

Article

Evaluating Adversarial Robustness of Deepfake Audio Detectors and Vocoder Fingerprint Detectors Against Universal Adversarial Perturbations

Quang Minh Tran , Wei Zong * , Yang-Wai Chow  and Willy Susilo 

Institute of Cybersecurity and Cryptology, School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia; qmt496@uowmail.edu.au (Q.M.T.); caseyc@uow.edu.au (Y.-W.C.); wsusilo@uow.edu.au (W.S.)

* Correspondence: wzong@uow.edu.au

Abstract

Audio deepfake and vocoder fingerprint detectors are increasingly used to identify synthetic speech and attribute it to its generating model. However, their robustness against adversarial perturbations remains unclear across attack algorithms, perturbation domains, detector representations, and vocoder types. This paper presents a focused, quality-aware evaluation of four representative adversarial attacks, namely the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD), and Carlini–Wagner (CW) attack, against audio deepfake and vocoder fingerprint detectors. Each attack is implemented in both the waveform domain and the short-time Fourier transform (STFT) magnitude domain. All attacks are optimized against Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST) under a targeted fake-to-real objective and are evaluated on synthetic speech generated by HiFi-GAN, Fullband MelGAN, StyleMelGAN, and Parallel WaveGAN. Attack performance is first measured on the source AASIST detector, after which black-box transferability is assessed on three target detector families: ResNet with Linear Frequency Cepstral Coefficient (LFCC) features, LCNN with Constant-Q Cepstral Coefficient (CQCC) features, and a bidirectional long short-term memory (BiLSTM) detector. The results show that adversarial effectiveness depends strongly on perturbation domain and detector representation. STFT-magnitude PGD transfers strongly to LFCC-based ResNet detectors but has limited effect on CQCC-based and recurrent detectors. In contrast, waveform-domain attacks produce broader transferability across feature-based detectors, with different attacks showing distinct ASR–quality trade-offs. Under the chosen waveform-domain budget, FGSM and BIM preserve transcription-level intelligibility while retaining meaningful black-box transferability, whereas CW provides the strongest overall source-detector and black-box attack performance. To distinguish effective adversarial perturbations from destructive signal degradation, we evaluate audio quality and intelligibility using word error rate (WER) and signal-to-noise ratio (SNR). Overall, the findings show that robustness claims in audio deepfake and vocoder fingerprint detection are limited when adversarial perturbations, black-box transferability, and audio quality are jointly considered.

Keywords: adversarial attacks; audio deepfake detection; vocoder fingerprints; black-box transferability; STFT-magnitude perturbation; waveform-domain perturbation; speech anti-spoofing



Academic Editor: Carlo Blundo

Received: 26 May 2026

Revised: 23 June 2026

Accepted: 27 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Neural vocoders have become a central component of modern speech synthesis systems, enabling the generation of highly natural speech waveforms from intermediate acoustic representations such as mel-spectrograms. Models such as HiFi-GAN, MelGAN variants, StyleMelGAN, and Parallel WaveGAN [1–4] can produce synthetic speech with high perceptual quality, making synthetic audio increasingly difficult to distinguish from genuine recordings. While these systems support legitimate applications in text-to-speech and voice conversion, they also increase the risk of audio deepfakes being used for impersonation, misinformation, and fraud [5–8].

To address this risk, audio deepfake detection methods aim to distinguish synthetic speech from real speech. Beyond binary real-versus-fake detection, a related forensic task is vocoder fingerprint detection, where the objective is to identify the vocoder or synthesis model responsible for generating a speech sample [9–11]. The motivation behind vocoder fingerprinting is that different neural vocoders may leave model-specific artifacts in the generated waveform [11,12]. These artifacts can appear in spectral structure, cepstral patterns, temporal dynamics, or other signal characteristics learned by the detector. If reliable, vocoder fingerprints can support both synthetic speech detection and source attribution.

However, the reliability of detecting vocoder fingerprints depends on whether these artifacts remain stable under perturbation. A detector may achieve high accuracy on clean synthetic speech but fails when the input is intentionally modified to disrupt the fingerprint cues used for classification. This is particularly important in adversarial settings, where an attacker may attempt to preserve the perceived speech content while causing the detector to misclassify the sample [13–15]. Therefore, evaluating clean accuracy alone is insufficient for assessing the practical robustness of vocoder fingerprint detectors.

Adversarial attacks provide a direct way to examine this vulnerability [13,16]. In audio, adversarial perturbations can be applied in different domains. Waveform-domain attacks modify the raw audio samples directly, while frequency-domain attacks modify a time-frequency representation such as the short-time Fourier transform (STFT) magnitude before reconstructing the waveform. These perturbation domains may interact differently with detector architectures. For example, waveform-based detectors such as Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST) [17] may rely on raw temporal cues, while feature-based detectors using linear frequency cepstral coefficient (LFCC) or constant-Q cepstral coefficient (CQCC) representations [18,19] may be more sensitive to spectral and cepstral distortions.

This transferability problem is central to realistic adversarial evaluation [20]. In practice, an attacker may not know the exact detector used by a forensic system. If adversarial audio optimized against one detector also fools other detectors, this indicates a broader weakness across detector families. Conversely, if transferability is limited, this suggests that robustness may depend strongly on the detector representation. Although recent work has begun to provide broader comparative evaluations of adversarial robustness in audio deepfake detection [15], less attention has been given to universal perturbations for vocoder fingerprint attribution and to how perturbations transfer across detectors with different cepstral, recurrent, and raw-waveform representations.

This paper presents a focused, quality-aware evaluation protocol of adversarial attacks against audio deepfake and vocoder fingerprint detectors. We evaluate four representative attacks: Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD), and Carlini–Wagner (CW). Each attack is implemented in both the waveform domain and the short-time Fourier transform (STFT) magnitude domain. All attacks are optimized against AASIST using a targeted fake-to-real objective. Attack performance is first measured on the source AASIST detector, and the resulting adversarial audio

is then evaluated for black-box transferability across three target detector families: ResNet with LFCC features, light convolutional neural network (LCNN) with CQCC features, and a bidirectional long short-term memory (BiLSTM)-based detector. Experiments are conducted on synthetic speech generated by HiFi-GAN, Fullband MelGAN, StyleMelGAN, and Parallel WaveGAN.

The contributions of this paper are threefold. First, rather than proposing a new adversarial attack algorithm, this study provides a focused, controlled evaluation protocol of representative adversarial attacks under a unified AASIST-source optimization setting. Second, we compare waveform-domain and STFT-magnitude-domain perturbations to examine how perturbation domain affects attack success, black-box transferability, and audio quality. Third, we evaluate transferability across detectors with different input representations and architectures, while also measuring intelligibility and signal distortion using word error rate and signal-to-noise ratio. Through this analysis, we show that adversarial robustness in audio deepfake and vocoder fingerprint detection depends jointly on attack algorithm, perturbation domain, detector representation, and audio quality.

2. Related Work

This section reviews the research areas that motivate the present study. It first discusses audio deepfake detection and vocoder fingerprinting, then summarizes detector representations used for synthetic speech detection. It then reviews adversarial attacks and transferability, with particular attention to prior work on adversarial robustness in audio deepfake detection and speech anti-spoofing.

2.1. Audio Deepfake Detection and Vocoder Fingerprints

Audio deepfake detection aims to determine whether a speech signal is genuine or synthetically generated. Existing detectors commonly approach this task either as binary real-versus-fake classification or as a more fine-grained source attribution problem. In the binary setting, the detector only decides whether an utterance is real or synthetic [12,17]. In the source attribution setting, the detector attempts to identify the synthesis model or vocoder that produced the speech. The latter task is closely related to vocoder fingerprint detection.

Vocoder fingerprints refer to model-specific traces that remain in generated speech after waveform synthesis [9–12]. These traces may arise from the architecture, training objective, upsampling strategy, spectral reconstruction loss, or adversarial training procedure used by a vocoder. For example, generative adversarial network (GAN)-based vocoders such as HiFi-GAN, MelGAN variants, StyleMelGAN, and Parallel WaveGAN reconstruct waveform details using different generator and discriminator designs. As a result, their outputs may contain distinguishable spectral or temporal artifacts that can be exploited by fingerprint detectors.

The existence of vocoder fingerprints makes source attribution possible, but it also raises an important robustness question. If the detector relies on subtle synthesis artifacts, then small adversarial perturbations may be able to disrupt those cues while preserving the apparent speech content. Therefore, clean classification accuracy alone does not establish that vocoder fingerprints are stable or reliable in adversarial settings. This motivates evaluating whether fingerprint detectors remain effective when synthetic audio is deliberately perturbed.

2.2. Detector Representations for Synthetic Speech Detection

Audio deepfake detectors differ in the representations they use to analyze speech. Feature-based detectors first transform the waveform into an acoustic feature representation

and then classify the extracted features. Common front-end features include mel-frequency cepstral coefficients (MFCCs), linear frequency cepstral coefficients (LFCCs), and constant-Q cepstral coefficients (CQCCs). These features compress spectral information into lower-dimensional representations that can be used by convolutional or recurrent classifiers.

LFCC and CQCC features are particularly relevant to vocoder fingerprint detection because they capture different frequency structures. LFCCs [18] use linearly spaced filters, which can preserve high-frequency and narrowband artifacts that may be useful for detecting synthetic speech. CQCCs [19] are based on the constant-Q transform and use a logarithmic frequency structure, which may emphasize different spectral patterns. As a result, perturbations that disrupt LFCC-based detectors may not necessarily transfer to CQCC-based detectors.

End-to-end detectors follow a different approach by operating directly on raw waveform inputs or minimally processed representations. These models learn task-specific features from the signal rather than relying on a fixed cepstral front end. AASIST is an example of this direction and is designed for speech anti-spoofing using raw waveform input. Compared with fixed-feature detectors, end-to-end detectors may exploit temporal, spectral, or phase-related cues that are not explicitly represented in hand-crafted features.

These differences in detector representation are central to adversarial transferability. An adversarial perturbation optimized against a raw-waveform detector may not affect a cepstral-feature detector in the same way, and an attack that disrupts LFCC features may not generalize to CQCC features or recurrent temporal models. For this reason, the present study evaluates multiple detector families, including AASIST, ResNet+LFCC, LCNN+CQCC, and a BiLSTM-based detector.

2.3. Adversarial Attacks and Transferability

Adversarial attacks modify an input signal to cause a machine learning model to make an incorrect prediction [13,21]. In audio deepfake detection, this means perturbing a speech waveform so that a detector fails to recognize it as synthetic or fails to attribute it to the correct vocoder. Such attacks are especially concerning when the perturbation preserves speech intelligibility while disrupting detector-relevant artifacts.

Gradient-based attacks provide a common framework for evaluating model robustness. FGSM [22] applies a single gradient-sign update, while BIM [23] extends this idea through multiple iterative updates. PGD [24] further strengthens the iterative attack by using random initialization and projection within a bounded perturbation region. CW [25] uses an optimization-based objective that directly manipulates the model logits while balancing classification success and perturbation size. These attacks differ in computational cost, optimization strength, and sensitivity to hyperparameters.

In the audio domain, adversarial perturbations can be applied in different signal domains. Waveform-domain attacks directly modify the audio samples, while frequency-domain attacks modify spectral representations such as STFT magnitude before reconstructing the waveform. This distinction is important because detectors do not all rely on the same signal evidence. Raw-waveform models may be affected by temporal or phase-related perturbations, while cepstral-feature detectors may be more sensitive to spectral magnitude changes.

Black-box transferability [26] is also important for practical evaluation. A perturbation optimized against one source detector may or may not transfer to other detectors with different representations and architectures. If transferability is high, the attack reveals a broader weakness across detector families. If transferability is limited, the vulnerability may be specific to the source detector or feature representation. Although recent work has begun to provide broader comparative evaluations of adversarial robustness in audio

deepfake detection, less attention has been given to the joint relationship between universal perturbations, perturbation domain, vocoder attribution, detector representation, and audio quality.

This work addresses that gap by comparing FGSM, BIM, PGD, and CW in both waveform and STFT-magnitude domains under a unified AASIST-source optimization protocol. The optimized universal perturbations are then evaluated across multiple vocoder fingerprint detectors and vocoder types without re-optimization. This design allows the study to examine source-detector attack success, black-box detector transferability, cross-vocoder transferability, and quality-aware attack effectiveness.

2.4. Adversarial Robustness in Audio Deepfake Detection

Recent studies have examined the vulnerability of audio deepfake detection systems under adversarial and anti-forensic conditions. Kawa et al. [14] evaluated adversarial attacks against multiple audio deepfake detection architectures in both white-box and transferability settings, and further investigated adversarial training as a defense strategy. Their results show that audio deepfake detectors can be vulnerable to adversarial perturbations and that robustness varies across detector architectures.

Farooq et al. [27] proposed a transferable adversarial attack framework for audio deepfake detection using surrogate models and a GAN-based generation strategy. Their work emphasizes the importance of black-box transferability and shows that adversarial examples can degrade the performance of state-of-the-art audio deepfake detectors beyond the source model. Wang et al. [28] studied BIM-based attacks against speech deepfake detectors in both time-domain and frequency-domain settings, showing that perturbation domain can affect attack effectiveness under white-box and black-box evaluation.

A closely related recent benchmark by Uddin et al. [15] provides a broad comparative study of anti-forensic attacks against audio deepfake detection systems. Their evaluation includes raw-audio and spectrogram-based detectors across multiple benchmark datasets, and considers both statistical attacks, such as pitch shifting, filtering, noise addition, and quantization, and optimization-based attacks, including FGSM, PGD, C&W, and DeepFool. This work provides valuable evidence that audio deepfake detectors remain vulnerable under diverse anti-forensic conditions.

More specifically, the attacks considered in this study are representative of the gradient-based and optimization-based methods commonly used to evaluate adversarial robustness. FGSM provides a single-step first-order attack, while BIM and PGD provide stronger iterative variants based on projected gradient updates. CW represents an optimization-based attack that directly targets the model decision boundary through a logit-margin objective. These attacks have also appeared in recent audio deepfake and speech anti-spoofing robustness studies. For example, Wang et al. [28] focused on BIM-based attacks against speech deepfake detectors in both time-domain and frequency-domain settings, while Uddin et al. [15] included FGSM, PGD, and CW as part of a broader benchmark of anti-forensic attacks against audio deepfake detection systems. The present work builds on this direction by evaluating these attacks as universal perturbations under a unified AASIST-source protocol and by examining their effects on both binary fake-to-real evasion and vocoder fingerprint attribution.

The present study differs from these works in focus and evaluation design. Rather than providing a broad benchmark of general binary audio deepfake detectors, this work focuses on universal adversarial perturbations for audio deepfake and vocoder fingerprint detectors. In particular, all attacks are optimized under a unified AASIST-source protocol and evaluated for black-box transferability across detectors with different representations, including LFCC-based, CQCC-based, and recurrent models. The study also examines

vocoder attribution failure and reports WER and SNR to distinguish quality-preserving adversarial perturbations from destructive signal degradation. This provides a focused, quality-aware evaluation of how universal perturbations affect both binary deepfake detection and vocoder fingerprint attribution.

3. Methodology

3.1. Overview

This study compares the robustness of audio deepfake and vocoder fingerprint detectors under multiple adversarial attack settings. Four attacks are evaluated: FGSM, BIM, PGD, and CW. Each attack is implemented in two perturbation domains: the waveform domain and the STFT-magnitude domain.

All attacks are optimized against AASIST using a targeted fake-to-real objective. The learned adversarial perturbations are then applied to synthetic speech generated by multiple vocoders and evaluated on both the source detector and independent target detectors. This design allows us to examine source-detector attack success, black-box transferability, and the effect of perturbation domain on detector robustness.

3.2. Problem Formulation

Let $x \in [-1, 1]^T$ denote a clean synthetic speech waveform of length T . The source detector is denoted as $f_s(\cdot)$, where f_s corresponds to AASIST in this study. AASIST is treated as a binary classifier with two classes:

$$0 = \text{fake}, \quad 1 = \text{real}. \quad (1)$$

For a synthetic input x , the true label is $y = 0$, and the target adversarial label is $y_t = 1$. The objective is to generate an adversarial waveform x_{adv} such that the source detector predicts the target class while the audio remains close to the original waveform:

$$x_{\text{adv}} = \mathcal{A}(x; \delta), \quad (2)$$

where $\mathcal{A}(\cdot)$ denotes the attack transformation and δ denotes the adversarial perturbation. The form of δ depends on the perturbation domain. In the waveform domain, δ is added directly to the raw audio samples. In the STFT-magnitude domain, δ is applied to the magnitude spectrum before waveform reconstruction.

3.3. Universal Perturbation Setting

All attacks are implemented as universal perturbations. Instead of optimizing a separate perturbation for each utterance, each attack learns one reusable perturbation from a source vocoder dataset:

$$\mathcal{D}_{\text{src}} = \{x_1, x_2, \dots, x_N\}. \quad (3)$$

The universal perturbation is optimized as:

$$\delta^* = \arg \min_{\delta \in \mathcal{S}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(f_s(\mathcal{A}(x_i; \delta)), y_t), \quad (4)$$

where δ^* is the optimized perturbation, $\mathcal{S}$ is the allowed perturbation set, and $\mathcal{L}(\cdot)$ is the attack loss. For FGSM, BIM, and PGD, cross-entropy loss toward the target real class is used. For CW, a logit-margin loss is used.

After optimization, the same perturbation δ^* is applied to evaluation audio without re-optimization:

$$x_{\text{adv}} = \mathcal{A}(x; \delta^*). \quad (5)$$

This universal setting creates a stricter transferability evaluation than per-sample optimization because the perturbation must generalize across utterances and vocoders. This choice is intentional because the study focuses on reusable perturbations and black-box transferability rather than on the maximum attack success achievable when each utterance is optimized independently.

3.4. Perturbation Domains

3.4.1. Waveform-Domain Perturbation

In the waveform-domain setting, the perturbation is added directly to the raw waveform:

$$x_{\text{adv}} = \text{clip}(x + \delta_w, -1, 1), \quad (6)$$

where δ_w is the waveform-domain perturbation. The clipping operation ensures that the adversarial waveform remains within the normalized amplitude range. The perturbation is constrained by an ℓ_∞ bound:

$$\|\delta_w\|_\infty \leq \epsilon. \quad (7)$$

This setting evaluates whether direct sample-level perturbations can disrupt detectors that operate on raw waveforms or on features extracted from waveforms.

3.4.2. STFT-Magnitude-Domain Perturbation

In the STFT-magnitude-domain setting, the perturbation is applied to the magnitude spectrum. Given a waveform x , its short-time Fourier transform is:

$$X = \text{STFT}(x). \quad (8)$$

The complex STFT is decomposed into magnitude and phase:

$$M = |X|, \quad P = \frac{X}{|X| + \eta}, \quad (9)$$

where η is a small constant used to avoid division by zero. The adversarial magnitude is computed as:

$$M_{\text{adv}} = M + \delta_M, \quad (10)$$

where δ_M is the STFT-magnitude perturbation. The adversarial waveform is reconstructed using the original phase:

$$x_{\text{adv}} = \text{ISTFT}(M_{\text{adv}} \cdot P). \quad (11)$$

Since STFT magnitudes are not bounded within a fixed amplitude range, a relative perturbation budget is used:

$$|\delta_M| \leq \epsilon_{\text{rel}} \cdot M_{\text{ref}}, \quad (12)$$

where M_{ref} is a reference magnitude template computed from the source training set, and ϵ_{rel} controls the perturbation strength.

3.5. Attack Algorithms

The four attacks are selected to cover representative categories of widely used adversarial optimization methods while keeping the evaluation controlled and comparable across perturbation domains. FGSM is included as a single-step gradient-sign attack and provides a low-complexity baseline [22]. BIM and PGD are included as iterative projected-gradient attacks, allowing the study to compare stronger multi-step variants against the one-step FGSM baseline [23,24]. CW is included as an optimization-based attack that uses a logit-margin objective rather than a direct cross-entropy gradient-sign update [25].

Together, these methods cover single-step, iterative, projected-gradient, and optimization-based attack families that are commonly used in adversarial robustness evaluation. They have also been considered in recent audio deepfake and speech anti-spoofing robustness studies [15,28].

Other adversarial attack techniques, such as DeepFool [29], query-based black-box attacks, psychoacoustic attacks, generative attacks, and adaptive defense-aware attacks, are not included in the present study. This exclusion is intentional rather than a claim that these methods are unimportant. The goal of this work is to provide a focused, quality-aware comparison of universal perturbations under a unified AASIST-source optimization protocol. Including additional attack families would introduce different assumptions, such as query access, perceptual masking constraints, learned attack generators, or defense-specific adaptation, which would make the comparison less controlled. These alternative attacks are therefore left for future work.

FGSM performs a single targeted gradient-sign update:

$$\delta^* = -\epsilon \cdot \text{sign}(\nabla_{\delta} \mathcal{L}_{\text{CE}}(f_s(\mathcal{A}(x; \delta)), y_t)). \quad (13)$$

BIM extends FGSM by applying iterative projected updates:

$$\delta^{k+1} = \Pi_S \left(\delta^k - \alpha \cdot \text{sign} \left(\nabla_{\delta} \mathcal{L}_{\text{CE}} \left(f_s(\mathcal{A}(x; \delta^k)), y_t \right) \right) \right), \quad (14)$$

where α is the step size and $\Pi_S(\cdot)$ projects the perturbation back into the allowed set S .

PGD follows the same iterative update rule as BIM but starts from a random initialization within the allowed perturbation set:

$$\delta^0 \sim \mathcal{U}(S). \quad (15)$$

CW is implemented as an optimization-based targeted attack using a logit-margin objective. Let z_{fake} and z_{real} denote the fake and real logits produced by AASIST. The CW loss is:

$$\mathcal{L}_{\text{CW}} = \max(z_{\text{fake}} - z_{\text{real}} + \kappa, 0), \quad (16)$$

where κ is a confidence margin. The total CW objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{D}(\delta) + c \cdot \mathcal{L}_{\text{CW}}, \quad (17)$$

where $\mathcal{D}(\delta)$ penalizes perturbation size and c controls the weight of the classification loss.

3.6. Transferability Evaluation

The optimized adversarial audio is evaluated under two criteria: source-detector evasion and black-box transferability. For AASIST, attack success is defined as fake audio being classified as real:

$$\text{ASR}_{\text{AASIST}} = \frac{N_{\text{fake} \rightarrow \text{real}}}{N}, \quad (18)$$

where $N_{\text{fake} \rightarrow \text{real}}$ is the number of fake samples predicted as real, and N is the total number of evaluated samples.

For multi-class vocoder attribution detectors, attack success is defined as incorrect vocoder attribution:

$$\text{ASR}_{\text{multi}} = 1 - \frac{N_{\text{correct}}}{N}, \quad (19)$$

where N_{correct} is the number of attacked samples still classified as their true vocoder class. This definition treats any incorrect vocoder prediction as a successful disruption of source attribution.

3.7. Audio Quality and Intelligibility Metrics

Attack success alone is insufficient because a detector may fail when the audio is heavily degraded. Therefore, adversarial effectiveness is evaluated together with speech intelligibility and signal distortion.

Word error rate (WER) is used to measure transcription-level intelligibility:

$$\text{WER} = \frac{S + D + I}{N_w}, \quad (20)$$

where S , D , and I are the numbers of substitutions, deletions, and insertions, respectively, and N_w is the number of words in the reference transcript. The WER increase is defined as:

$$\Delta\text{WER} = \text{WER}_{\text{adv}} - \text{WER}_{\text{clean}}. \quad (21)$$

Whisper WER is used as an automatic proxy for transcription-level intelligibility rather than as a complete substitute for human perceptual evaluation. It provides a consistent and reproducible way to compare whether adversarial perturbations affect the recognizability of speech content across attack settings. However, automatic WER does not fully capture subjective perceptual quality, naturalness, annoyance, or listener preference. Therefore, WER is interpreted together with SNR as an objective quality indicator, while subjective listening evaluation is left for future work.

Signal-to-noise ratio (SNR) is used to measure signal-level distortion between clean and adversarial waveforms:

$$\text{SNR} = 10 \log_{10} \left(\frac{\|x\|_2^2}{\|x - x_{\text{adv}}\|_2^2 + \eta} \right), \quad (22)$$

where η is a small constant used to avoid division by zero. Higher SNR indicates that the adversarial audio remains closer to the clean waveform.

4. Experimental Setup

This section describes the experimental conditions used to evaluate adversarial robustness. It introduces the dataset and vocoders, the source and target detector configurations, the attack hyperparameters, and the evaluation protocol used to measure attack success, black-box transferability, and audio quality.

4.1. Dataset and Vocoders

Experiments are conducted using synthetic speech derived from the LJSpeech corpus [30]. LJSpeech contains approximately 13,100 English utterances from a single female speaker recorded at 22.05 kHz. This corpus provides a controlled setting for evaluating vocoder-generated speech and adversarial perturbations because speaker identity, language, and recording conditions are held relatively constant. This makes it possible to focus the evaluation on the interaction between vocoder type, perturbation domain, attack algorithm, and detector representation. However, this controlled design also limits the generalizability of the results to multi-speaker, multilingual, noisy, compressed, or in-the-wild audio deepfake scenarios.

Four neural vocoders are evaluated: HiFi-GAN, Fullband MelGAN, StyleMelGAN, and Parallel WaveGAN. HiFi-GAN is used as the source vocoder for adversarial optimiza-

tion, while all four vocoders are used during evaluation. This setup allows the experiment to measure both same-vocoder performance and cross-vocoder transferability. After an attack is optimized on HiFi-GAN-generated speech, the resulting universal perturbation is applied to all vocoder outputs without re-optimization.

All adversarial audio is generated at 22.05 kHz. For detectors requiring a different sampling rate, such as AASIST, audio is resampled internally during evaluation. The attacks are applied after waveform synthesis, and no vocoder model weights are modified.

4.2. Detector Configuration

Four detector configurations are used: AASIST, ResNet + LFCC, LCNN + CQCC, and BiLSTM. These detectors are selected to cover different input representations and model architectures.

AASIST is used as the source detector for attack optimization. It is treated as a binary fake-versus-real classifier, where the target attack objective is to classify synthetic speech as real. The remaining detectors are used to evaluate black-box transferability.

The ResNet + LFCC detector is a multi-class vocoder attribution model referenced from baseline work [11]. Each waveform is converted into linear frequency cepstral coefficient (LFCC) features before classification. The LCNN + CQCC detector follows the same multi-class attribution setting but uses constant-Q cepstral coefficient (CQCC) features with a light convolutional neural network. The BiLSTM detector is included as a temporal sequence-based model to test whether attacks optimized against AASIST transfer to recurrent architectures.

For the multi-class detectors, the true label corresponds to the source vocoder of the evaluated sample. An attack is considered successful if the detector no longer predicts the correct vocoder class. For AASIST, an attack is considered successful if synthetic speech is classified as real.

4.3. Attack Configuration

Table 1 summarizes the hyperparameters used to generate universal perturbations in the waveform-domain and STFT-magnitude-domain settings. The same attack settings were applied across vocoders within each perturbation domain so that differences in attack success could be attributed to the interaction between attack algorithm, perturbation domain, vocoder type, and detector representation rather than to changes in optimization settings. The perturbation budgets are domain-specific because waveform-domain and STFT-magnitude-domain attacks operate on different signal representations; therefore, their raw ϵ values should not be interpreted as directly equivalent. For this reason, the attack results are interpreted together with WER and SNR measurements in addition to ASR.

The waveform-domain and STFT-magnitude-domain perturbation budgets have different physical meanings. In the waveform domain, ϵ directly bounds the maximum sample-level amplitude change in the normalized waveform. In the STFT-magnitude domain, ϵ_{rel} bounds the perturbation relative to a reference magnitude template, so it constrains the spectral-magnitude perturbation as a fixed proportion of the reference magnitude. Therefore, the raw numerical values of ϵ and ϵ_{rel} should not be interpreted as directly equivalent across domains.

The perturbation budgets are selected as quality-aware operating points for evaluating universal adversarial perturbations under each signal representation. For waveform-domain attacks, $\epsilon = 0.002$ is used to keep the perturbation within a low-WER regime while still allowing measurable detector disruption. For STFT-magnitude-domain attacks, $\epsilon_{\text{rel}} = 0.05$ is used to constrain the perturbation relative to the reference magnitude template. Changing these budgets would be expected to affect the distortion–effectiveness trade-off:

larger budgets may increase ASR but reduce SNR and increase WER, whereas smaller budgets may preserve audio quality but reduce attack success. Therefore, the cross-domain comparison in this study is interpreted through ASR together with WER and SNR, rather than through raw perturbation-budget values alone.

Four adversarial attacks are evaluated: FGSM, BIM, PGD, and CW. Each attack is implemented in both waveform-domain and STFT-magnitude-domain settings. All attacks are optimized against AASIST using a targeted fake-to-real objective and saved as universal perturbation checkpoints.

For waveform-domain attacks, the perturbation is added directly to the audio waveform. The universal perturbation is optimized over fixed-length chunks of 4 s. Since the sampling rate is 22,050 Hz, each waveform perturbation contains:

$$T_\delta = 4 \times 22,050 = 88,200 \quad (23)$$

samples. The adversarial waveform is then clipped to the valid amplitude range $[-1, 1]$ after perturbation.

For STFT-magnitude-domain attacks, the perturbation is applied to the magnitude spectrum while preserving the original phase for inverse STFT reconstruction. The STFT uses $n_{\text{fft}} = 1024$, hop length 256, and window length 1024. The magnitude perturbation is constrained using a relative perturbation budget based on a reference magnitude template computed from the source data.

Table 1. Attack hyperparameters used for waveform-domain and STFT-magnitude-domain attacks.

Parameter	Waveform Domain	STFT-Magnitude Domain
Sampling rate	22,050 Hz	22,050 Hz
Chunk length	4 s/88,200 samples	4 s/88,200 samples
STFT n_{fft}	–	1024
STFT hop length	–	256
STFT window length	–	1024
Perturbation budget	$\epsilon = 0.002$	$\epsilon_{\text{rel}} = 0.05$
BIM step size	$\alpha = 0.00025$	$\alpha_{\text{rel}} = 0.003$
PGD step size	$\alpha = 0.00025$	$\alpha_{\text{rel}} = 0.003$
BIM/PGD steps	5000	5000
Batch size	8	8
CW steps	5000	5000
CW learning rate	0.001	0.001
CW confidence margin	$\kappa = 0$	$\kappa = 0$
CW loss weight	$c = 1.0$	$c = 1.0$

The waveform-domain attacks are optimized using 4 s universal perturbations with $\epsilon = 0.002$. FGSM is implemented as a one-step universal attack. BIM and PGD are iterative projected-gradient attacks with step size $\alpha = 0.00025$, with PGD using random initialization within the allowed perturbation range. CW is implemented as an optimization-based attack using a targeted logit-margin objective and Adam optimization. During iterative attack optimization, random cropping is used so that the universal perturbation is exposed to different temporal regions of the source utterances.

4.4. Evaluation Protocol

The evaluation consists of three stages. First, each attack is optimized using HiFi-GAN-generated speech and AASIST as the source detector. Second, the saved universal perturbation checkpoint is applied to HiFi-GAN, Fullband MelGAN, StyleMelGAN, and Parallel

WaveGAN audio without additional optimization. Third, the clean and adversarial audio are evaluated on AASIST, ResNet+LFCC, LCNN+CQCC, and BiLSTM.

For each detector and vocoder, clean accuracy is first recorded to verify that the detector performs reliably before attack. Attack success rate (ASR) is then computed using the definitions from Section 3. For AASIST, ASR measures fake-to-real misclassification. For multi-class detectors, ASR measures incorrect vocoder attribution.

Audio quality and intelligibility are evaluated using Whisper [31] word error rate (WER) and signal-to-noise ratio (SNR). These metrics are reported alongside ASR to distinguish quality-preserving adversarial examples from attacks that mainly succeed through signal degradation.

4.5. Implementation Details and Reproducibility

The detector-training data were derived from a randomly selected subset of LJSpeech. A total of 3678 LJSpeech utterances were used to construct the detector-training corpus. Each selected utterance was included as a real speech sample and was also vocoded using the vocoder models available in the detector-training corpus. This produced a seven-class vocoder-fingerprint classification setting, consisting of HiFi-GAN [1], StyleMelGAN [3], Parallel WaveGAN [4], MelGAN [32], Multiband MelGAN [2], Fullband MelGAN [2], and real speech. The resulting file list was split into training, validation, and internal test partitions using stratified random sampling with a fixed random seed of 42. The resulting split contained 20,596 training files, 2575 validation files, and 2575 internal test files.

The adversarial robustness evaluation was conducted on a separate held-out evaluation subset. Specifically, 942 additional LJSpeech utterances were sampled from the remaining utterances not used to construct the detector-training corpus. These utterances were vocoded using the four vocoders evaluated in the main experiments: HiFi-GAN, Fullband MelGAN, StyleMelGAN, and Parallel WaveGAN. The attack evaluation is therefore restricted to these four vocoder sources. The remaining vocoder classes in the detector-training label space were used to train the multi-class attribution detectors but were not used as source vocoders in the reported adversarial evaluation. The real class is not treated as an attack source class, because the attacks are applied to synthetic speech and the AASIST source-detector objective is targeted fake-to-real evasion.

AASIST was adopted as a pretrained detector from the original AASIST work and was not retrained in this study. It was used as the source detector for adversarial optimization under a targeted fake-to-real objective. During AASIST evaluation, audio was resampled to 16 kHz and adjusted to the fixed input length required by the model using repeat-padding or cropping.

The ResNet + LFCC, LCNN + CQCC, and BiLSTM + LFCC detectors were implemented and trained from scratch for this study. The ResNet+LFCC detector follows the baseline vocoder-fingerprint setting and uses LFCC features as input. LFCC features were extracted at 22.05 kHz using $n_{\text{fft}} = 1024$, hop length 220, window length 551, 20 linear filters, and 20 cepstral coefficients. Static, delta, and delta-delta coefficients were concatenated to produce a 60-dimensional feature representation. The BiLSTM detector used the same LFCC feature representation, while the LCNN detector used CQCC features.

All trainable target detectors were optimized using cross-entropy loss and the Adam optimizer with learning rate 0.001 and momentum parameters (0.9, 0.98). A linear learning-rate decay schedule was applied during training, and gradient clipping with maximum norm 5.0 was used to improve training stability. ResNet + LFCC was trained for 100 epochs with batch size 256. LCNN + CQCC was trained for 100 epochs with batch size 64. BiLSTM + LFCC was trained for 60 epochs with batch size 128. Checkpoints were saved during

training, validation accuracy and validation loss were recorded at each epoch, and the final checkpoint was used for evaluation. No early stopping criterion was used.

The experiments were implemented using two Python environments. The custom vocoder fingerprint detectors were implemented in Python 3.9.21 with PyTorch 2.6.0+cu118, torchaudio 2.6.0+cu118, librosa 0.10.1, NumPy 1.24.3, and scikit-learn 1.6.1. The AASIST-based attack optimization and evaluation were conducted in Python 3.14.4 with PyTorch 2.11.0+cu128, torchaudio 2.11.0+cpu, librosa 0.11.0, NumPy 2.4.3, and scikit-learn 1.8.0. Whisper version 20250625 was used for WER-based intelligibility evaluation.

5. Results

This section reports the empirical results of the robustness evaluation. The results are first presented separately for STFT-magnitude-domain and waveform-domain attacks, followed by a cross-domain comparison, an analysis of black-box transferability, and an audio quality evaluation using WER and SNR.

5.1. Frequency-Domain Attack Results

Table 2 reports the ASR of STFT-magnitude-domain attacks across the four detectors and four vocoders. The strongest result is observed for PGD on ResNet + LFCC, where ASR reaches approximately 100% across all vocoders. Since the attack is optimized against AASIST rather than ResNet + LFCC, this indicates strong black-box transfer from the AASIST source detector to the LFCC-based attribution detector.

Table 2. Attack success rate (%) of STFT-magnitude-domain attacks across detectors and vocoders.

Detector	Vocoder	Clean Acc.	PGD	FGSM	BIM	CW
ResNet + LFCC	HiFi-GAN	99.3	99.99	6.00	2.00	6.00
	Fullband MelGAN	96.8	99.99	3.00	4.00	6.00
	StyleMelGAN	98.7	100.00	4.00	2.00	0.00
	Parallel WaveGAN	100.0	99.99	0.00	0.00	0.00
LCNN + CQCC	HiFi-GAN	96.0	0.00	7.00	8.00	6.00
	Fullband MelGAN	95.0	0.00	27.00	15.00	10.00
	StyleMelGAN	100.0	0.00	2.00	2.00	1.00
	Parallel WaveGAN	100.0	15.00	0.00	0.00	0.00
BiLSTM	HiFi-GAN	99.79	0.00	1.91	0.74	1.38
	Fullband MelGAN	99.47	0.00	4.03	1.70	1.17
	StyleMelGAN	100.0	0.00	0.11	0.42	0.00
	Parallel WaveGAN	100.0	0.00	0.00	0.00	0.00
AASIST	HiFi-GAN	94.0	34.00	6.69	6.37	6.37
	Fullband MelGAN	93.0	45.00	8.60	8.07	8.28
	StyleMelGAN	97.0	26.00	3.40	3.08	3.29
	Parallel WaveGAN	94.0	53.00	6.79	6.37	6.69

However, this transfer does not generalize uniformly across detectors. The same STFT-magnitude PGD perturbation has limited effect on LCNN+CQCC and no measurable effect on BiLSTM. For AASIST, PGD achieves moderate source-detector ASR, ranging from 26% to 53% across vocoders. FGSM, BIM, and CW remain weak in the STFT-magnitude setting, with low ASR on both AASIST and most black-box detectors. Overall, the frequency-domain results show that STFT-magnitude PGD is effective mainly against LFCC-based attribution, while the other attacks are largely ineffective under the tested budget.

5.2. Waveform-Domain Attack Results

Table 3 reports the ASR of waveform-domain attacks. Compared with the STFT-magnitude setting, waveform-domain perturbations produce broader transferability across feature-based detectors. ResNet+LFCC is highly vulnerable, with PGD and CW reaching near-complete or complete attribution failure across vocoders. FGSM and BIM also remain effective against ResNet + LFCC, although their ASR values vary across vocoders.

Table 3. Attack success rate (%) of waveform-domain attacks across detectors and vocoders.

Detector	Vocoder	Clean Acc.	PGD	FGSM	BIM	CW
ResNet + LFCC	HiFi-GAN	99.3	100.0	77.29	100.0	100.0
	Fullband MelGAN	96.8	92.0	74.85	83.02	100.0
	StyleMelGAN	98.7	100.0	99.69	100.0	100.0
	Parallel WaveGAN	100.0	100.0	93.00	99.69	100.0
LCNN + CQCC	HiFi-GAN	96.0	100.0	73.00	100.0	100.0
	Fullband MelGAN	95.0	100.0	87.00	100.0	100.0
	StyleMelGAN	100.0	100.0	100.0	100.0	100.0
	Parallel WaveGAN	100.0	100.0	6.00	100.0	100.0
BiLSTM	HiFi-GAN	99.79	99.79	99.79	99.68	100.0
	Fullband MelGAN	99.47	9.45	3.29	8.39	99.79
	StyleMelGAN	100.0	32.59	33.97	39.07	100.0
	Parallel WaveGAN	100.0	0.32	11.89	0.64	9.87
AASIST	HiFi-GAN	94.0	37.0	7.01	28.56	85.35
	Fullband MelGAN	93.0	39.0	9.02	31.00	89.60
	StyleMelGAN	97.0	24.0	3.29	17.62	71.66
	Parallel WaveGAN	94.0	38.0	7.22	30.57	87.69

LCNN + CQCC is also strongly affected by waveform-domain attacks. PGD, BIM, and CW reach 100% ASR across all vocoders, while FGSM produces high ASR for HiFi-GAN, Fullband MelGAN, and StyleMelGAN but transfers poorly to Parallel WaveGAN. This indicates that waveform-domain perturbations can disrupt CQCC-based attribution, although the transferability of single-step FGSM remains more vocoder-dependent.

The BiLSTM detector shows a more mixed pattern. It is highly vulnerable for HiFi-GAN, but its response varies substantially for the other vocoders. Fullband MelGAN and Parallel WaveGAN remain relatively robust to PGD, FGSM, and BIM, whereas CW achieves high ASR for most vocoders except Parallel WaveGAN. This suggests that recurrent temporal modelling changes the way adversarial perturbations transfer across vocoder sources.

For AASIST, CW is the strongest waveform-domain attack, reaching consistently high fake-to-real ASR across vocoders. PGD produces moderate ASR, while FGSM and BIM achieve lower source-detector ASR under the chosen waveform-domain budget. However, FGSM and BIM still transfer substantially to several black-box feature-based detectors. This shows that source-detector success does not necessarily predict black-box transferability.

5.3. Cross-Domain Comparison

Table 4 summarizes the average ASR across vocoders for both perturbation domains. The results show a clear domain effect. In the STFT-magnitude domain, high ASR is concentrated mainly in PGD against ResNet + LFCC. In the waveform domain, high ASR is distributed more broadly across attacks and detectors, especially for ResNet + LFCC and LCNN + CQCC.

The ranking of attacks also changes across domains. For AASIST, STFT-magnitude PGD is the strongest frequency-domain attack, whereas waveform-domain CW is the

strongest attack overall. Waveform-domain FGSM and BIM have lower source-detector ASR than CW, but they still show substantial transfer to feature-based black-box detectors. For BiLSTM, STFT-magnitude attacks have almost no effect, whereas waveform-domain attacks produce higher but more variable ASR. These results show that perturbation domain is a major factor in adversarial effectiveness and transferability.

Table 4. Average ASR (%) across vocoders for STFT-magnitude and waveform-domain attacks.

Domain	Detector	PGD	FGSM	BIM	CW
STFT-mag	ResNet + LFCC	99.99	3.25	2.00	3.00
	LCNN + CQCC	3.75	9.00	6.25	4.25
	BiLSTM	0.00	1.51	0.72	0.64
	AASIST	39.50	6.37	5.97	6.16
Waveform	ResNet + LFCC	98.00	86.21	95.68	100.00
	LCNN + CQCC	100.00	66.50	100.00	100.00
	BiLSTM	35.54	37.24	36.95	77.42
	AASIST	34.50	6.64	26.94	83.58

5.4. Black-Box Transferability Across Detectors

Table 5 compares source-detector ASR with black-box transfer ASR averaged across vocoders. Since all attacks are optimized against AASIST, the AASIST column represents source-detector performance, while ResNet + LFCC, LCNN + CQCC, and BiLSTM represent black-box target detectors.

Table 5. Source-detector ASR and black-box transfer ASR averaged across vocoders. AASIST is the source detector, while ResNet + LFCC, LCNN + CQCC, and BiLSTM are black-box target detectors.

Domain	Attack	AASIST ASR	ResNet + LFCC	LCNN + CQCC	BiLSTM	Avg. Black-Box ASR
STFT-mag	PGD	39.50	99.99	3.75	0.00	34.58
STFT-mag	FGSM	6.37	3.25	9.00	1.51	4.59
STFT-mag	BIM	5.97	2.00	6.25	0.72	2.99
STFT-mag	CW	6.16	3.00	4.25	0.64	2.63
Waveform	PGD	34.50	98.00	100.00	35.54	77.85
Waveform	FGSM	6.64	86.21	66.50	37.24	63.31
Waveform	BIM	26.94	95.68	100.00	36.95	77.54
Waveform	CW	83.58	100.00	100.00	77.42	92.47

In the STFT-magnitude domain, PGD achieves the strongest source-detector ASR and transfers almost completely to ResNet+LFCC. However, this transfer is not consistent across all black-box detectors, as the same perturbation has limited effect on LCNN + CQCC and BiLSTM. The other STFT-magnitude attacks show weak source-detector and black-box transfer performance. This indicates that STFT-magnitude PGD mainly exposes a representation-specific vulnerability in the LFCC-based detector rather than a uniformly transferable weakness across all detector families.

In contrast to the STFT-magnitude setting, waveform-domain attacks produce stronger black-box transfer overall. CW achieves the highest average black-box ASR, followed by PGD and BIM. Waveform-domain FGSM has low source-detector ASR on AASIST but still achieves substantial average black-box ASR. This confirms that source-detector evasion and black-box transferability should be evaluated separately. An attack that is weak against the source detector may still disrupt independent target detectors with different feature representations.

5.5. Audio Quality and Intelligibility

Table 6 reports Whisper WER, Table 7 reports SNR, and Table 8 summarizes the average WER and SNR across vocoders. The clean vocoded audio has an average WER of 3.96%, providing the baseline for intelligibility comparison.

Table 6. Whisper WER (%) of clean and adversarial audio across vocoders.

Domain	Attack	HiFi-GAN	Fullband MelGAN	StyleMelGAN	Parallel WaveGAN	Avg. WER
Clean	None	3.74	3.99	4.12	3.97	3.96
STFT-mag	FGSM	3.72	3.97	4.03	3.99	3.93
	BIM	3.72	3.98	4.02	4.06	3.95
	CW	3.74	4.01	4.05	4.01	3.95
	PGD	4.65	5.50	5.51	5.28	5.24
Waveform	FGSM	3.82	4.14	4.19	4.06	4.05
	BIM	3.75	4.08	4.07	4.02	3.98
	CW	3.85	4.18	4.17	4.09	4.07
	PGD	3.84	4.12	4.16	3.98	4.03

Table 7. Signal-to-noise ratio (SNR, dB) of adversarial audio across vocoders. Values are reported as mean $\pm$ standard deviation.

Domain	Vocoder	FGSM	BIM	PGD	CW
STFT-mag	HiFi-GAN	41.69 $\pm$ 1.24	46.31 $\pm$ 1.24	11.00 $\pm$ 1.28	49.40 $\pm$ 1.29
	Fullband MelGAN	41.80 $\pm$ 1.27	46.50 $\pm$ 1.27	10.90 $\pm$ 1.28	49.50 $\pm$ 1.33
	StyleMelGAN	42.80 $\pm$ 1.29	47.40 $\pm$ 1.28	11.90 $\pm$ 1.25	50.48 $\pm$ 1.37
	Parallel WaveGAN	42.00 $\pm$ 1.28	46.50 $\pm$ 1.28	11.00 $\pm$ 1.28	49.70 $\pm$ 1.34
Waveform	HiFi-GAN	30.92 $\pm$ 1.22	33.44 $\pm$ 1.22	32.90 $\pm$ 1.22	23.90 $\pm$ 1.24
	Fullband MelGAN	31.10 $\pm$ 1.26	33.62 $\pm$ 1.27	33.10 $\pm$ 1.26	24.10 $\pm$ 1.28
	StyleMelGAN	32.06 $\pm$ 1.30	34.58 $\pm$ 1.25	34.10 $\pm$ 1.21	25.00 $\pm$ 1.23
	Parallel WaveGAN	31.13 $\pm$ 1.27	33.65 $\pm$ 1.27	33.20 $\pm$ 1.26	24.10 $\pm$ 1.28

Table 8. Average WER and SNR across vocoders.

Domain	Attack	Avg. WER	Δ WER	Avg. SNR
STFT-mag	FGSM	3.93	−0.03	42.07
STFT-mag	BIM	3.95	−0.01	46.68
STFT-mag	PGD	5.24	1.28	11.20
STFT-mag	CW	3.95	0.00	49.77
Waveform	FGSM	4.05	0.09	31.30
Waveform	BIM	3.98	0.02	33.82
Waveform	PGD	4.03	0.07	33.33
Waveform	CW	4.07	0.12	24.28

For waveform-domain attacks, FGSM and BIM preserve transcription-level intelligibility under the chosen perturbation budget. Waveform-domain FGSM has an average WER of 4.05% and an average SNR of 31.30 dB, while waveform-domain BIM has an average WER of 3.98% and an average SNR of 33.82 dB. These WER values are close to the clean baseline of 3.96%, indicating that these attacks do not rely on severe intelligibility degradation.

Waveform-domain PGD and CW also keep WER close to the clean baseline, with average WER values of 4.03% and 4.07%, respectively. Among the waveform-domain attacks, BIM and PGD achieve the highest SNR values, while CW obtains the strongest source-detector and black-box ASR. Overall, these results show that waveform-domain attacks can retain meaningful transferability while preserving transcription-level intelli-

gibility, but the ASR–SNR–WER trade-off depends strongly on the attack algorithm and perturbation domain.

6. Discussion

The results support the need for a focused, quality-aware evaluation protocol when evaluating adversarial robustness in audio deepfake and vocoder fingerprint detection. Robustness is not determined by the detector architecture alone. Instead, it depends on the interaction between attack algorithm, perturbation domain, detector representation, transferability setting, and audio quality. This is important because a detector that appears robust under one attack configuration may remain vulnerable under another, and a perturbation that achieves high attack success may not necessarily represent the strongest quality-aware attack. The findings therefore show that adversarial robustness claims should be interpreted jointly with black-box transferability and signal-quality measurements.

6.1. Perturbation Domain and Detector Representation

The comparison between waveform-domain and STFT-magnitude-domain attacks shows that perturbation domain has a major effect on adversarial behavior. STFT-magnitude attacks mainly expose representation-specific vulnerability. In particular, STFT-magnitude PGD transfers strongly to ResNet+LFCC but has limited impact on LCNN+CQCC and BiLSTM. This suggests that the perturbation disrupts spectral structures that are highly relevant to LFCC-based attribution, but less damaging to CQCC-based or recurrent detectors.

Waveform-domain attacks produce a different pattern. They transfer more broadly to feature-based detectors, especially ResNet+LFCC and LCNN+CQCC. This indicates that direct waveform perturbations can disrupt downstream feature extraction even when the target detector itself does not operate directly on raw waveform input. Therefore, feature-based detectors should not be assumed to be protected from waveform-level adversarial manipulation.

These findings suggest that robustness evaluation should include more than one perturbation domain. A detector may appear robust against STFT-magnitude perturbations while remaining vulnerable to waveform-domain attacks, or vice versa. The results therefore support a domain-aware view of adversarial robustness, where attack effectiveness must be interpreted in relation to the signal representation being perturbed and the representation used by the target detector.

6.2. Source-Detector Success Does Not Guarantee Transferability

All attacks in this study are optimized against AASIST, but the strongest transfer effects do not always correspond to the strongest source-detector attacks. The clearest example is waveform-domain FGSM. Although it achieves only low fake-to-real ASR on AASIST under the chosen waveform-domain budget, it still causes substantial attribution failure on black-box feature-based detectors. This shows that source-detector ASR is not sufficient for evaluating the broader risk of an attack.

This result has practical implications for black-box adversarial evaluation. If evaluation only considers the source detector, an attack such as waveform-domain FGSM would appear weak. However, when evaluated against independent target detectors, the same perturbation produces substantial attribution failure. Similarly, STFT-magnitude PGD achieves only moderate source-detector ASR on AASIST but transfers almost completely to ResNet+LFCC. Therefore, black-box transferability should be evaluated explicitly across independent detectors rather than inferred from source-detector performance.

The detector-specific transfer patterns also suggest that different detector families rely on different fingerprint cues. LFCC-based, CQCC-based, recurrent, and raw-waveform detectors do not respond identically to the same adversarial perturbation. This means that robustness claims based on a single detector architecture are limited. A more reliable evaluation should include heterogeneous target detectors with different front-end representations and model structures.

6.3. Quality-Aware Interpretation of Attack Strength

The audio quality results show that high ASR alone is not enough to identify a strong adversarial attack. Attack effectiveness must be interpreted together with WER and SNR because different perturbation domains and attack algorithms produce different ASR–quality trade-offs. Under the chosen waveform-domain budget, FGSM, BIM, PGD, and CW all maintain average WER close to the clean baseline. The clean vocoded audio has an average WER of 3.96%, while waveform-domain FGSM, BIM, PGD, and CW achieve average WER values of 4.05%, 3.98%, 4.03%, and 4.07%, respectively. This indicates that the waveform-domain attacks evaluated here do not rely on severe transcription-level degradation.

However, the attacks still differ in their signal-level distortion and attack effectiveness. Waveform-domain BIM and PGD achieve the highest average SNR values, at 33.82 dB and 33.33 dB, respectively. FGSM also preserves relatively high SNR at 31.30 dB, while CW has a lower average SNR of 24.28 dB. In contrast, CW achieves the strongest overall source-detector and black-box attack performance. This demonstrates a clear ASR–SNR trade-off: the strongest detector-disrupting attack is not necessarily the one with the smallest signal-level distortion.

The STFT-magnitude-domain results show a different trade-off. FGSM, BIM, and CW preserve audio quality well, with WER values close to the clean baseline and high SNR, but they have low ASR in most detector settings. STFT-magnitude PGD is more effective, especially against ResNet+LFCC, but it increases average WER to 5.24% and reduces average SNR to 11.20 dB. This suggests that STFT-magnitude PGD produces stronger spectral distortion than the other STFT-magnitude attacks.

These findings show why adversarial audio evaluation should report both attack effectiveness and audio quality. Without WER and SNR, waveform-domain CW and STFT-magnitude PGD would appear strongest based mainly on ASR. With quality metrics included, the interpretation becomes more nuanced: CW provides the strongest attack performance, PGD and BIM provide stronger signal-preservation trade-offs in the waveform domain, and STFT-magnitude PGD exposes a representation-specific vulnerability in LFCC-based attribution at the cost of stronger spectral distortion. This supports the central evaluation framing of this study: adversarial robustness should not be judged from ASR alone, but from the joint relationship between attack success, transferability, intelligibility, and signal distortion.

6.4. Comparison with Prior Adversarial Robustness Studies

The findings of this study are broadly consistent with prior work showing that audio deepfake detectors and speech anti-spoofing systems can be vulnerable to adversarial perturbations. Kawa et al. [14] showed that adversarial attacks can reduce the reliability of audio deepfake detectors and that robustness differs across detector architectures. The present results support this observation: AASIST, ResNet + LFCC, LCNN + CQCC, and BiLSTM do not respond uniformly to the same perturbations, indicating that adversarial robustness is strongly architecture- and representation-dependent.

The transferability results are also consistent with studies emphasizing the importance of black-box evaluation. Farooq et al. [27] showed that adversarial examples can transfer beyond the source model and degrade the performance of independent audio deepfake detectors. Our results similarly show that source-detector performance does not fully predict black-box transferability. In particular, waveform-domain FGSM achieves only low fake-to-real ASR on AASIST but transfers substantially to ResNet + LFCC and LCNN + CQCC. This supports the broader conclusion that adversarial evaluation should include independent target detectors rather than relying only on source-detector ASR.

The observed domain-dependent behavior also aligns with Wang et al. [28], who reported that attack effectiveness against speech deepfake detectors can differ between time-domain and frequency-domain settings. Our results extend this observation by showing that domain effects are not uniform across detector representations. STFT-magnitude PGD transfers strongly to ResNet + LFCC but has limited effect on LCNN + CQCC and BiLSTM, while waveform-domain attacks produce broader transferability across feature-based detectors. This suggests that the perturbation domain interacts with both the source detector and the target detector representation.

Compared with the broad benchmark of Uddin et al. [15], which evaluates anti-forensic attacks against general audio deepfake detection systems, the present study provides a more focused analysis of universal perturbations and vocoder fingerprint attribution. Their results show that both raw-audio and spectrogram-based detectors can be vulnerable under statistical and optimization-based anti-forensic attacks. Our findings are consistent with this general vulnerability trend, but further show that attack strength should be interpreted together with speech intelligibility and signal distortion. In particular, the waveform-domain results show that attacks can preserve transcription-level intelligibility while still producing substantial black-box transferability, and that different attacks occupy different positions in the ASR–SNR–WER trade-off.

These comparisons place the present results in a broader research context. Prior work establishes that audio deepfake detectors are vulnerable to adversarial and anti-forensic perturbations. This study extends that evidence to a universal perturbation setting involving vocoder fingerprint attribution, cross-detector transferability, perturbation-domain comparison, and WER/SNR-based quality analysis.

6.5. Limitations and Generalizability

This study is conducted in a controlled single-speaker setting using synthetic speech derived from LJSpeech. This design is useful for isolating the effects of attack algorithm, perturbation domain, vocoder type, and detector representation because speaker identity, language, and recording conditions are held relatively constant. However, the findings should not be interpreted as fully representative of all audio deepfake detection scenarios. Multi-speaker corpora, multilingual speech, diverse accents, noisy recordings, compression artifacts, replay conditions, and in-the-wild social media audio may change both detector behavior and adversarial transferability. Future work should therefore evaluate whether the same robustness patterns hold across larger and more diverse datasets.

The comparison between waveform-domain and STFT-magnitude-domain attacks is also based on domain-specific perturbation budgets rather than a full exhaustive WER-matched budget sweep. The selected budgets provide quality-aware operating points for comparing ASR, WER, and SNR, but a complete WER-matched evaluation would require separately tuning the perturbation budget for each attack and perturbation domain. Such tuning may produce different ASR–SNR–WER trade-offs. Therefore, the present study reports WER and SNR alongside ASR so that attack effectiveness can be interpreted in relation to audio quality rather than from attack success alone.

This study uses Whisper WER and SNR as objective measures of transcription-level intelligibility and signal distortion. These metrics provide a reproducible way to compare attack settings, but they do not fully capture subjective perceptual quality, naturalness, annoyance, or listener preference. Subjective evaluation methods such as mean opinion score (MOS) tests [33] or Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)-style listening tests [34] would provide additional evidence about perceived audio quality. Such testing requires a separate human-subject evaluation protocol and is therefore left for future work.

Finally, this study focuses on universal perturbations rather than per-sample adversarial optimization. Prior audio deepfake adversarial studies have mainly considered input-specific or input-conditioned attack settings, including gradient-based attacks such as FGSM, PGD, BIM, C&W, and DeepFool, as well as GAN-based transferable attacks [14,15,27,28]. A per-sample attack setting would likely increase source-detector ASR because each perturbation could be optimized specifically for one utterance. However, it would evaluate a different threat model from the one considered here. The universal setting is more restrictive because a single perturbation must generalize across utterances and vocoders without re-optimization. Therefore, switching to per-sample attacks would likely change the absolute ASR and quality trade-offs, but it would not remove the main conclusion that adversarial robustness depends jointly on perturbation domain, detector representation, transferability, and audio quality.

Future work should extend this evaluation protocol to additional source detectors, defended pipelines, adaptive attacks, and explainability analysis of detector decisions and adversarial perturbations.

7. Conclusions

This paper presented a focused, quality-aware evaluation protocol of adversarial attacks against audio deepfake and vocoder fingerprint detectors. Four attacks, FGSM, BIM, PGD, and CW, were evaluated in both waveform-domain and STFT-magnitude-domain settings. All attacks were optimized against AASIST using a targeted fake-to-real objective. Attack performance was first measured on the source AASIST detector, and black-box transferability was then evaluated across detectors with different input representations, including ResNet + LFCC, LCNN + CQCC, and BiLSTM.

The results show that adversarial effectiveness depends strongly on the interaction between attack algorithm, perturbation domain, and detector representation. STFT-magnitude PGD transfers strongly to ResNet + LFCC but has limited effect on LCNN + CQCC and BiLSTM, indicating that frequency-domain vulnerability is representation-specific. In contrast, waveform-domain attacks produce broader transferability across feature-based detectors, particularly ResNet + LFCC and LCNN + CQCC. However, source-detector success does not always predict black-box transferability, as shown by waveform-domain FGSM, which achieves only low ASR against AASIST but transfers substantially to several black-box detectors.

The quality evaluation further shows that high ASR alone is insufficient for assessing adversarial audio. Under the chosen waveform-domain budget, FGSM, BIM, PGD, and CW preserve transcription-level intelligibility, with average WER values close to the clean baseline. However, the attacks differ in their ASR–SNR–WER trade-offs. CW provides the strongest overall source-detector and black-box attack performance, while BIM and PGD achieve stronger signal-preservation characteristics in the waveform domain. These findings highlight the need to evaluate adversarial attacks using both detection metrics and audio-quality metrics.

Overall, this study demonstrates that robustness claims in audio deepfake and vocoder fingerprint detection should not be based on a single attack, detector, perturbation domain, or quality metric. However, the experiments are conducted in a controlled single-speaker setting using LJSpeech, and the findings should therefore be interpreted within this scope. Future work should extend this evaluation protocol to larger and more diverse datasets, additional source detectors, more vocoder families, multilingual and multi-speaker speech, noisy or compressed audio, defended detection pipelines, adaptive attacks, and subjective listening evaluation.

Author Contributions: Conceptualization, Q.M.T.; methodology, Q.M.T.; software, Q.M.T.; formal analysis, Q.M.T.; visualization, Q.M.T.; writing—original draft preparation, Q.M.T.; writing—review and editing, W.Z. and Y.-W.C.; supervision, W.Z., Y.-W.C. and W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are derived from the publicly available LJSpeech dataset. The original LJSpeech corpus is available from its official public repository. The manuscript describes the main derivation and evaluation pipeline, including the source corpus, vocoders, perturbation domains, attack configuration, detector configuration, and evaluation protocol. The synthetic and adversarial audio evaluated in this study were generated from LJSpeech using the vocoder and adversarial attack procedures described in Sections 3 and 4. The main numerical results supporting the conclusions, including detector accuracies, attack success rates, word error rates, and signal-to-noise ratio measurements, are reported in the article. Due to the large size of the generated and adversarial waveform files and their status as derived data from a public third-party corpus and external model pipelines, the full derived audio files and additional evaluation outputs are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.5, 2026 version) to assist with language editing and structural organization. The authors have thoroughly reviewed and edited the AI-generated content and take full responsibility for the accuracy and integrity of the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kong, J.; Kim, J.; Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17022–17033.
2. Yang, G.; Yang, S.; Liu, K.; Fang, P.; Chen, W.; Xie, L. Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop, Shenzhen, China, 19–22 January 2021; pp. 492–498.
3. Mustafa, A.; Pia, N.; Fuchs, G. StyleMelGAN: An Efficient High-Fidelity Adversarial Vocoder with Temporal Adaptive Normalization. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 6034–6038.
4. Yamamoto, R.; Song, E.; Kim, J.-M. Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 6199–6203.
5. Li, M.; Ahmadiadli, Y.; Zhang, X.-P. A Survey on Speech Deepfake Detection. *ACM Comput. Surv.* **2025**, *57*, 1–38.
6. Shaaban, O.A.; Yildirim, R.; Alguttar, A.A. Audio Deepfake Approaches. *IEEE Access* **2023**, *11*, 132652–132682.
7. Sisman, B.; Yamagishi, J.; King, S.; Li, H. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *29*, 132–157.
8. Spanias, A.S. Speech Coding: A Tutorial Review. *Proc. IEEE* **1994**, *82*, 1541–1582.
9. Deng, J.; Ren, Y.; Zhang, T.; Zhu, H.; Sun, Z. VFD-Net: Vocoder Fingerprints Detection for Fake Audio. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024; pp. 12151–12155.

10. Li, F.; Chen, Y.; Liu, H.; Zhao, Z.; Yao, Y.; Liao, X. Vocoder Detection of Spoofing Speech Based on GAN Fingerprints and Domain Generalization. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–20.
11. Yan, X.; Yi, J.; Tao, J.; Wang, C.; Ma, H.; Wang, T.; Wang, S.; Fu, R. An Initial Investigation for Detecting Vocoder Fingerprints of Fake Audio. In Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Lisboa, Portugal, 14 October 2022; pp. 61–68.
12. Sun, C.; Jia, S.; Hou, S.; Lyu, S. AI-Synthesized Voice Detection Using Neural Vocoder Artifacts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 904–912.
13. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. Adversarial Attacks and Defences: A Survey. *arXiv* **2018**, arXiv:1810.00069.
14. Kawa, P.; Plata, M.; Syga, P. Defense Against Adversarial Attacks on Audio DeepFake Detection. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023; pp. 5276–5280.
15. Uddin, K.; Farooq, M.U.; Khan, A.; Malik, K.M. Adversarial Attacks on Audio Deepfake Detection: A Benchmark and Comparative Study. *arXiv* **2025**, arXiv:2509.07132.
16. Wang, X.; He, K. Enhancing the Transferability of Adversarial Attacks through Variance Tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1924–1933.
17. Jung, J.-W.; Heo, H.-S.; Tak, H.; Shim, H.-J.; Chung, J.S.; Lee, B.-J.; Yu, H.-J.; Evans, N. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23–27 May 2022; pp. 2405–2409.
18. Zhou, X.; Garcia-Romero, D.; Duraiswami, R.; Espy-Wilson, C.; Shamma, S. Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011; pp. 559–564.
19. Todisco, M.; Delgado, H.; Evans, N. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. In Proceedings of the Odyssey 2016: The Speaker and Language Recognition Workshop, Bilbao, Spain, 21–24 June 2016; pp. 283–290.
20. Demontis, A.; Melis, M.; Pintor, M.; Jagielski, M.; Biggio, B.; Oprea, A.; Nita-Rotaru, C.; Roli, F. Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks. In Proceedings of the 28th USENIX Security Symposium, Santa Clara, CA, USA, 14–16 August 2019; pp. 321–338.
21. Costa, J.C.; Roxo, T.; Proença, H.; Inacio, P.R.M. How Deep Learning Sees the World: A Survey on Adversarial Attacks and Defenses. *IEEE Access* **2024**, *12*, 61113–61136.
22. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
23. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial Examples in the Physical World. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 99–112.
24. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
25. Carlini, N.; Wagner, D. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 3–14.
26. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical Black-Box Attacks against Machine Learning. In Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates, 2–6 April 2017; pp. 506–519.
27. Farooq, M.U.; Khan, A.; Uddin, K.; Malik, K.M. Transferable Adversarial Attacks on Audio Deepfake Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, Tucson, AZ, USA, 28 February–4 March 2025; pp. 1640–1649.
28. Wang, W.E.; Salvi, D.; Negroni, V.; Leonzio, D.U.; Bestagini, P.; Tubaro, S. BIM-Based Adversarial Attacks Against Speech Deepfake Detectors. *Electronics* **2025**, *14*, 2967.
29. Moosavi-Dezfooli, S.-M.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
30. Ito, K.; Johnson, L. The LJ Speech Dataset. Available online: <https://url.au.m.mimecastprotect.com/s/ENimCD1vkkT5Mk71BIxfXcjNlnw?domain=keithito.com> (accessed on 25 May 2026).
31. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
32. Kumar, K.; Kumar, R.; de Boissiere, T.; Gustin, L.; Teoh, W.Z.; Sotelo, J.; de Brebisson, A.; Bengio, Y.; Courville, A. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14881–14892.

33. International Telecommunication Union. *Recommendation ITU-T P.800.1: Mean Opinion Score (MOS) Terminology*; International Telecommunication Union: Geneva, Switzerland, 2016.
34. International Telecommunication Union. *Recommendation ITU-R BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*; International Telecommunication Union: Geneva, Switzerland, 2015.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.