Speaker Privacy and Security in the Big Data Era: Protection and Defense against Deepfake

Liping Chen*, Kong Aik Lee†, Zhen-Hua Ling*, Xin Wang‡, Rohan Kumar Das§, Tomoki Toda¶, Haizhou Li‖
* University of Science and Technology of China, China
Emails: {lipchen, zhling}@ustc.edu.cn
† The Hong Kong Polytechnic University, China
Email: kong-aik.lee@polyu.edu.hk
‡ National Institute of Informatics, Japan
Email: wangxin@nii.ac.jp
§ Fortemedia, Singapore
Email: ecerohan@gmail.com
¶ Nagoya University, Japan
Email: tomoki@icts.nagoya-u.ac.jp
‖ School of Data Science, Chinese University of Hong Kong, China
Email: haizhouli@cuhk.edu.cn

***Abstract*—In the era of big data, remarkable advancements have been achieved in personalized speech generation techniques that utilize speaker attributes, including voice and speaking style, to generate deepfake speech. This has also amplified global security risks from deepfake speech misuse, resulting in considerable societal costs worldwide. To address the security threats posed by deepfake speech, techniques have been developed focusing on both the protection of voice attributes and the defense against deepfake speech. Among them, the voice anonymization technique has been developed to protect voice attributes from extraction for deepfake generation, while deepfake detection and watermarking have been utilized to defend against the misuse of deepfake speech. This paper provides a short and concise overview of the three techniques, describing the methodologies, advancements, and challenges. A comprehensive version, offering additional discussions, will be published in the near future.**

I. Introduction

In the past years, the evolution of the Internet has enabled people to generate and share an increasing amount of speech data online. Simultaneously, advancements in computational power have significantly enhanced the development of deep neural networks (DNNs), providing effective tools for processing the information conveyed by big speech data, resulting in substantial progress in speech technology [1]–[3]. As a crucial component of the information conveyed by speech signals, modeling techniques for speaker attributes have gone through remarkable advancements, prompting the development of related speech techniques. For example, given a speech segment of only a few seconds, the speaker's identity can be recognized using speaker recognition techniques. Moreover, it can also be utilized to synthesize the speaker's speech using personalized speech generation techniques, generating deepfake speech of the speaker. Notably, the emergence and rapid development of large speech generation models [4]–[7] have greatly enhanced the fidelity and speaker similarity regarding voice and speaking style of the generated speech.

Given that voice and speaking style are unique to each individual, thereby conveying speaker privacy information, the misuse of deepfake speech poses significant security threats globally, as extensively reported in media sources. For instance, an individual's deepfake speech may be exploited for fraudulent activities or disseminated to harm her/his reputation. When public figures are targeted, deepfake speech becomes a powerful tool for manipulating public sentiment and opinions. In the big data era, the security challenges posed by deepfakes have drawn governmental attention worldwide, driving the formulation of regulatory frameworks, such as the General Data Protection Regulation in European Union [8], the Interim Regulations for the Management of Generative Artificial Intelligence Services in China [9], the Act on the Protection of Personal Information in Japan [10] and the Personal Data Protection Act in Singapore [11].

The security threats posed by deepfake speech have also attracted the attention of the research community, prompting the development of protection and defense techniques against it. Specifically, the protection techniques aim to protect speaker attributes from being extracted and exploited for deepfake generation, among which voice anonymization provides a viable solution [12]. The defense techniques are developed to prevent the use of deepfake speech for malicious purposes, wherein both deepfake detection [13] and watermarking [14] are among the viable techniques. Deepfake detection operates in a passive manner, without prior knowledge of the speech synthesis system; whereas watermarking can function proactively, as the inaudible watermark is embedded into the generated speech data before it is distributed to the public.

Fig. 1 illustrates an example for the applications of the three techniques for the security of speaker privacy against deepfake speech. In the absence of security techniques, leveraging the speaker attributes conveyed in the speech utterance of a speaker, deepfake speech can be generated by the attacker using personalized speech generation techniques that replicate the speaker's voice and mimic her/his speaking style. The deepfake speech can then be maliciously utilized to impersonate the speaker and attack automatic speaker verification (ASV) systems. To prevent the extraction of voice attributes for deepfake speech generation, voice anonymization techniques can be utilized to obscure or eliminate these attributes in the speaker's utterance, thereby preventing the synthesis of speech in their voice. Additionally, deepfake detection can be employed before the ASV system to examine whether the input speech is synthetic. Furthermore, by integrating watermarking techniques into the personalized speech generation process, deepfake speech can be marked as synthetic rather than genuine. Before the ASV system, the watermark can be detected, thereby further inhibiting its potential use for impersonation. This paper provides a concise overview of the three techniques, outlining the methodologies, advancements, and challenges.

This work was supported in part by the National Key Research and Development Program of China (Project No. 2024YFE0217200), the Innovation and Technology Fund of the Hong Kong SAR (Project No. MHP/048/24), and JST PRESTO (Project No. JPMJPR23P9).

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

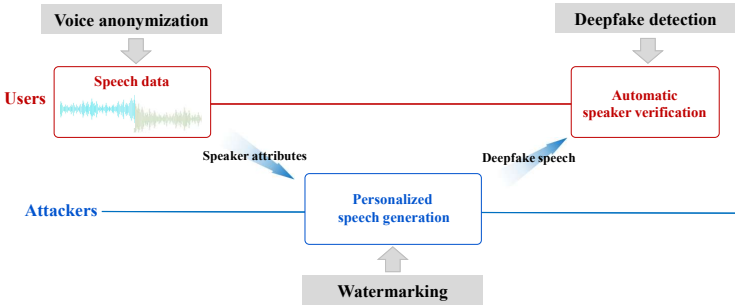

Fig. 1: Illustrative example of applications for voice anonymization, deepfake detection, and watermarking techniques, shown in the grey boxes, for the security of speaker privacy. Voice anonymization and deepfake detection are utilized at the user end, while watermarking is applied during the generation of deepfake speech.

II. Voice Anonymization

Voice anonymization is a technique that can be dated back to the 1980s, when speech signals were represented as analog signals [15]–[17]. At that time, the speech signals were modified to protect the speaker attributes within them, resulting in degraded speech quality. In recent years, advancements in neural networks and speech generation techniques have significantly enhanced the development of voice anonymization [18], [19]. As voice attributes are detectable by both human auditory systems and machine algorithms, voice anonymization can be realized in both synchronous [20]–[25] or asynchronous [26]–[32] manners. Synchronous anonymization protects voice attributes from being correctly perceived by both human hearing and machine algorithms, whereas asynchronous anonymization preserves human perception and solely protects against machine algorithm extraction.

The requirements for anonymized speech are defined from two aspects: voice privacy protection capability and utility [12], [33]. Voice privacy protection requires that the voice attributes can not be correctly extracted and utilized by machine algorithms. In terms of human perception, synchronous voice anonymization is required to alter the original voice attributes, while asynchronous voice anonymization needs to preserve them as perceived by human listeners. The utility depends on the applications and may require that anonymized speech preserves the non-voice attributes of the original speech, including the speech quality, linguistic content, and prosody, among others.

This section outlines the generative and adversarial approaches to voice anonymization, which currently attract the most attention in the research community. Specifically, the generative approach produces anonymized speech as synthesized output, applicable for both synchronous and asynchronous voice anonymization. The adversarial approach obtains anonymized speech by adding perturbations to the original signal, facilitating asynchronous voice anonymization.

A. Generative voice anonymization

Based on a speech generation framework that disentangles and represents speaker attributes as embedding vectors, voice anonymization can be accomplished by substituting the original speaker embedding with that of a pseudo-speaker. This mechanism facilitates generative solutions for both synchronous and asynchronous voice anonymization, wherein the anonymized output is synthesized speech. In synchronous voice anonymization, the pseudo-speaker embedding is obtained as a representation of a speaker that is distinct from the original speaker [19], [34], [35]. As reported in the VoicePrivacy Challenge 2024 [36], the voice anonymization approaches demonstrated effectiveness in protecting speaker privacy while preserving linguistic content. As voice anonymization focuses on protecting the voice attribute of speakers, the disentanglement of voice attribute from the others forms the crux of the generative methodology. Due to the correlation between voice and prosody attributes, both encoded in prosodic features like pitch and energy, preserving original prosody while preventing speaker attribute leakage through these features presents a significant challenge.

In asynchronous voice anonymization, limited by the requirement to preserve human perception of the original speaker, the pseudo-speaker embedding is obtained through constrained modifications of the original one. Existing studies have shown that machine recognition of original speaker attributes can be effectively obscured in asynchronously anonymized speech, with their human perception preserved [31], [32]. However, compared to synchronous voice anonymization, additional challenges exist including: 1) due to the lack of disentanglement of machine and human-perceived attributes within the speaker embedding, the modifications on speaker embedding have to reach trade-offs between the obscuration of machine perception with the preservation of human perception; 2) in scenarios where attackers can access the anonymization system to train speaker attribute extractors, the effectiveness in voice protection drops significantly.

B. Speaker-adversarial speech

The speaker-adversarial speech provides an alternative solution to asynchronous voice anonymization. As discovered in [37], neural network models were susceptible to adversarial perturbations in input samples [37], leading to investigations into adversarial attacks on speaker recognition models [38]–[40]. The ability of speaker-adversarial speech to deceive speaker recognition further enables its application in voice anonymization [26]–[30]. Existing studies have demonstrated that speaker-adversarial speech effectively prevents the accurate extraction of voice attributes by the speaker extractor that is used for adversarial perturbation generation. Besides, its efficacy in preserving the human perception of speaker attributes and the utility of the original speech has been validated. However, in its application in voice anonymization, the technique faces challenges including: 1) the limited transferability of the perturbation leads to a substantial reduction in its capability of preventing the voice attributes from being extracted by speaker extractors that were not involved in the adversarial perturbation generation process, 2) similar to the generative approach, speaker-adversarial speech loses its voice privacy protection effectiveness if an attacker gains access to its generation system and uses the adversarially generated speech to train a speaker attribute extractor.

Speaker privacy covers beyond voice to include speaking style, lexical preferences, and grammatical choices; therefore, voice anonymization alone is insufficient to fully protect speaker privacy, necessitating more complex approaches than these mentioned methods.

III. Deepfake detection

Deepfake detection processes an input utterance and yields an answer indicating whether the utterance was uttered by a human speaker or not. While the formulation is not flawless, deepfake detection as a binary classification task has received growing attention in the past decade [13], [41]. Its assumed application includes what Fig. 1 illustrates – the deepfake detector rejects any input trial that is unlikely to be a human speech before the trial is delivered to an ASV system. In that context, 'deepfake detection' was often referred to as speech anti-spoofing [42]. Another application is to protect the human ears, for example, by tagging the sound track of a video as deepfake on social media platforms. It is in the second application that we see an increased number of incidents caused by high-quality deepfake speech contents [43]. In the rest of this section, however, we use the term deepfake detection to cover both applications.

A. Progress of deepfake detection

Approaching deepfake detection as a binary classification task allows us to plug various modules into the machine learning pipeline, which accelerate the research and development iteration in the field. Feature engineering has advanced from purely using digital signal processing (DSP) algorithms [44] (e.g., linear-frequency cepstrum coefficients [45]) to the hybrid of DSP and deep learning methods (e.g., trainable filters integrated in a convolution network [46]). The latest trend is to extract features using pre-trained self-supervised-learning (SSL)-based models [47]. Potentially due to the SSL training based on a large-scale of human speech data, the SSL-based models seem to be able to extract features that better reveal the artifacts in the frequency band of speech sounds [48], [49].

On the classifiers, we have witnessed the paradigm shift from linear models (e.g., Gaussian mixture model) to various types of DNNs. Many of them are borrowed from computer vision or deep learning fields, e.g., light convolutional neural network (CNN) [50], ResNet [51], graph-based network [52], and the latest state-space models [53]. Combining the SSL-based feature extractor and the latest DNN-based classifier appears to be the state-of the-art paradigm [41].

The progress of deepfake detection is also supported by databases. Notable databases are those from the ASVspoof challenges [54] and audio deepfake detection challenges [55]. It is based on the shared databases (as well as evaluation protocols) that results from different papers are compared. Nowadays, many research groups are creating new databases that span different languages [56], newest fake speech generation methods [56], [57], more challenge acoustic conditions [58], all of which address new research questions.

B. Challenges and future directions

High-performing detectors can now attain significantly low equal error rates on evaluation datasets with limited complexity, such as ASVspoof 2019. From this paragraph, however, we dive into the potential limitations and issues.

1) Short-cut learning: Classifiers are prone to short-cut learning, and binary deepfake detectors are no exception. Specifically, a deepfake detector may overfit to the artifact that only presents in a particular training set but is irrelevant to the decision [59] (e.g., the length of the non-speech region [60]). Although the detector may perform well on a test set with a similar artifact, it cannot make useful decisions when the short-cut does not exist in a different test set. Even worse, the detector can be easily fooled by an attacker who adversarially uses the artifact to mislead the detector. Avoiding short-cut learning in model training is a good research direction. For studies not directly addressing the short-cut learning issue, we call for evaluation using multiple test sets from different data sources (e.g., those from the ASVspoof plus other non-ASVspoof test sets [61]). This avoids over-optimistic results caused by short-cut learning on a particular database.

2) Explainability: Modern detectors are predominantly black-box models that output decision scores without revealing the underlying reasoning. This does not help when evidence for the decision is required, for example, by the system user. Some studies try to interpret the detectors' behaviors using simpler models, but they are unlikely to be logically solid [62]. A few recent studies explored explainable models by design [63]. There is also an effort of using audio large language model to produce text-based explanation [64]. These are encouraging directions that are indispensable for decisions making with evidence.

3) Task definition: While a binary classification task is easy to work with, the definition of the two classes is equivocal. Following the convention, we may start by defining the negative class as 'being generated by a model given a text or a source voice prompt'. Then the positive class will be simply the logic negation of the negative. However, should a human-uttered utterance be treated as 'fake' if it is compressed using a DNN-based speech codec? Note that the DNN-based decoder in many speech codec share the same DNN architecture (e.g., HiFiGAN [65]) as many speech synthesis systems. As another example, should an anonymized utterance be tagged as fake?

What make things worse is that a binary deepfake detector is also expected to generalizable to any unseen attack. However, this goal may be untestable — we cannot claim a deepfake detector to be generalizable to *any* unseen fake speech unless we evaluate on *all* possible fake speech, but there will always be new fake speech created in the future that is not test.

A safer approach may be to define the two classes based on the applications, but the detector has to be rebuilt whenever the class definition is updated. Alternatively, we may shift from the binary classification task to different paradigms, e.g., spoofing-oriented verification [66] and multi-class source tracing [67], [68]. We may also hide information into speech data to signify the source (either human or a particular generator). Then the deepfake detection becomes a watermarking task that will be explained in the next section.

Practitioners working on standard databases may not bother with the above issues, but they are unavoidable when deploying deepfake detectors in real applications. Hence, we encourage the readers to keep an eye on potential issues.

IV. Watermarking

Watermarking involves embedding imperceptible messages into a data sample and extracting this message at a later stage [69]. The technique was initially applied to multimedia data for copyright protection, authentication, digital ownership management, among others [69], but it recently found application in proactive image and speech deepfake detection. For example, watermark can be embedded into synthetic speech [70], [71], indicating that the speech

Simpler models approximate the black-box detector's decision in a neighborhood of the feature space. If the approximation is sufficiently accurate, we should use the simpler models rather than the black-box detector; if the approximation is inaccurate, the interpretation does not make sense.

is generated rather than authentic, thereby providing a viable solution for preventing the misuse of both voice and speaking style information of the original speaker. While the requirements may vary among the applications, in most of cases watermark for speech data must be imperceptible and robust. An imperceptible watermark message should not degrade the perceptual quality of the carrier speech data. A robust watermark should persist even if the carrier speech signal is subjected to intentional manipulation or unintentional degradation, for example, MP3 compression and low-pass filtering.

A. Post-processing and collaborative approaches

Speech watermark using classical signal-processing-based methods have been extensively investigated before the deep learning and big data era [72]. Despite well-crafted algorithms and thorough theoretical analysis [73] (with certain assumptions on channel noise), the classical methods seem to be unable to produce watermark robust to many types of degradation [14], [71], [74], [75]. Hence, some recent studies started to investigate deep-learning-based speech watermark.

Popular designs can be roughly categorized into post-processing and collaborative types. The former takes speech data (either synthetic or real) as system input and add watermark in a post-processing manner [70], [71], [76]–[78]. Most of these methods use an encoder-decoder-like DNN — the watermark bit string is added to the output of the speech encoder, and the watermarked speech waveform is reconstructed by the decoder. There is also a method adding learnable perturbation to the input speech waveform [78]. The post-processing approaches can be easily applied to any speech synthesis system, and many of them demonstrated higher robustness than signal-processing-based ones [71], [75].

The collaborative approach focuses on watermarking the speech generation model [79]–[82]. The generation model and a watermark detector are jointly optimized so that the generated speech carries imperceptible latent information that can be extracted by the watermark detector (but not any other detector). This is different from the post-processing approach, wherein the generation model is independent from the watermarking algorithm. Although the joint optimization complicates the training procedure, the collaborative approach seems to be more robust against many degradation types than the post-processing methods [14].

B. Challenges and future directions

Unfortunately, watermark produced by the recent DNN-based algorithms are not sufficiently robust. Among these, processing with a vocoder or codec results in strong distortion of the watermark information [83]. The DNN-based codec is particularly 'effective' in removing the watermark [14], [84]. The lack of general robustness is problematic if the watermarked data is distorted by similar vocoder or codec during the transmission or under attacker. Besides, the defense against adversarial attacks on the watermark detectors remains an open question that awaits future investigation.

Even being sufficient robust against unintentional or intentional degradation, the current speech watermark algorithms do not guarantee security [85]. For example, in application where watermarks are added to real and synthetic speech data, an attacker may add another layer of conflicting watermark (e.g., adding a 'real' watermark to a synthetic utterance that already carries the 'fake' watermark). This collusion attack is beyond the issue of robustness against vocoder or codec and may not be easily fixed by optimizing the watermark encoder and decoder. Instead, protocols of embedding and authenticating the watermark may be necessary [85], [86].

The terms 'intentional' and 'unintentional' are defined from the attacker's perspective. For example, an attacker may intentionally use MP3 compression to destroy the watermark; degradation may also be unintentional if the watermarked data has to be compressed using MP3 when being transmitted.

V. Conclusions

This paper provides a concise overview of techniques employed to address the security threats caused by deepfake speech, including voice anonymization, deepfake detection, and watermarking. Their methodologies, advancements, and challenges are discussed. Moreover, potential challenges may arise in integrating the three techniques into a unified system, necessitating further investigation and not addressed in this paper.

References

[1] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[2] Y. Kumar, A. Koul, and C. Singh, "A deep learning approaches in text-to-speech system: A systematic review and recent research perspective," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 15 171–15 197, 2023.

[3] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[4] S. Chen, C. Wang, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705–718, 2025.

[5] Z. Du, Y. Wang, Q. Chen, *et al.*, "CosyVoice 2: Scalable streaming speech synthesis with large language models," *arXiv*, 2024. eprint: 2412.10117.

[6] Z. Jiang, Y. Ren, R. Li, *et al.*, "MegaTTS 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis," *arXiv*, 2025. eprint: 2502.18924.

[7] Z. Ju, Y. Wang, K. Shen, *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *Proc. International Conference on Machine Learning*, 2024.

[8] "General Data Protection Regulation," 2016. [Online]. Available: https://gdpr-info.eu/.

[9] "Interim Regulations for the Management of Generative Artificial Intelligence Services." [Online]. Available: https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

[10] "Act on the Protection of Personal Information," 2003. [Online]. Available: https://www.cas.go.jp/jp/seisaku/hourei/data/APPI.pdf.

[11] "Personal Data Protection Act 2012," 2012. [Online]. Available: https://sso.agc.gov.sg/Act/PDPA2012.

[12] N. Tomashenko, B. M. L. Srivastava, X. Wang, *et al.*, "Introducing the voiceprivacy initiative," *Proc. Interspeech*, 2020, pp. 1693–1697.

[13] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *arXiv*, 2023. eprint: 2308.14970.

[14] P. O'Reilly, Z. Jin, J. Su, and B. Pardo, "Deep audio watermarks are shallow: Limitations of post-hoc watermarking techniques for speech," *The 1st Workshop on GenAI Watermarking*, 2025.

[15] R. Cox, D. Bock, J. J. K. Bauer and, and J. Snyder, "The analog voice privacy system," *Proc. ICASSP*, 1986, pp. 341–344.

[16] N. S. Jayant, R. V. Cox, B. J. McDermott, and A. Quinn, "Analog scramblers for speech based on sequential permutations in time and frequency," *Bell System Technical Journal*, vol. 62, no. 1, pp. 25–46, 1983.

[17] R. V. Cox and J. M. Tribolet, "Analog voice privacy systems using TFSP scrambling: Full duplex and half duplex," *The Bell System Technical Journal*, vol. 62, no. 1, pp. 47–61, 1983.

[18] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, and et al, "Introducing the voiceprivacy initiative," *Proc. Interspeech*, 2020, pp. 1693–1697.

[19] F. Fang, X. Wang, J. Yamagishi, *et al.*, "Speaker anonymization using x-vector and neural waveform models," *Proc. SSW10*, 2019, pp. 155–160.

[20] T. Vaidya, M. Sherr, M. Todisco, A. Nautsch, and N. Evans, "You talk too much: Limiting privacy exposure via voice input," *IEEE Security and Privacy Workshops*, 2019.

[21] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Speech sanitizer: Speech content desensitization and voice anonymization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 6, pp. 2631–2642, 2021.

[22] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," *Proc. Interspeech*, 2021, pp. 1099–1103.

[23] J. Qian, H. Du, J. Hou, *et al.*, "Voicemask: Anonymize and sanitize voice input on mobile devices," *CoRR*, vol. abs/1711.11460, 2017.

[24] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," *Proc. the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 82–94.

[25] X. Miao, R. Tao, C. Zeng, and X. Wang, "A benchmark for multi-speaker anonymization," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 3819–3833, 2025.

[26] M. Chen, L. Lu, J. Wang, *et al.*, "VoiceCloak: Adversarial example enabled voice de-identification with balanced privacy and utility," 2, vol. 7, 2023, pp. 1–21.

[27] P. Cheng, Y. Wu, Y. Hong, *et al.*, "UniAP: Protecting speech privacy with non-targeted universal adversarial perturbations," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 1, pp. 31–46, 2024.

[28] S. Chen, L. Chen, J. Zhang, K. Lee, Z. Ling, and L. Dai, "Adversarial speech for voice privacy protection from personalized speech generation," *Proc. ICASSP*, 2024, pp. 11 411–11 415.

[29] Z. Zhang, Q. Yang, D. Wang, *et al.*, "Mitigating unauthorized speech synthesis for voice protection," *Proc. the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, 2024, pp. 13–24.

[30] L. Chen, C. Guo, R. Wang, K. A. Lee, and Z.-H. Ling, "Any-to-any speaker attribute perturbation for asynchronous voice anonymization," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 7736–7747, 2025.

[31] R. Wang, L. Chen, K. A. Lee, and Z.-H. Ling, "Asynchronous voice anonymization using adversarial perturbation on speaker embedding," *Proc. Interspeech*, 2024, pp. 4443–4447.

[32] R. Wang, L. Chen, K. A. Lee, and Z.-H. Ling, "Asynchronous voice anonymization by learning from speaker-adversarial speech," *IEEE Signal Processing Letters*, vol. 32, pp. 1905–1909, 2025.

[33] N. Tomashenko, X. Miao, P. Champion, *et al.*, "The VoicePrivacy 2024 Challenge evaluation plan," *arXiv*, 2024. eprint: 2404.02677.

[34] B. M. L. Srivastava, M. Maouche, M. Sahidullah, *et al.*, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2383–2395, 2022.

[35] L. Chen, W. Gu, K. A. Lee, W. Guo, and Z.-H. Ling, "Pseudo-speaker distribution learning in voice anonymization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 272–285, 2025.

[36] N. Tomashenko, "Overview of the voiceprivacy 2024 challenge," 2024. [Online]. Available: https : / / www . voiceprivacychallenge.org/vp2024/docs/VPC-2024-.pdf.

[37] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations*, 2014.

[38] X. Zhang, X. Zhang, M. Sun, X. Zou, K. Chen, and N. Yu, "Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition," *Complex & Intelligent Systems*, vol. 9, no. 1, pp. 65–79, 2023.

[39] Abdullah *et al.*, "Hear "No Evil", See "Kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems," *IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 712–729.

[40] X. Li *et al.*, "Adversarial attacks on GMM i-vector based speaker verification systems," *Proc. ICASSP*, 2020, pp. 6579–6583.

[41] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A survey on speech deepfake detection," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–38, 2025.

[42] Z. Wu, T. Kinnunen, N. Evans, *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," *Proc. Interspeech*, 2015, pp. 2037–2041.

[43] McAfee, *Beware the Artificial Impostor: A McAfee Cybersecurity Artificial Intelligence Report*, 2023.

[44] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: From the perspective of ASVspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, e2, 2020.

[45] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," *Proc. Interspeech*, 2015, pp. 2087–2091.

[46] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," *Proc. ICASSP*, 2020, pp. 6369–6373.

[47] A. Mohamed, H.-y. Lee, L. Borgholt, *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[48] X. Wang and J. Yamagishi, "Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures," *Proc. Odyssey*, 2022, pp. 100–106.

[49] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *Proc. Odyssey*, 2022, pp. 112–119.

[50] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the

ASVspoof2019 challenge," *Proc. Interspeech*, 2019, pp. 1033–1037.

[51] H. Zeinali, T. Stafylakis, G. Athanasopoulou, *et al.*, "Detecting spoofing attacks using VGG and SincNet: BUT-Omilia submission to ASVspoof 2019 challenge," *Proc. Interspeech*, 2019, pp. 1073–1077.

[52] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," *Proc. ICASSP*, 2022, pp. 6367–6371.

[53] Y. Xiao and R. K. Das, "XLSR-Mamba: A Dual-Column Bidirectional State Space Model for Spoofing Attack Detection," *IEEE Signal Processing Letters*, vol. 32, pp. 1276–1280, 2025.

[54] X. Wang, H. Delgado, H. Tak, *et al.*, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *ASVspoof Workshop 2024*, 2024, pp. 1–8.

[55] J. Yi, J. Tao, R. Fu, *et al.*, *ADD 2023: The second audio deepfake detection challenge*, 2023. eprint: 2305.13774.

[56] N. M. Müller, P. Kawa, W. H. Choong, *et al.*, "MLAAD: The Multi-Language Audio Anti-Spoofing Dataset," *Proc. IJCNN*, 2024, pp. 1–7.

[57] Y. Lu, Y. Xie, R. Fu, *et al.*, "Codecfake: An initial dataset for detecting llm-based deepfake audio," *Proc. Interspeech*, 2024, pp. 1390–1394.

[58] J.-w. Jung, Y. Wu, X. Wang, *et al.*, "SpoofCeleb: Speech Deepfake Detection and SASV In The Wild," *IEEE Open Journal of Signal Processing*, vol. 6, pp. 68–77, 2025.

[59] M. Sahidullah, H.-j. Shim, R. G. Hautamäki, and T. H. Kinnunen, "Shortcut Learning in Binary Classifier Black Boxes: Applications to Voice Anti-Spoofing and Biometrics," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–16, 2025.

[60] N. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is silver, silence is golden: What do ASVspoof-trained models really learn?" *Proc. ASVspoof challenge workshop*, 2021, pp. 55–60.

[61] S. Arena, *Speech arena: Speech deepfake leaderboard*, 2025. [Online]. Available: https://huggingface.co/spaces/Speech-Arena-2025/Speech-DF-Arena.

[62] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[63] J. Mishra, M. Chhibber, H.-j. Shim, and T. H. Kinnunen, "Towards explainable spoofed speech attribution and detection: A probabilistic approach for characterizing speech synthesizer components," *Computer Speech & Language*, vol. 95, p. 101 840, 2026.

[64] H. Gu, J. Yi, C. Wang, *et al.*, "ALLM4ADD: Unlocking the Capabilities of Audio Large Language Models for Audio Deepfake Detection," *Proc. ACMMM*, 2025.

[65] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[66] T. Chen and E. Khoury, "Spoofprint: A new paradigm for spoofing attacks detection," *Proc. SLT*, 2021, pp. 538–543.

[67] N. Klein, T. Chen, H. Tak, R. Casal, and E. Khoury, "Source tracing of audio deepfake systems," *Proc. Interspeech*, 2024.

[68] N. M. Müller, P. Kawa, S. Hu, *et al.*, "A new approach to voice authenticity," *Proc. Interspeech*, 2024, pp. 2245–2249.

[69] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," *Proc. IEEE International Conference on Image Processing*, vol. 3, 1996, pp. 243–246.

[70] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, "Detecting voice cloning attacks via timbre watermarking," *Proc. Network and Distributed System Security Symposium*, 2023.

[71] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, "Wavmark: Watermarking for audio generation," *arXiv*, 2023. eprint: 2308.12770.

[72] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. Thing, "Twenty years of digital audio watermarking—a comprehensive review," *Signal processing*, vol. 128, pp. 222–242, 2016.

[73] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1127–1141, 1999.

[74] M. Yamni, A. Daoui, H. Karmouni, *et al.*, "An efficient watermarking algorithm for digital audio data in security applications," *Scientific Reports*, vol. 13, no. 1, p. 18 432, 2023.

[75] Y. Wen, A. Innuganti, A. B. Ramos, H. Guo, and Q. Yan, *SoK: How robust is audio watermarking in generative AI models?* 2025. eprint: 2503.19176.

[76] P. O'Reilly, Z. Jin, J. Su, and B. Pardo, "Maskmark: Robust neuralwatermarking for real and synthetic speech," *ICASSP*, 2024, pp. 4650–4654.

[77] R. S. Roman, P. Fernandez, A. Défossez, T. Furon, T. Tran, and H. Elsahar, "Proactive detection of voice cloning with localized watermarking," *Proc. ICML*, 2024.

[78] S. Wu, J. Liu, Y. Huang, H. Guan, and S. Zhang, "An Audio Watermarking Algorithm Based on Adversarial Perturbation," *Applied Sciences*, vol. 14, no. 16, p. 6897, 2024, ISSN: 2076-3417.

[79] L. Juvela and X. Wang, "Collaborative watermarking for adversarial speech synthesis," *Proc. ICASSP*, 2024, pp. 11 231–11 235.

[80] L. Juvela and X. Wang, "Audio codec augmentation for robust collaborative watermarking of speech synthesis," *Proc. ICASSP*, 2025.

[81] X. Cheng, Y. Wang, C. Liu, D. Hu, and Z. Su, "HiFi-GANw: Watermarked speech synthesis via fine-tuning of HiFi-GAN," *IEEE Signal Processing Letters*, vol. 31, pp. 2440–2444, 2024.

[82] J. Zhou, J. Yi, T. Wang, *et al.*, "Traceablespeech: Towards proactively traceable text-to-speech with watermarking," *Proc. Interspeech*, 2024, pp. 2250–2254.

[83] H. Liu, M. Guo, Z. Jiang, L. Wang, and N. Gong, "Audiomarkbench: Benchmarking robustness of audio watermarking," *Advances in Neural Information Processing Systems*, vol. 37, pp. 52 241–52 265, 2024.

[84] Y. Özer, W. Choi, J. Serrà, M. K. Singh, W.-H. Liao, and Y. Mitsufuji, "A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs?" *Proc. Interspeech*, 2025.

[85] F. Cayre, C. Fontaine, and T. Furon, "Watermarking security: Theory and practice," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3976–3987, 2005.

[86] M. Steinebach and J. Dittmann, "Watermarking-based digital audio data authentication," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 10, p. 252 490, 2003.