SiFMimicEvader: Evading Fake Voice Detection with Adversarial Neural Mimicry Attacks

Xuan Hai
Xin Liu*
Zihao Zhang
Ziyao Yu
Xiangzhen Kong
haix2024@lzu.edu.cn
bird@lzu.edu.cn
zhzihao2023@lzu.edu.cn
yuzy2024@lzu.edu.cn
kongxzh2024@lzu.edu.cn
Lanzhou University
Lanzhou, China

Song Li
The State Key Laboratory of Blockchain and Data Security, Zhejiang University
Hangzhou High-Tech Zone (Bin jiang) Institute of Blockchain and Data Security
Hangzhou, China
songl@zju.edu.cn

Weina Niu
University of Electronic Science and Technology of China
Chengdu, China
vinusniu@uestc.edu.cn

Rui Zhou
Qingguo Zhou
zr@lzu.edu.cn
zhouqg@lzu.edu.cn
Lanzhou University
Lanzhou, China

Abstract

The application of deep learning in voice cloning has significantly enhanced the quality of cloned voices. While advanced voice cloning technologies are widely applied across various domains, they also pose serious security challenges such as producing natural Deepfakes. In response, numerous studies have focused on detecting fake voices, with many reporting outstanding performance. However, is the issue truly resolved? This paper introduces Adversarial Neural Mimicry Attack (ANMA) which leverages a specialized model to predict the behavior of other similar models, transforming black-box attacks into white-box scenarios indirectly. Based on ANMA and Speaker-irrelative Features (SiFs), we propose a novel black-box attack framework called SiFMimicEvader, designed to evade fake voice detectors with high success rates and minimal query requirements. The framework utilizes speech representation models as the breakthrough to predict the behaviors of fake voice detectors and employs a series of SiFs editing operations as perturbations to deceive these detectors. Experimental results demonstrate the effectiveness of SiFMimicEvader, achieving an average attack success rate exceeding 50% across various detectors, significantly outperforming other attack methods, while also showing great performance in audio quality and query scale, indicating its high availability in real-world scenarios.

CCS Concepts

• Information systems → Multimedia information systems; **• Security and privacy → Social network security and privacy**.

[*] Corresponding authors

MM '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755595

Keywords

AI-Synthesized voice, DeepFake, AI-synthesized voice detection, Adversarial attack

ACM Reference Format:
Xuan Hai, Xin Liu, Zihao Zhang, Ziyao Yu, Xiangzhen Kong, Song Li, Weina Niu, Rui Zhou, and Qingguo Zhou. 2025. SiFMimicEvader: Evading Fake Voice Detection with Adversarial Neural Mimicry Attacks. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755595

1 Introduction

Voice has long been a fundamental medium for information exchange in human society, serving critical functions in digital systems, including real-time communication, identity authentication, etc [7, 39, 47]. The development of high-quality voice cloning technology has enabled its broad use across various fields. However, the rapid spread of voice cloning technology also introduces substantial security challenges, particularly by harming content integrity and trust-building processes. With the in-depth application of deep learning technology, voice cloning has greatly improved in both quality and naturalness. These advanced systems [32, 33, 37, 40, 41, 44] can clone vocal timbre, intonation, and prosody with remarkable precision, making it increasingly difficult for listeners to distinguish between human and cloned fake voices. Numerous real-world cases [19, 31] have shown that advanced voice cloning has been widely used to commit fraud, bypass voice authentication systems, and spread political disinformation.

To address these risks, relevant researchers have conducted a large number of studies on fake voice detection technology in recent years. Fake voice detection is a binary classification task aimed at determining whether audio is AI-generated or genuine. Early studies primarily focus on traditional speech features, such as Mel-Frequency Cepstral Coefficients (MFCC), to identify fake voices [2, 14]. In recent years, end-to-end (E2E) approaches have become increasingly popular which leverage deep neural network (DNN) models to automatically extract speech features and directly differentiate between real and synthetic speech. Many of these studies claim that their proposed methods achieve high effectiveness and deliver excellent performance in evaluation experiments. Some even report Equal-Error Rates (EER) below 1%. It indicates that existing fake voice detectors can accurately identify fake voices, offering robust technical support for mitigating the risks associated with voice cloning abuse. However, the actual situation may not be as optimistic as it appears. Recent research [28] introduces the concept of speaker-irrelative features (SiFs), referring to non-speaker-related information such as current noise, background noise, etc. The research highlights that existing fake voice detectors rely heavily on SiFs, leading to significant robustness issues. Furthermore, another study [16] shows that removing specific SiFs from input audio leads to a marked decline in the performance of these detectors.

These studies indicate that existing fake voice detectors pay more attention to SiFs and fail to capture the essential difference between fake and human voices, which makes it theoretically possible to deceive them by leveraging SiFs. The study [17] proposes an attack framework based on SiFs and traditional adversarial attack algorithms, achieving a high attack success rate. However, this approach requires a large number of queries to the target detector, and the generation of attack samples is time-intensive, significantly restricting its practical usability. To address the limitations of existing solutions, this paper proposes a novel attack design. Our approach is inspired by a neuroscience study using a brain-like neural network model to predict monkey brain behavior. Similarly, we aim to design a model to predict fake voice detector behavior, transforming black-box attacks into white-box attacks. We observe that recent advanced fake voice detectors increasingly rely on pre-trained speech representation models as feature extractors. These models typically have a large number of parameters and play a critical role in influencing detector decisions. Theoretically, by designing a fake voice detector that leverages these speech representation models, we can predict the behavior of other detectors utilizing similar models.

Based on this observation and SiFs, we propose a novel black-box attack framework, SiFMimicEvader, targeting fake voice detectors. This framework achieves a high success rate in black-box attack scenarios, requiring few queries or, in some cases, none at all. Specifically, SiFMimicEvader uses a generation model to create noise perturbations and incorporates a fake voice detector with a pre-trained speech representation model. This detector mimics the behavior of other detectors and functions as a discriminator to guide the generation model's training. Furthermore, we introduce a series of SiFs editing operations to enhance the attack's effectiveness.

The main contributions of this paper are summarized as follows: 1) We analyze the emerging trends in fake voice detection and **propose utilizing speech representation models, which are widely used in this field, to predict the behavior of fake voice detectors,** thereby indirectly transforming a black-box attack into a white-box attack; 2) We propose SiFMimicEvader, a new black-box attack framework designed to evade fake voice detection, which leverages multiple SiFs as attack features and enables real-time generation of attack samples with minimal query overhead; 3) We conduct a series of experiments to evaluate the success rate, attack sample quality, and query efficiency of our framework, and the results show that our method significantly outperforms existing approaches.

2 Related Works

2.1 Fake Voice Detection

Early synthetic voice detection methods focused on traditional features like spectral differences. With deep learning advancements, mainstream methods now use deep neural networks, categorized into four types: traditional feature-based methods, computer vision (CV)-based methods, E2E-based methods, and other novel methods.

Traditional feature-based methods: These methods convert raw audio into traditional audio features like spectrum and design back-end models for classification. In 2019, Alzantot et al. [2] proposed a detection method based on ResNet by combining multiple traditional features. Li et al. [25] improved ResNet with Res2Net for multi-scale feature learning, enhancing generalization against unknown attacks. In 2021, Gao et al. [12] extracted artifacts on log-Mel spectrograms using 2D discrete cosine transform (DCT), forcing back-end networks to learn advanced representations from long-term modulation patterns of audio inputs.

CV-based methods: These approaches convert audio features into images and borrow deep-learning models from image processing for detection. In 2019, Farid et al. [1] converted bispectral analysis features into images and classified them using SVM, which is the first CV-based approach. In 2021, Ballesteros et al. [3] proposed Deep4SNet, which uses histograms to represent voice data distributions and employs a Convolutional Neural Network (CNN)-based back-end model for classification.

E2E-based methods: These approaches process raw audio directly for detection without requiring additional feature extraction and are the most widely used approaches in recent years. In 2020, Tak et al. [35] introduced RawNet2, an end-to-end AI-synthetic audio detection system using sinc convolutions and residual blocks. In 2021, they designed RawGAT-ST [34], a model based on graph attention networks. In 2023, Ding et al. [11] proposed SAMO, which employed multi-center clustering to detect synthetic audio. Guo et al. [15] utilized the pre-trained audio representation model WavLM for AI-synthetic audio detection. In 2024, Zhang et al. [45] proposed an end-to-end detection model including a sensitive layer selection (SLS) module based on the pre-trained model XLS-R.

Other methods: Some detection methods adopt unconventional approaches. In 2020, Wang et al. [43] developed DeepSonar, a system that employs neuronal activity from an AI-driven speaker recognition system, considering it as a crucial feature for detecting synthetic audio. In 2022, Blue et al. [5] modeled audio from an articulatory phonetics perspective for detection. However, the detection time of this approach is notably high, making it challenging to implement in real-world scenarios.

2.2 Attacks on Audio Security Systems

In general, audio security systems can be divided into two types: speaker verification systems and fake voice detection systems. The former is primarily used to determine whether the speaker's identity matches expectations, while the latter is focused on detecting whether a piece of audio has been artificially synthesized. These two systems can both function independently and work together to form a high-security voice authentication system.

Speaker verification systems attacks: Replay attacks, once a classic method to target speaker verification systems, are ineffective against dynamic passphrases and have largely faded in practice. Researchers have since turned to adversarial generation techniques. Kreuk et al. [23] used the Fast Gradient Sign Method (FGSM) to perform both white-box and black-box attacks on speaker verification systems, demonstrating the capability of adversarial attacks to deceive such systems. Li et al. [26] extended FGSM-based attacks by enhancing the transferability of adversarial samples to other speaker verification systems. Tian et al. [38] proposed a black-box attack method for speaker verification systems within a feedback-controlled VC framework. With advancements in AI voice cloning, attackers now focus on generating synthetic audio to bypass verification systems, as discussed earlier.

Synthetic audio detection attacks: Before the rise of voice cloning technology, voice-based spoofing was costly and ineffective, leading to late-starting research on attacks targeting synthetic audio detection systems, with attack theories and methods being more traditional. In 2019, Liu et al. [27] proposed attack schemes based on FGSM and Projected Gradient Descent (PGD). In 2022, Liu et al. [29] first introduced a speaker-irrelative feature-based evasion theory at Black Hat USA, presenting a proof-of-concept framework. In 2023, Kassis et al. [22] successfully bypassed commercial security-critical voice authentication systems by deploying both AI synthetic audio detection and speaker verification systems by utilizing multiple SiFs during voice synthesis. Liu et al. [17] developed a targeted evasion method that can defeat any black-box fake voice detection system by combining fuzz testing with speaker-irrelative feature theory. In 2024, Zuo et al. [49] introduced an adversarial attack method based on TTS technology, which is the first to perform both adversarial and spoofing attacks using any speech content and timbre.

3 Motivation and Insight

Recent research shows high accuracy in fake voice detection under ideal conditions, proving detectors can effectively distinguish human voices from fake voices. However, several recent research studies have challenged this perspective. A recent study [28] introduces the concept of SiFs and highlights that existing fake voice detectors tend to overfit to SiFs, resulting in significant robustness issues. Another study [16] shows that removing specific SiFs led to a notable decline in the performance of existing fake voice detectors. This suggests that existing fake voice detectors rely heavily on SiFs rather than capturing the fundamental distinctions between fake and genuine voices.

In theory, if fake voice detectors rely on SiFs to distinguish fake voices, they can be deceived by altering the SiFs in the fake voices. Currently, most attacks on AI systems are based on traditional adversarial attacks. However, this approach has several notable limitations: **1) Query limitation:** Real-world fake voice detectors often restrict the number of queries per user within a set timeframe. Traditional adversarial attacks, which rely on multi-round iterative optimization and numerous queries, are thus limited in effectiveness. **2) Real-time limitation:** Many scenarios, like voice calls or online conferences, demand real-time attacks. However, most traditional adversarial methods require extensive iterations taking several minutes, making them impractical for such settings. **3) Universality limitation:** Existing attack methods typically need optimization for specific detectors or samples. Attackers seeking minimal interaction to maintain stealth and reduce costs find these methods inadequate, as they lack broad applicability.

In 2019, a neurological study [4] demonstrated that DNN models structured similarly to the brain could predict brain activities, implying that designing DNNs resembling the target model may also enable activity prediction. Inspired by this study, we propose a new attack methodology called adversarial neural mimicry attack (ANMA) to deceive the fake voice detectors. ANMA utilizes the speech representation models which are widely used in fake voice detectors in recent years as the breakthrough to mimic the fake voice detectors. Speech representation models are typically trained using unsupervised learning on large-scale datasets to extract meaningful representations of speech. These models are widely used in fake voice detectors due to their strong feature extraction capabilities. Most of these models are open-source, trained on the same public datasets, and have similar structures. As a result, their outputs are often comparable, meaning that using a different model can still produce similar results.

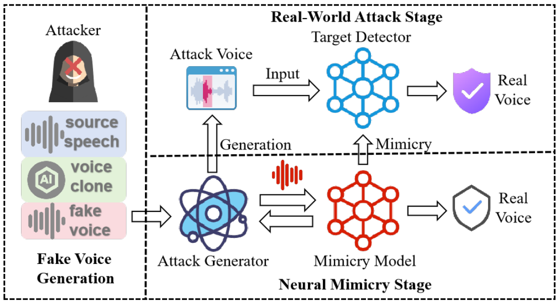


Figure 1: The ANMA attack process

The attack process of ANMA is illustrated in Figure 1. The mimicry model is a fake voice detector that uses a speech representation model as its feature extractor. During the neural mimicry stage, the attacker creates attack samples by mixing fake voice perturbations generated by the attack generator. The generator is optimized until it deceives the mimicry model, after which it is used to attack the target detector. Using ANMA, the attacker can overcome the limitations outlined above. For the query limitation, minimal interaction during optimization reduces the need for numerous queries. To address real-time constraints, a GAN-based attack generator can produce samples instantly. For universality, the mimicry model can theoretically replicate detectors using speech representation models, as these models yield similar outputs and significantly influence detector behavior. Additionally, for detectors without speech representation models, the mimicry model can still partially replicate their behavior. This is because both types of detectors perform the same task, and the speech representation model extracts richer feature information, often encompassing the features emphasized by other models.

4 Methodology

4.1 Overview

In this paper, we propose a novel attack framework, SiFMimicEvader, designed to deceive fake voice detectors. 'SiF' stands for

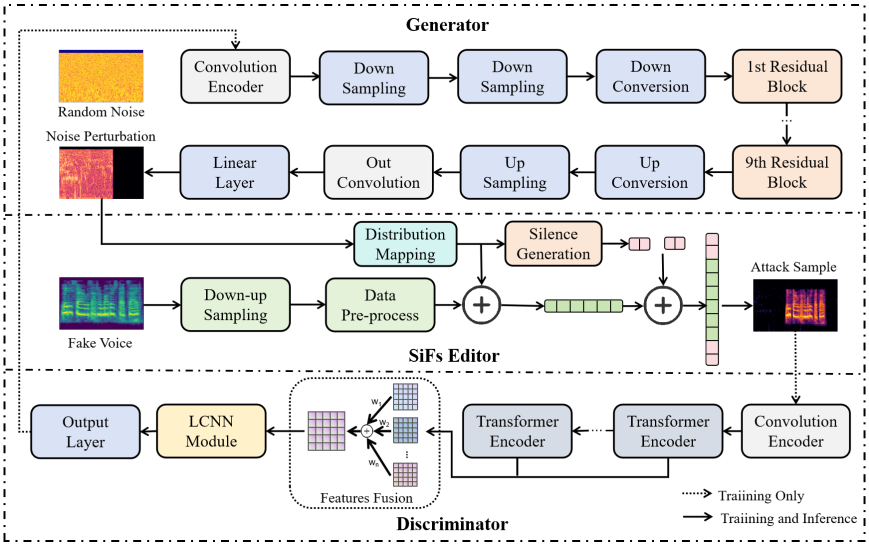


Figure 2: The SiFMimicEvader framework

speaker-irrelative features, 'Mimic' reflects the ANMA design principles, and 'Evader' signifies the attack strategy aimed at evading fake voice detection. The attack goal of SiFMimicEvader is to evade fake voice detection. To be specific, let x represent a fake voice, $D(\cdot)$ represent a fake voice detector, and $E(\cdot)$ represent SiFMimicEvader. The attack process can be represented as:

$$D(E(x)) = Real \quad Diff(x, E(x)) < threshold \tag{1}$$

Our approach leverages ANMA and SiFs as its foundation. Specifically, our attack is based on two main prerequisites: 1) Existing fake voice detectors tend to overfit to SiFs, meaning that modifying these specific SiFs can influence their judgments; 2) Recent advanced fake voice detectors frequently use speech representation models as feature extractors. Developing a fake voice detector based on a pre-trained speech representation model can effectively replicate the behavior of these detectors.

The structure of SiFMimicEvader is shown in Figure 2. SiFMimicEvader is a GAN-based attack framework that manipulates specific SiFs in the input fake voice to execute attacks. We select the background noise, silence segments before and after human voice, and the sampling rate of the audio file as the target features which are extremely representative SiFs.

SiFMimicEvader comprises three main components: a noise perturbation generator, a SiFs editor, and a mimic discriminator. The noise perturbation generator is a DNN-based audio data generation model that transforms random input data into an ideal noise perturbation. The SiFs editor is responsible for performing a series of editing operations on SiFs of the target fake voice to deceive the fake voice detectors. SiFMimicEvader takes the target fake voice as input. First, a data pre-processing model is used to remove specific SiFs from the input fake voice. Next, a silence perturbation is generated based on the noise perturbation which is the output of the noise perturbation generator. These two perturbations are then combined with the input fake voice to produce the attack sample. Finally, SiFs editor resamples the attack sample to further increase the attack success rate. The mimic discriminator is a fake voice detector built on ANMA principles to replicate the behavior of other detectors. It comprises a speech representation model as the upstream feature extractor and a downstream network for additional feature processing. This component is used exclusively during the training phase to develop a high-quality noise perturbation generator.

4.2 SiFs Selection

SiFMimicEvader aims to edit the specific SiFs of the input fake voice to deceive the fake voice detector into making wrong judgments. The selection of the SiFs used to attack follows the following two basic principles: 1) The SiFs perturbation must make sure that it does not affect the content of the input fake voice; 2) The perturbation should cause as little damage to the hearing as possible to ensure the stealth of the attack.

Following these principles, we conduct extensive research and experimental testing on fake voice detectors and select three representative SiFs: background noise, silence segments before and after human voice, and sampling rate. The rationale behind these choices is outlined below.

Background noise: Human voice recordings inevitably contain background noise due to environmental factors and device limitations, even in professional settings where thermal noise from electronics persists. Similarly, fake voices generated by synthesis algorithms exhibit mechanical noise, varying by algorithm. The differing noise distributions between human and fake voices are critical for detectors to distinguish them. However, many detection systems overfit to background noise [28], making it theoretically possible to add specific noise perturbations to fake voice samples to mislead detectors.

Silence segments before and after human voice: Human voice recordings often include silence segments before and after speech due to recording start and end times, whereas fake voices typically lack such segments. This difference in silence patterns can lead to detector overfitting, as noted in recent studies [6, 30, 46, 48]. While removing silence using voice activity detection (VAD) might seem beneficial, it actually degrades detector performance, causing high false positive rates in real-world scenarios. Thus, silence segments represent a potential vulnerability for targeted attacks.

Sampling rate: Most fake voice detectors process audio inputs of a fixed length, truncating or padding clips to standardize their size. However, these detectors often do not enforce a uniform sampling rate, instead using the original file's rate. This can cause distortions when the actual sampling rate differs from the assumed rate. For instance, truncating a file to 40,000 data points results in 2.5 seconds at 16 kHz but only 1.75 seconds at 22.05 kHz. Such discrepancies alter the proportion of silence segments, potentially impacting detector accuracy and reliability.

4.3 SiFs Editor

The SiFs editor is responsible for modifying specific SiFs of the input fake voice. The processing flow of the SiFs editor consists of the following steps. First, the input audio signal x with an original sampling rate f_s is subjected to resampling. The signal is upsampled to a target rate $f_u = 22.05\,\text{kHz}$:

$$x_u = \text{Upsample}(x_d, f_u). \tag{2}$$

This operation aims to increase the proportion of silence perturbation in subsequent processing if target detectors do not enforce a fixed sampling rate for input.

Next, the resampled signal x_u undergoes denoising using a denoising function $\mathcal{D}(\cdot)$:

$$x'_d = \mathcal{D}(x_u). \tag{3}$$

This process, implemented via a wavelet tool [9], eliminates background noise.

Following denoising, the noise perturbation P generated by the noise perturbation generator is introduced. To reduce the impact of perturbation P on the input voice, it is mapped to a target distribution characterized by mean μ_t and standard deviation σ_t. The mapping operation is defined as:

$$P' = \mu_t + \tau_t \cdot \frac{P - \mu_p}{\sigma_p}, \tag{4}$$

where μ_p and σ_p are the mean and standard deviation of the original perturbation distribution. Then a length adjustment operation is used to make the perturbation length match the length of the input:

$$P'' = \begin{cases} \text{cut}(p', \dim(x'_d)) & \text{if } \dim(P') \geq \dim(x'_d), \\ \text{pad}(P', \dim(x'_d)) & \text{if } \dim(P') < \dim(x'_d). \end{cases} \tag{5}$$

The function $\text{pad}(x, t)$ extends x to length t by repeating its content, while $\text{cut}(x, t)$ extracts a segment of length t from x. Additionally, $\dim(x)$ denotes the length of x. The adjusted perturbation P'' is added to the denoised signal:

$$x_m = x'_d + P''. \tag{6}$$

A fixed-length silence perturbation S is then generated by extracting a noise perturbation segment of duration $T_s = 1.5$ s:

$$S = \text{ExtractSegment}(P', T_s). \tag{7}$$

This silence perturbation is appended to both ends of the perturbed signal x_m, forming the final adversarial sample:

$$x_{att} = S \| x_m \| S, \tag{8}$$

where $\|$ denotes the concatenation operation.

4.4 Noise Perturbation Generator

The noise perturbation generator is designed to produce ideal noise perturbations in real-time while minimizing the number of required queries to the greatest extent possible. To achieve this, we employ a DNN-based model as the generator. For our implementation in this paper, we adopt the same generator structure as used in the study [21]. The work introduces a GAN-based framework to achieve high-quality voice conversion and we refer to its generator network structure in this work.

Specifically, the noise perturbation generator processes a mel-frequency spectrum input of size 35x128. For the input, we experiment with various options and ultimately select random data of size 35 × 128 for our attack. Alternative inputs, such as a real voice sample from the same speaker, are also tested but result in suboptimal performance. Initially, a convolutional encoder filters out noise information and aggregates features from the input spectrum. The spectrum is then transformed into sequence data through two down-sampling blocks and a conversion block. Subsequently, a residual module, consisting of nine residual blocks, performs a deep transformation on the sequence distribution. Following this, an up-sampling and conversion operation restores the spectrum data. Finally, the output layer maps the spectrum to 1D voice data. Leveraging the noise perturbation generator, SiFMimicEvader learns the noise distribution most likely to induce anomalies in the output of the upstream speech representation model of fake voice detectors.

4.5 Mimic Discriminator

The mimic discriminator aims to mimic the behavior of the fake voice detectors including the speech representation models to indirectly convert black-box attacks to white-box attacks. It is a binary classification model exclusively utilized during training to guide the optimization of the noise perturbation generator by providing feedback on its outputs. The design of the discriminator can be divided into two parts: upstream selection and downstream design.

Upstream selection: There are widely used speech representation models in multiple downstream applications in the audio field. These models have similar structures and are trained with specific open-source datasets. To achieve high mimic similarity, we select the wav2vec2.0-xlsr [8] as our upstream feature extractor, trained on a massive dataset spanning 436k hours of speech data across 128 languages, making it the most widely used pre-trained upstream model in fake voice detection. The model has three parameter sizes: 300M, 1B, and 2B. We use the 300M version, the most commonly used, in our implementation. The output of the wav2vec2.0-xlsr is a 24-layer feature map. Some existing detectors choose to fuse all or part of the layer features, while others select only the last layer map. To increase the versatility of the attack, we design a feature fusion module to fuse the last 12-layer maps. Let $[F_1, F_2, ..., F_n]$ represent the feature map of each layer and $[w_1, w_2, ...w_n]$ represent the fusion weights. The specific process of feature fusion is as follows:

$$\text{Fusion Map} = \sum_{i=1}^{n} w_i F_i \tag{9}$$

Downstream design: The objective of the downstream model is to identify anomalous patterns or deviations in the output of the upstream feature extractor. In this paper, we utilize a light-CNN (LCNN)-based downstream network, as described in [24], which is also part of a baseline system from the ASVspoof2021 challenge [10]. The downstream network consists of multiple convolution-based blocks for feature aggregation and a Bi-directional Long Short-Term Memory (BiLSTM) module to model cross-frame patterns, ultimately converting frame-level features into an utterance-level representation.

4.6 Training and Fine-Tuning

The optimization of SiFMimicEvader includes two phases: training and fine-tuning. The detailed process of the two phases is as follows:

Training: During the training stage, we adopt a GAN-like approach. We use a pre-trained mimic discriminator, which is trained on the ASVspoof2019 LA dataset, as the discriminator. Unlike standard GANs, the parameters of the discriminator remain fixed throughout the entire training process. We utilize the ASVspoof2019 LA training subset as the training dataset. First, attack samples are generated through the noise perturbation generator and SiFs editor. These attack samples are then input into the discriminator to obtain probability information. The parameters of the noise perturbation generator are subsequently updated using this probability information and a cross-entropy loss function. Furthermore, SiFMimicEvader can also be trained in a pure black-box scenario, where the attacker only has access to labels without any probability information.

Fine-tuning: The mimic discriminator can mimic the fake voice detectors' behavior including the speech representation model to some extent. However, for detectors without speech representation models, the mimicry may not be as accurate. Additionally, variations in the downstream design of detectors based on speech representation models can introduce additional behavioral differences. To optimize the attack effectiveness, we fine-tune the SiFMimicEvader after training for different attack targets. Specifically, we use the target detector as the discriminator to fine-tune the noise generator. The fine-tuning process is similar to the training. Furthermore, SiFMimicEvader can also be fine-tuned in a pure black-box scenario, where the attacker only has access to labels without any probability information. The fine-tuning can significantly optimize the attack effect of SiFMimicEvader with a small number of queries such as hundreds or thousands.

5 Evaluation

Our evaluation aims to answer the following research questions:

- **RQ1:** Is SiFMimicEvader effective in attack performance with low overhead?
- **RQ2:** Does the attack perturbation added by SiFMimicCracker significantly impact the quality of the audio?
- **RQ3:** How much do different SiFs contribute to the attack's effectiveness?

5.1 Evaluation Setup

In this subsection, we will briefly introduce the evaluation environment information, the implementation details of SiFMimicEvader used in this evaluation, baseline detectors' information, and the datasets used in the experiment. The specific information is as follows:

Implementation Details: During the training stage of the noise perturbation generator, we train for 30 epochs using the Adam optimizer with an initial learning rate of 0.0001. The loss function used is cross-entropy. In the fine-tuning stage, most conditions remain the same as in the training stage, except for a reduced learning rate of 0.00001. For the LCNN network used in the mimic discriminator, we adopt the same structure as the LFCC-LCNN, a baseline system from the ASVspoof2021 challenge, but replace its feature extractor with the wav2vec2.0-xlsr-300m model.

Target detectors: We select seven high-performance fake voice detectors from recent top conferences, journals, and challenges as evaluation targets. Specifically, five detectors lack speech representation models: RawNet2 [35], RawGAT-ST [34], AASIST [20], Raw-pc-darts [13], and TSSDNet [18]. Two detectors, Tak-SSL [36] and SLSforADD [45], incorporate speech representation models. All detectors with speech representation models use the same pre-trained model, wav2vec2.0-xlsr, which is the most widely used in this field. This open-source model, trained on the largest dataset in the field, achieves state-of-the-art performance and is the preferred choice for nearly all detectors using speech representation models. For all target detectors, we use the authors' open-source implementations and default parameters. We determine the judgment threshold by calculating the EER on the ASVspoof2019 LA evaluation subset. If a detector's output score exceeds its EER threshold, the input is classified as a real voice.

Datasets: For baseline detectors, except for two detectors built by us, we use the authors' pre-trained weights for evaluation. All of the baseline detectors are trained on the ASVspoof2019 LA dataset. For the SiFMimicEvader, we also use the ASVspoof2019 LA dataset. We select all of the spoof samples from the training subset as the dataset during the training stage and use the corresponding samples in the development subset during the fine-tuning stage. To simulate a query-limited scenario, we select 4,800 samples and ensure an equal number of samples from each spoof algorithm during the fine-tuning stage.

Metrics: We define the success of our attack as the Success Acceptance Rate (SAR), which represents the probability that an attack sample is classified as a human voice sample. Specifically, let the Acceptance Account (AA) denote the number of attack samples classified as human voices, and the Rejection Account (RA) denote the number of failed attack samples. SAR can then be calculated as follows:

$$SAR = \frac{AA}{AA + RA} \tag{10}$$

5.2 Attack Performance Evaluation (RQ1)

In this evaluation, we focus on answering RQ1: Is SiFMimicEvader effective in attack performance with low overhead? To evaluate attack performance, we use all spoof samples from the ASVspoof2019 LA evaluation subset, comprising 63,882 audio files. To mitigate the impact of original silence segments, we preprocess the samples

Table 1: The performance comparison result.

Attack Method	Target Detectors							Average
	RawNet2	RawGAT-ST	AASIST	Raw-pc-darts	TSSDNet	Tak-SSL	SLSforADD	
SiF-DeepVC	**81.68%**	37.90%	12.34%	24.28%	6.41%	0.35%	0.26%	23.32%
SiFDetectCracker	69.50%	51.19%	57.45%	**70.08%**	50.48%	18.73%	6.00%	40.42%
Kassis et al.	9.27%	0.99%	3.01%	3.50%	24.65%	0.25%	0.01%	5.96%
SiFMimicEvader	65.22%	55.84%	**41.21%**	63.83%	50.82%	**62.34%**	31.22%	52.92%
SiFMimicEvader (label-only)	66.79%	**59.00%**	41.17%	63.88%	**51.44%**	62.17%	**55.90%**	**57.19%**

using Sound Exchange (SoX), a widely adopted audio processing tool, to remove silence segments prior to evaluation. We select three advanced attack methods for comparison: SiF-DeepVC [28], SiFDetectCracker [17], and Kassis et al [22]. SiF-DeepVC and Kassis et al. are universal attack methods, not optimized for a specific target. SiFDetectCracker, based on traditional adversarial attack algorithms, reports excellent performance. All implementations are collected from the authors' open-source repositories, with attack parameters matching those in the original versions.

SiF-DeepVC and Kassis et al. generate attack samples rapidly, so we use the same test data size as SiFMimicEvader. In contrast, SiFDetectCracker requires iterative optimization, consuming about five minutes per audio attack, making large-scale evaluation impractical. This makes evaluating it on a dataset with tens of thousands of samples impractical. Therefore, we evaluate SiFDetectCracker on 195 spoof samples, selected using the method reported in their paper. Since different target detectors exhibit varying levels of misjudgment, we remove the misjudged samples of each detector from their respective datasets to eliminate this influence.

The results in Table 1 show that SiFMimicEvader achieves excellent performance, with average SAR exceeding 50% regardless of using probability or label-only feedback, outperforming existing methods. Notably, the label-only version performs better, suggesting that too much target information may interfere with the discriminator's learned features. While some attacks (e.g., SiF-DeepVC and SiFDetectCracker) perform well on detectors without speech representation models (e.g., >80% SAR on RawNet2), their success drops sharply (<20% SAR) on those with such models, highlighting the robustness of representation-based detectors. Moreover, SiF-DeepVC produces noticeable perturbations, and SiFDetectCracker requires excessive queries and 5 minutes per attack. In contrast, our label-only method achieves over 50% SAR on representation-based targets, demonstrating ANMA's effectiveness in enhancing attack performance against advanced detectors.

Most recent advanced fake voice detectors rely on speech representation models to enhance feature extraction capabilities, resulting in superior performance compared to detectors without such models. This trend suggests that integrating speech representation models will be a key development in the design of future fake voice detectors. The superior attack performance against targets including speech representation models further demonstrates that ANMA holds significant potential against emerging detectors in the future. Additionally, the high SAR when attacking detectors without upstream models further shows that our approach can deceive a wide range of fake voice detectors.

Query time represents a crucial performance metric for evaluating the efficiency of the attack. SiFMimicEvader requires no queries during the training stage and only a limited number of queries during the fine-tuning stage. To assess the extent of SiFMimicEvader's reliance on queries, we evaluate its attack performance using varying fine-tuning sample sizes. The result is shown in Figure 3. Overall, the SAR shown in this figure demonstrates that SiFMimicEvader does not rely on large-scale query operations in most cases. For most detectors without speech representation models, increasing the fine-tuning sample size has minimal impact on attack performance. This suggests that the unreasonable judgment factors present in these models, which can be exploited for the attack, also exist in our discriminator. Consequently, fine-tuning does not provide additional useful information to enhance the attack. For other detectors, the attack performance positively correlates with the sample size. Except for SLSforADD, significant SAR gains are achieved after fine-tuning with just 1,200 queries, a feasible number for real-world attacks. Even for SLSforADD, the least sensitive detector, SAR exceeds 20% after fine-tuning with 1,200 queries. Furthermore, queries are required solely during the fine-tuning stage, allowing the attacker to distribute query requests over an extended period. Once fine-tuning is completed, the attacker can generate attack samples for any voice input without the need for additional queries. This demonstrates that our attack can adapt to attack scenarios with strict requirements on the number of queries.

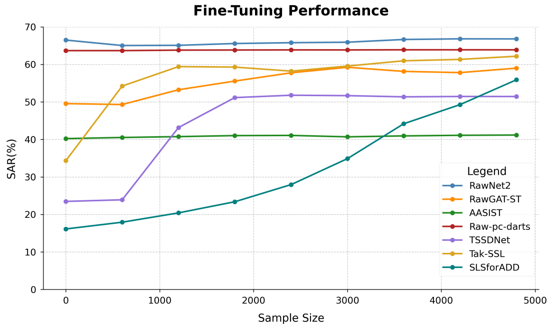


Figure 3: The fine-tuned attack performance with different training sample sizes.

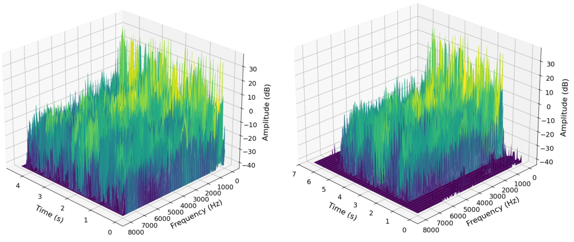
5.3 Quality Evaluation (RQ2)

This evaluation primarily addresses RQ2: Does the attack perturbation added by SiFMimicCracker significantly impact the quality of the audio? To evaluate the quality of our attack samples, we use both objective and subjective methods. The objective evaluation focuses on two metrics: the signal-to-noise ratio (SNR) and the similarity score (SS), which ranges from -1 to 1, with higher values indicating greater audio similarity. The SNR measures the perturbation level relative to the original audio, while the SS, computed using an ASV system [42], quantifies the similarity between original and attack samples. For subjective evaluation, 30 volunteers assessed the attack samples using a structured questionnaire without prior knowledge of the evaluation's purpose to ensure unbiased feedback. We randomly selected 5 original samples and 7 corresponding attack samples from the attack performance data for different target detectors, pairing each original with its attack sample. For each pair, participants answered two multiple-choice questions: one on the similarity between the samples and the other on the noticeability of added noise perturbations. For the questions, a scoring scale from 0 to 5 is used, where 0 indicates 'completely different' and 5 indicates 'completely consistent' with regard to similarity.

Table 2: The quality evaluation result.

Objective Evaluation			Subjective Evaluation	
Targets	**SNR**	**SS**	**Similarity**	**Noise**
RawNet2	8.89	0.73	**3:** 32.68% **2:** 58.57% **0-1:** 8.57%	**3:** 31.43% **2:** 52.68% **0-1:** 15.71%
RawGAT-ST	7.83	0.75		
AASIST	7.68	0.77		
Raw-pc-darts	7.76	0.74		
TSSDNet	8.56	0.74		
Tak-SSL	7.56	0.75		
SLSforADD	7.41	0.77		
Average	7.96	0.75		

The quality evaluation results are shown in Table 2. In the objective evaluation, the average SNR is approximately 8, indicating that while the attack samples contain some noise, it is unlikely to impair human perception of the underlying information. The SS metric supports this observation, showing a similar trend. In the subjective evaluation, the vast majority of participants perceived our attack audio as very similar to the original audio, with little to no noticeable noise. This suggests that the perturbation introduced by our attack is difficult for humans to detect under most circumstances. Additionally, when we proactively asked participants about their impressions of the attack samples, most of them believed the noise they heard was simply background noise from the device.



(a) Original

(b) SiFMimicEvader

Figure 4: The spectra of the original sample and attack sample generated by SiFMimicEvader.

The spectrum of our attack sample is shown in Figure 4. We randomly select an attack sample and its corresponding original sample to visualize their spectra. As shown, the main components of the speech spectra are nearly identical, further confirming that the noise perturbation has minimal impact on the original audio information. The primary difference is that the attack sample includes additional silence before and after the audio, which is typically not perceived as abnormal.

5.4 Ablation Evaluation (RQ3)

In this evaluation, we mainly focus on RQ3: How much do different SiFs contribute to the attack's effectiveness? To analyze the impact of each component on attack performance, we design a series of comparison groups. In each group, a specific component of the attack framework is removed, while all other conditions are kept consistent with those used in the attack performance evaluation. Specifically, we establish five experimental groups: the raw group, no noise group, no silence group, no resampling group, and no fine-tuning group. The raw group maintains the same conditions as the attack performance evaluation, while the other groups exclude a specific SiFs operation, as indicated by their respective names. Since the silence segment perturbation inherently includes the noise perturbation, we utilize a segment consisting entirely of zeros as the silence segment perturbation in the no noise group.

Table 3: Ablation evaluation result. The metric is SAR. C1-C4 represent no noise group, no resampling group, no silence group, and no fine-tuning group respectively.

Targets	Groups				
	Raw	C1	C2	C3	C4
RawNet2	66.79%	64.66%	68.67%	0.75%	66.48%
RawGAT-ST	59.00%	50.35%	66.65%	0.10%	49.55%
AASIST	41.17%	40.48%	41.55%	0.35%	40.22%
Raw-pc-darts	63.88%	63.60%	54.65%	0.12%	63.67%
TSSDNet	51.44%	3.99%	54.77%	8.47%	23.48%
Tak-SSL	62.17%	28.59%	53.75%	0.00%	34.35%
SLSforADD	55.90%	3.41%	47.47%	0.02%	16.14%
Average	57.19%	36.44%	55.36%	1.40%	41.99%

The ablation evaluation results are presented in Table 3. Overall, all comparison groups show varying degrees of performance decline. In the no noise group, the average performance drop exceeds 20%, highlighting the critical role of noise perturbation in the attack. On the other hand, the results for the no silence group demonstrate that the silence segments before and after the human voice are crucial to the success of our attack. Successfully executing the attack without silence segment perturbation—relying solely on noise perturbation and resampling—is almost impossible. This suggests that the performance decline observed in the no noise group is primarily due to changes in the noise within the silence segments. In the no fine-tuning group, the SAR remains above 40%, outperforming other comparison attack methods in the attack performance evaluation. This demonstrates that our attack maintains a significant performance advantage even without fine-tuning. Consequently, it achieves excellent performance even under the most stringent constraints.

6 Conclusion

In this paper, we introduce a novel attack strategy called ANMA, which leverages open-source speech representation models commonly used in fake voice detection to transform black-box attacks into white-box attacks. Building on this concept, we design a new black-box attack framework, SiFMimicEvader, combined with SiFs to deceive fake voice detectors. The evaluation results demonstrate that SiFMimicEvader outperforms other attack methods, achieving an attack success rate exceeding 60%. Furthermore, it eliminates the need for large-scale targeted queries and exhibits high robustness, underscoring its strong potential for real-world application.

Acknowledgments

This work is supported by Gansu Provincial Department of Education Young Doctoral Fund Project No. 2025CXZX049, Lanzhou University Fundamental Research Funds for Central Universities No. lzujbky-2025-it02 and No. lzujbky-2025-jdzx10, the National Natural Science Foundation of China under Grant No. 62302438, Supercomputing Center of Lanzhou University, and the Open Research Fund of The State Key Laboratory of Blockchain and Data Security, Zhejiang University.

References

[1] Ehab A AlBadawy, Siwei Lyu, and Hany Farid. 2019. Detecting AI-Synthesized Speech Using Bispectral Analysis.. In *CVPR workshops*. 104–109.

[2] Moustafa Alzantot, Ziqi Wang, and Mani B Srivastava. 2019. Deep residual neural networks for audio spoofing detection. *arXiv preprint arXiv:1907.00501* (2019).

[3] Dora M Ballesteros, Yohanna Rodriguez-Ortega, Diego Renza, and Gonzalo Arce. 2021. Deep4SNet: deep learning for fake speech classification. *Expert Systems with Applications* 184 (2021), 115465.

[4] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. 2019. Neural population control via deep image synthesis. *Science* 364, 6439 (2019), eaav9436.

[5] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. 2022. Who are you (i really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction. In *31st USENIX Security Symposium (USENIX Security 22)*. 2691–2708.

[6] Bhusan Chettri, Emmanouil Benetos, and Bob LT Sturm. 2020. Dataset Artefacts in anti-spoofing systems: a case study on the ASVspoof 2017 benchmark. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 3018–3028.

[7] Tripti Choudhary, Vishal Goyal, and Atul Bansal. 2023. WTASR: Wavelet Transformer for Automatic Speech Recognition of Indian Languages. *Big Data Mining and Analytics* 6, 1 (2023), 85–91. doi:10.26599/BDMA.2022.9020017

[8] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979* (2020).

[9] Ingrid Daubechies et al. 2023. PyWavelets: Wavelet Transforms in Python. https://github.com/PyWavelets/pywt. Accessed on 2023-12-24.

[10] Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, and Others. 2021. ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv preprint arXiv:2109.00535* (2021).

[11] Siwen Ding, You Zhang, and Zhiyao Duan. 2023. Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[12] Yang Gao, Tyler Vuong, Mahsa Elyasi, Gaurav Bharaj, and Rita Singh. 2021. Generalized spoofing detection inspired from audio generation artifacts. *arXiv preprint arXiv:2104.04111* (2021).

[13] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2021. Raw differentiable architecture search for speech deepfake and spoofing detection. *arXiv preprint arXiv:2107.12212* (2021).

[14] Alejandro Gomez-Alanis, Antonio M Peinado, Jose A Gonzalez, and Angel M Gomez. 2019. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In *Proc. Interspeech*, Vol. 2019. 1068–1072.

[15] Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. 2024. Audio Deepfake Detection With Self-Supervised Wavlm And Multi-Fusion Attentive Classifier. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 12702–12706.

[16] Xuan Hai, Xin Liu, Zhaorun Chen, Yuan Tan, Song Li, Weina Niu, Gang Liu, Rui Zhou, and Qingguo Zhou. 2024. Ghost-in-Wave: How Speaker-Irrelative Features Interfere DeepFake Voice Detectors. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[17] Xuan Hai, Xin Liu, Yuan Tan, and Qingguo Zhou. 2023. SiFDetectCracker: An Adversarial Attack Against Fake Voice Detection Based on Speaker-Irrelative Features. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8552–8560.

[18] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. 2021. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters* 28 (2021), 1265–1269.

[19] The Wall Street Journal. 2019. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. https://www.wsj.com/articles/fraudsters-use-ai-to-mimicceos-voice-in-unusual-cybercrime-case-11567157402

[20] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6367–6371.

[21] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2019. Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *arXiv preprint arXiv:1907.12279* (2019).

[22] Andre Kassis and Urs Hengartner. 2023. Breaking Security-Critical Voice Authentication. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 951–968.

[23] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. 2018. Fooling end-to-end speaker verification with adversarial examples. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 1962–1966.

[24] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. 2019. STC antispoofing systems for the ASVspoof2019 challenge. *arXiv preprint arXiv:1904.05576* (2019).

[25] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. 2021. Replay and synthetic speech detection with res2net architecture. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6354–6358.

[26] Xu Li, Jinghua Zhong, Xixin Wu, Jianwei Yu, Xunying Liu, and Helen Meng. 2020. Adversarial attacks on GMM i-vector based speaker verification systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6579–6583.

[27] Songxiang Liu, Haibin Wu, Hung-yi Lee, and Helen Meng. 2019. Adversarial attacks on spoofing countermeasures of automatic speaker verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 312–319.

[28] Xin Liu, Yuan Tan, Xuan Hai, Qingchen Yu, and Qingguo Zhou. 2023. Hidden-in-Wave: A Novel Idea to Camouflage AI-Synthesized Voices Based on Speaker-Irrelative Features. In *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 786–794.

[29] Xin Liu, Xiaokang Zhou, and Qingguo Zhou. 2022. Human or Not: Can You Really Detect the Fake Voices? doi:10.13140/RG.2.2.19572.83849

[30] Nicolas M Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. 2021. Speech is silver, silence is golden: What do ASVspoof-trained models really learn? *arXiv preprint arXiv:2106.12914* (2021).

[31] NBC News. 2024. Fake Biden robocall telling Democrats not to vote is likely an AI-generated deepfake. https://www.nbcnews.com/tech/misinformation/joe-biden-new-hampshire-robocall-fake-voice-deep-ai-primary-rcna135120

[32] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[33] Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116* (2023).

[34] Hemlata Tak, Jee-weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2021. Graph attention networks for anti-spoofing. *arXiv preprint arXiv:2104.03654* (2021).

[35] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6369–6373.

[36] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop*.

[37] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[38] Xiaohai Tian, Rohan Kumar Das, and Haizhou Li. 2019. Black-box attacks on automatic speaker verification using feedback-controlled voice conversion. *arXiv preprint arXiv:1909.07655* (2019).

[39] Nik Vaessen and David A Van Leeuwen. 2022. Fine-tuning wav2vec2 for speaker recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7967–7971.

[40] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* 12 (2016).

[41] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023).

[42] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. [n. d.]. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. *arXiv preprint arXiv:2303.00332* ([n. d.]).

[43] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. 2020. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia*. 1207–1216.

[44] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[45] Qishan Zhang, Shuangbing Wen, and Tao Hu. 2024. Audio deepfake detection with self-supervised XLS-R and SLS classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6765–6773.

[46] Yuxiang Zhang, Zhuo Li, Jingze Lu, Hua Hua, Wenchao Wang, and Pengyuan Zhang. 2023. The Impact of Silence on Speech Anti-Spoofing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[47] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng. 2022. Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification. *arXiv preprint arXiv:2203.15249* (2022).

[48] Yuxiang Zhang12, Wenchao Wang12, and Pengyuan Zhang12. 2021. The effect of silence and dual-band fusion in anti-spoofing system. In *Proc. Interspeech*.

[49] Chu-Xiao Zuo, Zhi-Jun Jia, and Wu-Jun Li. 2024. AdvTTS: Adversarial Text-to-Speech Synthesis Attack on Speaker Identification Systems. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4840–4844.