-proactive Attacks

This section presents a review on traditional, or **non-proactive** spoofing attacks that use limited prior knowledge of the attacked ASV system. These attacks can be broadly divided into four categories, impersonation, replay, VC and TTS [4]. Impersonation is commonly referred as mimicry, where the attacker attempts to mimic the voice characteristics of the target speaker. Replay attacks are executed by replaying the previously recorded speech of the target speaker. Finally, VC and TTS attacks aim at modifying source speaker identity to that of a target speaker, and to produce text in a given target speaker's voice, respectively.

Fundamentally, crafting non-proactive attacks lacks a direct optimization target related to the attacked ASV system (such as false acceptance rate). Rather, such attacks represent ideas or technology originally designed with completely different aims and purposes in mind; they are taken *as-is* to execute stress tests on ASV systems. For instance, mimicry takes place in acting and stand-up comedy without any reference to ASV systems. Similarly, VC and TTS technology researcher may not consider themselves as developers of 'ASV attack technology' (any more than knife or gun manufacturers may consider themselves as developers of 'murder technology'). Finally, speech recorders and loudspeakers (used in mounting of replay attacks) is technology intended to reproduce recorded or transmitted speech, music, or any other audio to a human listener as faithfully as possible. The fact that TTS, VC and replay attacks *do* compromise the security of ASV systems is a lucky side-product[1] rather than the original aim. In the following subsections, we group the non-proactive attacks into three categories based on the their target to attack ASV with or without spoofing countermeasures.

2.1. Attacks on ASV

Different kinds of non-proactive attacks are investigated on ASV systems without spoofing countermeasures to showcase the impact of such attacks. Concerning mimicry attacks, impersonators tend to make vocal caricatures of their target speakers by mimicking high-level speaker cues such as prosody, accent, pronunciation and lexicon, more than the low-level spectral cues used by ASV systems. As a result, impersonation is not a consistent approach to attack ASV [19, 20].

In contrast to the mimicry attacks, the attacks generated by a VC or TTS system are optimized for both speaker similarity and quality. The aim of the former, a relevant concern for ASV, is to generate or modify speech so that it sounds as if spoken by a given target. This is often done by empirically minimizing a spectral distance measure [21, 22] between the synthesized (or modified) and target speech, as a proxy of time-consuming perceptual experiments. Even if simple spectral distance measures have only a weak connection to speaker similarity computations implemented in ASV systems, several studies indicate that ASV systems are nonetheless vulnerable to these attacks [23–26]. Further, modern VC and TTS systems are not tailored for a fixed set of speakers. High-quality target speaker voice can be generated by adapting an average voice model trained with multi-speakers' data towards the desired target [27] or by conditioning the model using a global (utterance-level) speaker variable [28]. These speaker-conditioning variables are similar (or even same) as *speaker embeddings* used in ASV systems. These developments have made ASV and TTS/VC technologies closer to each other, imposing imminent threat to ASV systems.

[1] Implied by *desirable* properties of the original technology, such as accurate reproduction of stored audio or target speaker voice timbre.

ASV systems are also vulnerable to replay attacks which use pre-recorded speech samples of the target speaker [29]. As replayed samples contain strong traits of the target speaker, they pose a critical threat on any unprotected ASV system, most notably text-independent ASV and text-dependent ASV without protection against wrong passphrase. For ASV systems protected against wrongly spoken passphrase, the replay attacks require the pre-recorded samples of the same spoken content and are unflexible. On the other hand, the attacks derived using VC and TTS systems can be performed by only knowing the lexical information of the target speaker in a such scenario.

2.2. Attacks on Spoofing Countermeasures

Spoofing countermeasures are introduced to the ASV systems to protect them from various attacks. An attacker may also try to attack only the spoofing countermeasures with non-proactive attacks that are not easily detectable. The studies on the first edition of ASVspoof challenge indicated that a *unit-selection* based attack, produced by concatenation of time-domain waveform samples confused many of the spoofing countermeasures [5]. On the other hand, attacks synthesized by using vocoders showed less threat to the spoofing countermeasures [5]. A further study [30] on the second edition of *voice conversion challenge* [31] suggested that modern waveform filtering based samples might be comparably harder to detect than traditional vocoded samples.

For replay attack countermeasures, the replay configuration plays a key role. An analysis presented in [32] suggested that replay speech generated with high quality recording and playback devices in clean environment can be particularly difficult to detect. This applies to *any* kind of countermeasure as the artifacts distinguishing replay speech from the bonafide speech is minimal in such a scenario; whenever the replayed speech becomes digitally indistinguishable from bonafide speech no (low-level) countermeasure will be able to detect it.

2.3. Attacks on ASV with Spoofing Countermeasures

We have discussed attacks on ASV and countermeasures separately, but they could also be performed on combined systems consisting of ASV and countermeasures. There are no extensive studies on this direction. Although [33] suggests that ASV with a spoofing countermeasure in combination might be less vulnerable to the attacks, the severity of attacks on ASV with spoofing countermeasure depends on the nature of their combination approach, which deserves further research.

3. Spoofing with Adversarial Attacks

We now turn our attention to proactive, or adversarial attacks, which have been explored for the case of TTS, VC and impersonation attacks. As far as the authors are aware of, replay attacks have not yet been investigated in an adversarial context.

3.1. Attacks on ASV

Optimizing input signals with partial or full knowledge of the attacked system is not a new concept in itself. For instance, *artificial signals* (that may bear no resemblance to human speech) have been successfully used to attack ASV [34]. A key difference in the adversarial attacks, however, is that the new signals are required to remain unnoticeable to human eye or ear — being perceptually indistinguishable from natural signals. A study [14] later generated adversarial samples by adding a perceptually indistinguishable structured noise to the original

test examples for attacking an end-to-end ASV system. Adversarial training uses the so-called *fast gradient sign method* (FGSM) [12] with white-box and black-box attacks in a cross-corpora and cross-feature setting considering the same ASV. The studies demonstrated the ability of the adversarial attacks to deceive ASV systems. Another recent study [35] also used FGSM to perform white-box and black-box attacks. It extended studies on adversarial *transferability* from one ASV to attack another ASV system [35].

Another adversarial attack against ASV, 'FakeBob', is addressed in [36]. This study uses black-box attacks by adding small perturbation to generate adversarial samples too, but considered different cases for practical scenario. These include studies with various ASV architectures (including commercial systems), transferability of attacks, practicality of over-the-air through replay and imperceptibility based on human perception. A further study, explored the real-time nature and feasibility of adversarial attacks replaying over-the-air by modeling room impulse response (RIR) during adversarial training [37, 38].

The authors of [39] investigated the effect of *dictionary attacks* on ASV. This kind of attack allows targeting large speaker population without having specific knowledge of individuals or their speech models [40]. They selected a set of non-target trials that have high false acceptance in a population for an ASV system. Given such a trial and the training utterances of the speaker population, a time-domain waveform, *master voice*, is learned by adding adversarial perturbations to maximize the spectrogram similarity. The time-domain waveform is generated by spectrogram inversion once the similarity exceeds a threshold to have a close match to a number of speakers in the population. The adversarial optimization of dictionary attacks were found to be imperative for deceiving ASV systems.

A verification-to-synthesis attack using white-box ASV is carried out in [16]. In this adversarial attack, a VC system is trained using white-box ASV model without target speaker training data (unlike traditional VC systems). As the trained network may distort the phonetic properties of the input voice, an automatic speech recognition model is also included as part of optimization to regulate loss of phonemic information. The output voice thus produced is not only able to deceive the ASV system, but also maintains the perceptual quality.

The authors of [18] studied black-box attacks on ASV through *feedback-controlled* VC framework. The authors treat the ASV system as a black-box with access to its detection score only. This score is taken as a feedback to train the VC system. The objective function for training the feedback-controlled VC is jointly optimized with the feedback ASV score. The results indicated that black-box attacks can degrade ASV performance. Additionally, listening experiments suggested that adversarial examples are indistinguishable from the VC examples generated without ASV feedback.

The above studies typically assume that the system accessed or queried by the attacker is the same the as the attacker finally wishes to attack. In contrast to this assumption, there are also studies that assume that the attacker *cannot* access the attacked system itself, but another ASV system, used as a proxy of the attacked one. The authors of [15] consider mimicry attacks where they find the closest target speaker for given attacker using a proxy ASV system. However, when asked to mimic their selected target speakers, the attackers did not manage to increase the detection score. The study indicated that mimicry, even when assisted by ASV-based, may not fool ASV. But an ASV system can be definitely used to assist the attacks on another ASV system.

3.2. Attacks on Spoofing Countermeasures

Adversarial attacks solely on spoofing countermeasures have received less attention. The authors of [41] proposed an adversarial training method for statistical parametric speech synthesis, where the loss function for training is modified by adding a weighted loss using an anti-spoofing system. As the loss function minimizes generation error as well as makes the distribution of synthetic speech close to that of natural speech, it is also able to deceive the anti-spoofing system apart from producing an improved speech quality. This work is extended for a generative adversarial network based synthetic speech generation framework, which also proved to be effective to increase the spoofing rate [42].

A recent work in [17] conducts white-box and black-box attacks on spoofing countermeasures. The authors consider one of the strong anti-spoofing system based on *light convolutional neural network* (LCNN) [43] to carry out the adversarial attacks with the FGSM and the projected gradient descent methods. The studies conducted with both white-box and black-box attacks indicate that the well performing spoofing countermeasures can be fooled by generating adversarial samples. Further, listening test revealed that the adversarial samples are indistinguishable from non-proactive samples.

3.3. Attacks on ASV with Spoofing Countermeasures

Adversarial attacks could also be carried out on ASV with spoofing countermeasure by leveraging from any prior information the attack has about either system. As far as the authors are aware, there is currently no (reported) research on this direction. However, as many real-world systems combine ASV and countermeasures, future work should address attacks (both non-proactive and proactive) against combined system.

4. Defenses to Adversarial Attacks on ASV

The spoofing conducted using adversarial attacks discussed in the previous section projects the weak spots of ASV. Although many countermeasures for non-proactive attacks are available, countermeasures for adversarial attacks need attention as well. In the field of machine learning, various defense mechanisms are employed to handle adversarial attacks [9, 46]. They can be categorized into *passive* and *proactive* defenses. The former aims to counter adversarial attacks without modifying the attacked system model. Proactive defenses, in turn, aim at training new models that are robust to adversarial samples. Motivated by such directions, there are some recent works that explore defense mechanisms against adversarial attacks on ASV.

Adversarial regularization is addressed in [44] to protect end-to-end ASV from adversarial attacks. The studies first generate adversarial samples by FGSM and *local distributional smoothness* (LDS) [47] method that are found to fool the ASV system. Therefore, the model is retrained with adversarial regularization as a defense mechanism. This mechanism aims at finding a worst spot around the current data point, and then optimize using this worst data point to derive a robust model [44]. The regularization is studied for both methods (FGSM-REG and LDS-REG) and is found to improve ASV performance against adversarial attacks.

Spoofing countermeasures also require defense mechanisms against adversarial attacks. A passive defense method namely, *spatial smoothing* [48] and another proactive method namely, *adversarial training* are studied to defend adversarial attacks for spoofing countermeasures [45]. The former is a sim-

Table 1: *ASV and spoofing countermeasure (CM) performance before (B), after (A) adversarial attacks, and post defense (D) applied (if any) in different metrics, accuracy (ACC), equal error rate (EER), attack success rate (ASR), spoofing rate (SR) and score comparison.*

Adversarial Attack/ Defense	Attack Type	ASV/CM System	Corpus	Performance (B/A/D)	Metric
Adding perturbation [14]	White/black-box	End-to-end	YOHO, NTIMIT	87.50/25.75/- (white-box)	ACC (%)
Adding perturbation [35]	White/black-box	x-vector, i-vector	VoxCeleb1	7.20/8.83/- (black-box, i-vector)	EER (%)
Adding perturbation [36]	Black-box	i-vector, GMM-UBM	LibriSpeech	-/70/- (i-vector)	ASR (%)
Adding perturbation with RIR [37]	White-box	x-vector	VCTK	10/50/-	ASR (%)
Adding perturbation with RIR [38]	White-box	x-vector	VCTK	1.33/90.19/-	ASR (%)
Dictionary attack [39]	White-box	VGGVox	VoxCeleb2	-/20 (female), 10 (male)/-	SR (%)
VC with feedback loss [16]	White-box	d-vector	Japanese data	NA	Scores
Feedback-controlled VC [18]	Black-box	i-vector	ASVspoof 2019	29.25/30.73/-	EER (%)
ASV assisted mimicry [15]	Black-box	x-vector, i-vector	VoxCeleb, self-collected	NA	Scores
TTS with feedback loss [41, 42]	White-box	DNN	ATR Japanese	NA	SR plot
Adding noise [17]	White/black-box	LCNN, SENet	ASVspoof 2019	3.87/4.69/- (white-box, LCNN)	EER (%)
Adding perturbation/ FGSM-REG, LDS-REG [44]	White-box	End-to-end	TIMIT	4.87/11.89/8.31 (FGSM-REG)	EER (%)
Adding perturbation/spatial smoothing, adversarial training [45]	White-box	SENet, VGG	ASVspoof 2019	99.97/48.32/93.76 (Adversarial training, SENet)	ACC (%)

ple method, where a slicing window moves over the power spectrum, then performs smoothing by use of filters such as median, mean and Gaussian, commonly used in the field of image processing. The general idea behind these simple noise suppression techniques is to suppress the impact of adversarial perturbations that are noise-like. The latter, in turn, leverages from adversarial samples at the training stage to improve robustness against attacks. Both methods are investigated on two spoofing countermeasures and are found as effective defense methods against the adversarial attacks.

5. Summary and Discussion

Before concluding, we summarize the cited literature on adversarial attacks and their defenses in Table 1. Generally, the studies are diverse in terms of the adversarial attacks, attacked systems, datasets and metrics. While these differences make it impossible to compare different studies, the available performance numbers within studies (when available) suggest that the adversarial attacks can severely degrade the ASV performance and that defenses are required for safeguarding systems from such attacks. Further, although the white-box attacks suggest a higher relative threat (as one might expect), black-box attacks might be more realistic; if the attacker already has full access to the system details, does he/she need to bother about generating spoofed samples?

The defense mechanisms in [44, 45] contributes to defend adversarial attacks as observed from Table 1. However, these methods learn to resist *particular kind of attack* in most cases. Therefore, such defense methods might be less effective when the attacker changes the settings of the attack [49].

To address such problem ensemble adversarial training is employed that generates a larger adversarial training examples by attacking several different models and then train the model by transferring the examples [46]. This kind of defense mechanism might be more favorable from the outlook of practical systems, where the nature of adversarial attacks is always unknown. The challenges associated with unknown attacks has already been noted in the context of ASVspoof challenges. The evaluation data (provided without ground-truth to participants) have purposefully included some 'surprises' — attacks not included in the training data, and these have turned out difficult to detect.

We find the adversarial attacks that are proactive in nature, have a definite impact for knowing the weak spots of ASV systems as discussed throughout the paper. Nevertheless, the defense mechanisms to tackle such attacks are more imperative for improving system robustness in real-world scenario. This remains as an important direction for futuristic ASV systems.

6. Conclusions

The overview presented in this work shows that the proactive or adversarial attacks have a higher threat to ASV than the non-proactive attacks. However, they are less explored and the existing studies are dispersed across different dataset designs, different ways to evaluate various attacks and their defenses. Further, considering the practicality of adversarial attacks and their defenses, there is a need to have a common protocol, performance metric, and corpus for future research. The special session on *The Attacker's Perspective on ASV* in Interspeech 2020 organized by the authors is a small step towards this direction.

7. Acknowledgements

This research work is supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain) and in part by the Academy of Finland (Proj. No. 309629 "NOTCH: NOn-cooperaTive speaker CHaracterization").

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12 – 40, 2010.

[2] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," in *SLTC Newsletter*, February 2013.

[3] "ISO/IEC 30107-1:2016, Information technology-Biometric presentation attack detection-part1:framework." [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-1:ed-1:v1:en

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

[5] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanili, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, June 2017.

[6] 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," in *Interspeech 2019*, 2019, pp. 1008–1012.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR 2014*, 2014.

[8] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[9] 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*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.

[10] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *AsiaCCS 2017*, 2017, pp. 506–519.

[11] B. S. Vivek, K. R. Mopuri, and R. V. Babu, "Gray-box adversarial training," in *Computer Vision – ECCV 2018*, 2018, pp. 213–228.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR 2015*, 2015.

[13] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *IEEE CVPR 2018*, 2018, pp. 1625–1634.

[14] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *IEEE ICASSP 2018*, 2018, pp. 1962–1966.

[15] V. Vestman, T. Kinnunen, R. G. Hautamäki, and M. Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech & Language*, vol. 59, pp. 36 – 54, 2020.

[16] T. Nakamura, Y. Saito, S. Takamichi, Y. Ijima, and H. Saruwatari, "V2S attack: building DNN-based voice conversion from automatic speaker verification," in *10th SSW 2019*, 2019, pp. 161–165.

[17] S. Liu, H. Wu, H. Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in *IEEE ASRU 2019*, 2019, pp. 312–319.

[18] X. Tian, R. K. Das, and H. Li, "Black-box attacks on automatic speaker verification using feedback-controlled voice conversion," in *Speaker Odyssey 2020*, 2020.

[19] Yee Wah Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Intelligent Multimedia, Video and Speech Processing 2004*, 2004, pp. 145–148.

[20] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13 – 31, 2015.

[21] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[22] Hui Ye and S. Young, "High quality voice morphing," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–9.

[23] J. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Interspeech 2007*, 2007, pp. 2053–2056.

[24] T. Kinnunen, Z. Wu, K. A. Lee, F. Sedlak, E. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *IEEE ICASSP 2012*, 2012, pp. 4401–4404.

[25] P. L. D. 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*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[26] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector plda: towards unifying speaker verification and transformation," in *2017 IEEE ICASSP 2017*, 2017, pp. 5535–5539.

[27] X. Tian, J. Wang, H. Xu, E.-S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data," in *Odyssey 2018*, 2018, pp. 227–232.

[28] W. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *ICLR 2019*, 2019.

[29] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311–5327, May 2016.

[30] T. Kinnunen, J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, and Z. Ling, "A spoofing benchmark for the 2018 voice conversion challenge: Leveraging from spoofing countermeasures for speech artifact assessment," in *Odyssey 2018*, 2018, pp. 187–194.

[31] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Odyssey 2018*, 2018, pp. 195–202.

[32] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: metadata analysis and baseline enhancements," in *Odyssey 2018*, 2018, pp. 296–303.

[33] P. Korshunov and S. Marcel, "Joint operation of voice biometrics and presentation attack detection," in *IEEE International Conference on BTAS 2016*, 2016, pp. 1–6.

[34] F. Alegre, R. Vipperla, N. Evans, and B. Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," in *EUSIPCO 2012*, 2012, pp. 36–40.

[35] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM i-vector based speaker verification systems," in *IEEE ICASSP 2020*, 2020, pp. 6579–6583.

[36] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real Bob? adversarial attacks on speaker recognition systems," *ArXiv*, vol. abs/1911.01840, 2019.

[37] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *ACM HotMobile 2020*, 2020, pp. 9–14.

[38] 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 *IEEE ICASSP 2020*, 2020, pp. 1738–1742.

[39] M. Marras, P. Korus, N. Memon, and G. Fenu, "Adversarial Optimization for Dictionary Attacks on Speaker Verification," in *Interspeech 2019*, 2019, pp. 2913–2917.

[40] A. Roy, N. Memon, and A. Ross, "Masterprint: Exploring the vulnerability of partial fingerprint-based authentication systems," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 9, pp. 2013–2025, 2017.

[41] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," in *IEEE ICASSP 2017*, pp. 4900–4904.

[42] ——, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions Audio, Speech & Language Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[43] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Interspeech 2019*, 2019, pp. 1033–1037.

[44] Q. Wang, P. Guo, S. Sun, L. Xie, and J. H. Hansen, "Adversarial Regularization for End-to-End Robust Speaker Verification," in *Interspeech 2019*, 2019, pp. 4010–4014.

[45] H. Wu, S. Liu, H. Meng, and H. yi Lee, "Defense against adversarial attacks on spoofing countermeasures of ASV," in *IEEE ICASSP 2020*, 2020, pp. 6564–6568.

[46] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *ICLR 2018*, 2018.

[47] T. Miyato, S. ichi Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional smoothing with virtual adversarial training," *ArXiv*, vol. abs/1507.00677, 2015.

[48] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *NDSS 2018*, 2018.

[49] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *ICLR 2017*, 2017.