ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447190

# ADVTTS: ADVERSARIAL TEXT-TO-SPEECH SYNTHESIS ATTACK ON SPEAKER IDENTIFICATION SYSTEMS

*Chu-Xiao Zuo*[†]*, Zhi-Jun Jia*[†]*, Wu-Jun Li*

National Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology, Nanjing University, China
{zuochuxiao, jiazhijun}@smail.nju.edu.cn, liwujun@nju.edu.cn

## ABSTRACT

Speaker identification (SI) systems have been widely employed in real-world applications. However, recent research has demonstrated that SI systems are vulnerable to two prevalent attacks even without providing feedback to the attacker: the transfer-based adversarial attack and the speech synthesis spoofing attack. The transfer-based adversarial attack faces the challenges of collecting natural speech for specific content and timbre. In contrast, the speech synthesis spoofing attack can synthesize speech for any content and timbre but can be detected by audio deepfake detectors (ADD). In this paper, we propose a novel method, called adversarial text-to-speech synthesis (AdvTTS), for attacking SI systems. AdvTTS combines the strengths of transfer-based adversarial attacks and speech synthesis spoofing attacks, by synthesizing transferable attack speech with local surrogate models. AdvTTS is the first attack method that can conduct both adversarial and spoofing attacks with any speech content and timbre. AdvTTS can deceive SI systems with high-quality speech while evading ADD detection. Experiments show that AdvTTS can outperform other baselines for spoofing attacks, and can outperform the baselines for adversarial attacks with the combination of projected gradient descent (PGD).

***Index Terms—*** Adversarial attack, spoofing attack, speaker identification, speech synthesis

## 1. INTRODUCTION

Speaker identification (SI) systems have been widely employed in real-world applications, including authentication for financial transactions and voice-controlled devices. Ensuring the reliability and robustness of SI systems is critical, necessitating a comprehensive evaluation against potential threats or attacks.

Recent research has demonstrated the vulnerability of SI systems against adversarial attack [1, 2] which is performed by adding imperceptible perturbations onto natural samples. The adversarial attack is designed for the goal of inducing a misalignment between human perception and models. Initially proposed for image tasks, various gradient-based adversarial attack methods [3, 4, 5] have been proposed under the white-box scenario. Later research extends these attacks to speech data for attacking the SI system [6, 7, 8]. However, in practical scenarios, it is typically infeasible to obtain the gradients of the target systems. Query-based adversarial attacks, such as FakeBob [9] and SMACK [10], address this challenge by estimating gradients through continuous queries of similarity scores from the SI system under the black-box scenario. Nevertheless, in some practical scenarios, e.g., attacking the commercial system Microsoft Azure, the similarity scores are inaccessible, and the risk of being detected also increases with high-frequency queries [11].

Transfer-based adversarial attacks, such as SUETA [12], generate transferable adversarial samples on a local surrogate model. Although feedback from the target system is unnecessary, the transferability of transfer-based adversarial attacks is limited by the discrepancy between the surrogate and target models. Another limitation of transfer-based adversarial attacks lies in the collection process of natural speech. More specifically, in scenario where the SI system is deployed with a speech recognition module, attackers may be required to provide the system with speech that has specific content and timbre (e.g., forensic identification). However, collecting corresponding natural speech with specific content and timbre may pose a challenge for the attacker.

In contrast, speech synthesis spoofing attacks have been empirically verified [13] for their ability to deceive SI systems by imitating the timbre of a target speaker. Recent advancement in zero-shot text-to-speech (TTS) synthesis models, such as YourTTS [14], has improved the risk of producing high-quality audio with any speech content and timbre. Nevertheless, the synthesized speech can be detected by audio deepfake detectors (ADD) [15].

In this paper, we propose a novel method, called adversarial text-to-speech synthesis (AdvTTS), for attacking SI systems. AdvTTS combines the strengths of transfer-based adversarial attacks and speech synthesis spoofing attacks, by synthesizing transferable attack speech with local surrogate models.

[†] Equal contribution.

The main contributions of AdvTTS are listed as follows:

- AdvTTS is the first attack method that can conduct both adversarial and spoofing attacks with any speech content and timbre.
- AdvTTS can deceive SI systems with high-quality speech while evading ADD detection.
- Experiments show that AdvTTS can outperform other baselines for spoofing attacks, and can outperform the baselines for adversarial attacks with the combination of projected gradient descent (PGD).

## 2. PRELIMINARIES

### 2.1. Speaker identification systems

This paper focuses on one of the fundamental tasks of the SI system: close-set identification (CSI). CSI identifies a speaker from an enrolled set without rejection. To formally define the problem, we denote an input speech utterance as $\boldsymbol{x}$, and a group of speakers enrolled in the SI system as $G$. The speakers in $G$ are numbered by $1, 2, ..., K$. An SI model $S$ is employed in the system to extract speaker embeddings. The model outputs the similarity scores, denoted as $S(\boldsymbol{x}) = (s_1, ..., s_K)$. Specifically, we use $[S(\boldsymbol{x})]_i = s_i$ to denote the similarity score between the speaker embedding of $\boldsymbol{x}$ and that of the $i$-th enrolled speaker. For the CSI task, the SI system $F$ returns the speaker with the highest similarity score: $F(\boldsymbol{x}) = \arg\max_{i \in G} [S(\boldsymbol{x})]_i$.

### 2.2. Threat model

**Attacker's goals.** This paper focuses on the targeted attack. In targeted attack, the attacker aims to craft an attack speech $\boldsymbol{x}'$ that does not belong to the authentic voice of a target speaker $y$, but will be classified by the SI system $F$ as the target speaker. Furthermore, the attacker also aims to prevent the attack speech from being detected by an ADD model $D$. The goal can be formulated as $F(\boldsymbol{x}') = y$ and $D(\boldsymbol{x}') < \theta$, where $\theta$ is a detection threshold and the detector will reject any input with an output score above $\theta$.

**Attacker's knowledge.** This paper focuses on the black-box scenario where feedback from the target SI and ADD models is unavailable. Similar to the condition of transfer-based adversarial attacks, attackers can leverage surrogate models and collect utterances of target speakers.

**Attacker's capabilities.** The crafted attack speech should preserve high acoustic quality. We achieve this by applying a state-of-the-art (SOTA) zero-shot TTS model.

**Attack type.** Depending on whether the timbre of input speech is the same as a target speaker, the targeted attack on SI systems can be further divided into two types:

- Adversarial attack. The timbre of the attack speech is different from that of the target speaker. The attack speech will cause a misalignment between human perception and the SI system.
- Spoofing attack. The timbre of the attack speech resembles that of the target speaker.

Note that the traditional adversarial attacks on the SI system are not designed with the goal of deceiving the ADD. This is because the traditional adversarial attacks are conducted only on natural speech. In this paper, we show that an adversarial attack can also be conducted in the sampling process of a TTS model. Therefore, the adversarial attack is also required to deceive the ADD.

### 2.3. Transfer-based adversarial attack

Transfer-based adversarial attacks utilize white-box methods to generate transferable adversarial samples on a local surrogate model. In this paper, we leverage the classical white-box adversarial attack method, projected gradient descent (PGD) [4], to attack a local surrogate SI model in the sampling process of our method. PGD iteratively perturbs a sample based on the gradient of a loss function $\mathcal{L}$. In each iteration, PGD projects the perturbation to fall in an $\epsilon$-ball. The attack step at iteration $i$ is formally defined as:

$$\boldsymbol{x}^i = \prod_{B(\boldsymbol{x}^0, \epsilon)} (\boldsymbol{x}^{i-1} + \eta \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}^{i-1}} \mathcal{L}(\boldsymbol{x}^{i-1}, y))), \quad (1)$$

where $B(\boldsymbol{x}^0, \epsilon)$ is the $\epsilon$-ball centered at the natural sample $\boldsymbol{x}^0$, $y$ is the label of $\boldsymbol{x}^0$, $\eta$ is the step size, and $\prod$ is the projection operator.

## 3. METHOD

The zero-shot TTS model can synthesize speech of any content and timbre using only one reference speech segment of the target timbre. We leverage a pre-trained SOTA zero-shot TTS model, YourTTS [14], as the backbone model to synthesize attack speech. The YourTTS model can be simplified as a pipeline through five modules, including an input encoder ($En$), an alignment model ($A$), a flow model ($Fl$), a posterior encoder ($Pe$), and a Hifi-GAN [16] based decoder ($De$). Given a reference speech $\boldsymbol{x}$ and any speech content $\boldsymbol{c}$, the yourTTS model synthesizes speech in the following sampling process:

$$\begin{aligned} &\boldsymbol{e}_t, \boldsymbol{e}_s = En(\boldsymbol{c}, \boldsymbol{x}), \quad \boldsymbol{z}_a = A(\boldsymbol{e}_t, \boldsymbol{e}_s, \boldsymbol{z}_r), \\ &\boldsymbol{z}_f = Fl(\boldsymbol{z}_a, \boldsymbol{e}_s), \quad \boldsymbol{x}' = De(\boldsymbol{z}_f, \boldsymbol{e}_s), \end{aligned} \quad (2)$$

where $\boldsymbol{e}_t$ is the text embedding, $\boldsymbol{e}_s$ is the reference speaker embedding, $\boldsymbol{z}_r$ is a random noise vector, $\boldsymbol{z}_a$ and $\boldsymbol{z}_f$ are the intermediate features, and $\boldsymbol{x}'$ is the synthesized speech. The module $Pe$, used to align the distribution of feature $\boldsymbol{z}_f$ with

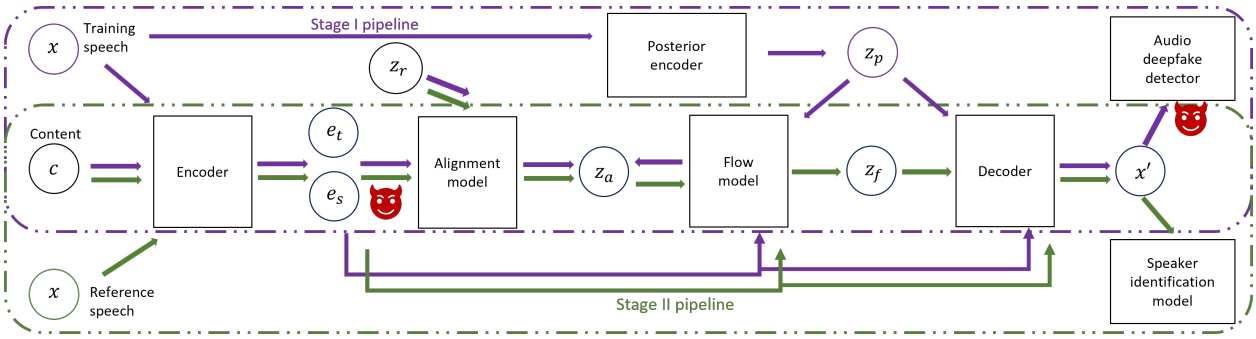


**Fig. 1**. The architecture of AdvTTS based on the YourTTS structure.

that of the learned features $\boldsymbol{z}_p$ for natural speech, is only employed in the training process.

Based on yourTTS model, we propose AdvTTS as a two-stage attack method. In the first stage, we perform adversarial fine-tuning on the yourTTS model with the goal of deceiving a surrogate ADD model. In the second stage, we perturb the high-level feature $\boldsymbol{e}_s$ during the sampling process, with the goal of deceiving a surrogate SI model. The whole architecture of AdvTTS is illustrated in Fig. 1.

### 3.1. Stage I: adversarial fine-tuning on the TTS model

The goal of deceiving the ADD model $D$ is independent of specific speakers. Attackers possess the capability to tune their TTS model, enabling all generated outputs to evade ADD detection. Consequently, rather than performing an adversarial attack against the ADD model in the sampling process, we propose to fine-tune the entire yourTTS model. Once we have obtained a model capable of evading ADD detection, we can proceed to reach the goal of deceiving the SI system. The adversarial fine-tuning is performed with the following loss function: $\mathcal{L}^{Adv} = \mathcal{L}^{Train} + \alpha \cdot CE(D(\boldsymbol{x}), 0)$, where $\mathcal{L}^{Train}$ is the training loss for YourTTS model [14], $CE$ is the cross-entropy loss function, $D(\boldsymbol{x})$ is the output score of the ADD detector for input $\boldsymbol{x}$ in the training set, and $\alpha$ is a hyper-parameter for tuning the adversarial strength.

### 3.2. Stage II: perturbing the high-level feature

In the sampling process, we perturb the reference speaker embedding $\boldsymbol{e}_s$, which is directly related to the identity information of the synthesized speech. Since the alignment model $A$ and random noise $\boldsymbol{z}_r$ are used for increasing text-to-speech alignment and diversity, we fix $\boldsymbol{z}_r$ and keep the output of $A$ fixed as $\boldsymbol{z}_a^0$ after initialization. We then conduct a PGD-based attack against a local surrogate SI model $S$ with the following loss function :

$$\mathcal{L}^{SI}(\boldsymbol{e}_s, y) = -\max(-[S(De(Fl(\boldsymbol{z}_a^0, \boldsymbol{e}_s), \boldsymbol{e}_s))]_y + \kappa,\ 0) \tag{3}$$

**Algorithm 1** Stage II of AdvTTS

**Input:**
Reference speech $\boldsymbol{x}$, speech content $\boldsymbol{c}$, target speaker $y$; perturbation budget $\epsilon$, number of iterations $N$, step size $\eta$, momentum factor $\beta$.

**Output:**
Attack speech utterance $\boldsymbol{x}'$;

$\boldsymbol{e}_t^0, \boldsymbol{e}_s^0 \leftarrow En(\boldsymbol{c}, \boldsymbol{x})$;
Sample random noise $\boldsymbol{z}_r$;
Initialize $\boldsymbol{z}_a^0 \leftarrow A(\boldsymbol{e}_t^0, \boldsymbol{e}_s^0, \boldsymbol{z}_r), \boldsymbol{g}^0 \leftarrow 0$;
**for** $i = 0$ to $N - 1$ **do**
  $\boldsymbol{g}^{i+1} \leftarrow \beta \cdot \boldsymbol{g}^i + \nabla_{\boldsymbol{e}_s^i} \mathcal{L}^{SI}(\boldsymbol{e}_s^i, y)$;
  $\boldsymbol{e}_s^{i+1} \leftarrow \prod_{B(\boldsymbol{e}_s^0, \epsilon)} (\boldsymbol{e}_s^{i+1} + \eta \cdot \text{sign}(\boldsymbol{g}^{i+1}))$;
**end for**
$\boldsymbol{x}' \leftarrow De(Fl(\boldsymbol{z}_a^0, \boldsymbol{e}_s^N), \boldsymbol{e}_s^N)$;
**return** $\boldsymbol{x}'$

where $\kappa$ is a confidence parameter. To enhance the transferability, we also adopt the momentum optimization strategy following the settings in [12]. The algorithm of stage II is illustrated in Algorithm 1.

To conduct the adversarial attack, we collect speech from a different speaker $\tilde{y} \neq y$ as the reference speech. We perform PGD on $\boldsymbol{e}_s$ without imposing perturbation constraints. As observed in our experiment, the output timbre could be different from the reference speech, but not identical to that of the target speaker. To conduct the spoofing attack, we collect speech from the target speaker as the reference speech. We also perform PGD on $\boldsymbol{e}_s$ with a perturbation budget $\epsilon$ to prevent significant variation in the speech timbre. Furthermore, the speech synthesized by AdvTTS can also be integrated with PGD by adding perturbation onto the output wave.

## 4. EXPERIMENT

### 4.1. Experimental setting

**Dataset.** We use librispeech [17] to train and test the SI systems. The dataset is split into training, test, enrollment, and

**Table 1**. The detection results (in percent) of the white-box RawNet2 model and the black-box LFCC-LCNN model.

| Spoofing attacks | White-box | | Black-box | |
|---|---|---|---|---|
| | DR | EER | DR | EER |
| TTS | 87 | 12.85 | 72.44 | 27.51 |
| AdvTTS stage I | **14.22** | **85.66** | 49.88 | 50.08 |
| AdvTTS | 15.71 | 84.24 | **49.07** | **50.85** |

**Table 2**. The ASR (in percent) of AdvTTS transferring across different SI models for spoofing attack. The numbers followed by * represent the ASR of local white-box attacks.

| Source/Target | X-Vector | D-TDNN | ECAPA |
|---|---|---|---|
| AdvTTS stage I | 49.44 | 52.56 | 64.78 |
| X-Vector | 70* | 64.11 | 77.33 |
| D-TDNN | 62.89 | 70.89* | **77.56** |
| ECAPA | **64.44** | **65.78** | 86.56* |

impostor subsets following the procedure in [9]. The adversarial fine-tuning is operated on the libriTTS and VCTK [18].

**Model structure.** We select three SOTA SI models: X-Vector [19], D-TDNN [20], and ECAPA-TDNN (ECAPA) [21]. The models are trained with the AAM-Softmax loss function [20, 22]. For the ADD model, we use the two pretrained baseline models of the DF track in the ASVSpoof 2021 [15], including a RawNet2 used as the white-box surrogate detector and an LFCC-LCNN used as the black-box detector.

**Attack setting.** For the adversarial fine-tuning in stage I, we employ the AdamW optimizer with betas set to 0.8 and 0.99 and weight decay set to 0.01. The learning rate is set to 2e-5 and decays exponentially with a gamma factor of 0.999875. For perturbing the high-level feature in stage II, we set $\epsilon = 0.005$, $\eta = \epsilon/4$, and $N = 20$. We set $\kappa = 1$ in Eq. (3). We randomly collect 10 speech utterances from the target speaker as a substitution for the unknown enrolled speech.

**Evaluation metrics and baselines.** We employ the attack success rate (ASR) to evaluate the effectiveness of attacks. For evaluating the detection results of the ADD, we employ the detection rate (DR) and equal error rate (EER). The threshold is set according to EER, i.e., $\text{DR} \approx 1 - \text{EER}$. The strength of attack against the defense of the detector is evaluated by ASR under detection (AUD), defined as $\text{AUD} = \text{ASR} \times (1 - \text{DR})$. AUD is the most important metric for measuring the attack performance. We also employ the NISQA [23] to evaluate the speech quality.

### 4.2. Results

Table 1 presents the detection results for speech synthesized by yourTTS, the model that only undergoes a fine-tuning process in stage I (denoted as AdvTTS stage I), and AdvTTS. Compared to yourTTS, AdvTTS achieves significantly lower DR and higher EER on both the white-box RawNet2 detector and the black-box LFCC-LCNN detector. The EER of the black-box detector approximately approaches 50%, indicating difficulty in distinguishing natural speech from AdvTTS-generated speech and showing the effectiveness of AdvTTS in evading ADD detection. In Table 2, we show the ASR of spoofing attacks conducted on different surrogate SI models. The results show that the speech generated by AdvTTS can transfer across different SI models. Furthermore, the stage II of AdvTTS can enhance the transferability beyond stage I. Table 3 presents the results of spoofing attacks averaged on different surrogate and target models. AdvTTS exhibits significant superiority over yourTTS except for the ASR. Particularly, stage I helps AdvTTS evade the ADD detection but harms the ASR by up to 16% compared to yourTTS. Furthermore, AdvTTS achieves the highest AUD which reflects the performance to achieve the ultimate goal of the targeted attack.

**Table 3**. The performance (in percent, except for NISQA) of spoofing attacks under the black-box scenario.

| | AUD | ASR | DR | EER | NISQA |
|---|---|---|---|---|---|
| Natural speech | * | * | * | * | 3.62 |
| yourTTS | 20.03 | **72.67** | 72.44 | 27.51 | 3 |
| AdvTTS stage 1 | 27.86 | 55.59 | 49.89 | 50.08 | 3.05 |
| AdvTTS | **34.98** | 68.69 | **49.07** | **50.85** | **3.22** |

**Table 4**. The performance (in percent, except for NISQA) of adversarial attacks under the black-box scenario.

| | AUD | ASR | DR | EER | NISQA |
|---|---|---|---|---|---|
| Natural speech | * | * | * | * | 3.62 |
| PGD | 16.91 | 35.37 | **52.19** | **47.78** | **3.51** |
| AdvTTS | 14.01 | 30.98 | 54.77 | 45.19 | 2.87 |
| AdvTTS + PGD | **17.21** | **38.11** | 54.84 | 45.09 | 2.87 |

For adversarial attacks, although PGD cannot generate speech with any speech content and timbre, we still conduct an experimental comparison between AdvTTS and PGD on the speech contents and timbres that appear in the test set, which is actually unfair for AdvTTS. The results in Table 4 show that PGD outperforms AdvTTS in AUD, but the combination of AdvTTS with PGD can achieve higher AUD compared to PGD.

## 5. CONCLUSION

In this paper, we propose AdvTTS, the first attack method that can conduct both adversarial and spoofing attacks with any speech content and timbre. AdvTTS can deceive SI systems with high-quality speech while evading ADD detection.

**Acknowledgements** This work is supported by NSFC Project (No.62192783, No.12326615) and Fundamental Research Funds for the Central Universities (No.020214380108). Wu-Jun Li is the corresponding author.

## 6. REFERENCES


[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[3] Nicholas Carlini and David A. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE S&P*, 2017, pp. 39–57.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[5] Francesco Croce and Matthias Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *ICML*, 2020, vol. 119, pp. 2206–2216.

[6] Felix Kreuk, Yossi Adi, Moustapha Cissé, and Joseph Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *ICASSP*, 2018, pp. 1962–1966.

[7] Qing Wang, Pengcheng Guo, and Lei Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," in *INTERSPEECH*, 2020, pp. 4228–4232.

[8] Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Feng Wang, and Jiashui Wang, "Towards understanding and mitigating audio adversarial examples for speaker recognition," *IEEE TDSC*, pp. 1–17, 2022.

[9] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *IEEE S&P*, 2021, pp. 694–711.

[10] Zhiyuan Yu, Yuanhaur Chang, Ning Zhang, and Chaowei Xiao, "SMACK: semantically meaningful adversarial audio attack," in *USENIX Security*, 2023.

[11] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y. Zhao, "Blacklight: Defending black-box adversarial attacks on deep neural networks," *CoRR*, vol. abs/2006.14042, 2020.

[12] Chu-Xiao Zuo, Jia-Yi Leng, and Wu-Jun Li, "Speaker-specific utterance ensemble based transfer attack on speaker identification," in *INTERSPEECH*, 2022, pp. 3203–3207.

[13] Phillip L. De Leon, Vijendra Raj Apsingekar, Michael Pucher, and Junichi Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *ICASSP*, 2010, pp. 1798–1801.

[14] Edresson Casanova, Julian Weber, Christopher Dane Shulby, Arnaldo Cândido Júnior, Eren Gölge, and Moacir A. Ponti, "Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *ICML*, 2022, vol. 162, pp. 2709–2720.

[15] Xuechen Liu, Xin Wang, Md. Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas W. D. Evans, Andreas Nautsch, and Kong Aik Lee, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE TASLP.*, vol. 31, pp. 2507–2522.

[16] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[18] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[19] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[20] Ya-Qi Yu and Wu-Jun Li, "Densely connected time delay neural network for speaker verification," in *INTERSPEECH*, 2020, pp. 921–925.

[21] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *INTERSPEECH*, 2020, pp. 3830–3834.

[22] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[23] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *INTERSPEECH*, 2021, pp. 2127–2131.