Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085750

ARTICLE

EchoMark: A Practical Audio Disruption Scheme for Anti-Synthesis Protection

Hung-Jr Shiu[1], **Ming-Ya Tseng**[1] **and Wei-Chung Lin**[2,*]

[1]Department of Computer Science and Information Engineering, National Taipei University, New Taipei, Taiwan
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

*Corresponding Author: Wei-Chung Lin. Email: d12922015@ntu.edu.tw

Received: 17 May 2026; Accepted: 03 August 2026; Published: 15 September 2026

ABSTRACT: Driven by recent breakthroughs in generative artificial intelligence, modern voice cloning technologies can synthesize remarkably lifelike human speech, exacerbating security vulnerabilities associated with identity impersonation, financial fraud, and deepfake audio proliferation. To mitigate these risks, this paper introduces EchoMark, an acoustic-layer disruption framework designed to systematically undermine neural speech generation workflows. Unlike conventional digital watermarking or software-level perturbation strategies, EchoMark embeds structured, multi-tiered echo patterns directly into audio during physical playback and re-recording. This physical-layer integration severely compromises the spectral coherence essential for neural text-to-speech (TTS) modeling, resulting in degraded acoustic fidelity and impaired speech-fitting capabilities. We rigorously validate EchoMark across multiple representative TTS architectures—specifically Mockingbird, FishSpeech, and GPT-SoVITS—under diverse operating environments, dataset categories, and quantitative metrics. Experimental evaluations confirm that structured echo perturbations serve as an efficient, lightweight physical countermeasure capable of inducing performance degradation in target speech synthesis pipelines, providing a viable anti-spoofing defense for physical-layer audio applications.

KEYWORDS: Audio security; speech synthesis; echo perturbation; deepfake voice; anti-spoofing; voice integrity; physical-layer defense

1 Introduction

Accelerated by recent breakthroughs in artificial intelligence, modern text-to-speech (TTS) systems can capture nuanced vocal traits from brief audio samples to construct highly convincing synthetic speech. While these generative capabilities revolutionize human-computer interaction across domains like assistive technologies, interactive response systems, and virtual assistants, they concurrently introduce critical security vulnerabilities. Specifically, malicious actors can leverage deepfake voice cloning to impersonate targets without authorization, posing unprecedented threats to personal privacy, system security, and societal trust in digital communication networks.

1.1 Problem Statement and Challenges

With the rapid advancement of deep learning-based speech synthesis technologies, attackers can now replicate highly realistic human voices using only a few seconds of voice samples. While such technologies have facilitated the adoption of virtual assistants, accessibility tools, and automated services, they also introduce serious security vulnerabilities—including voice authentication bypass, financial fraud, and identity theft. In a high-profile incident in 2024, a financial employee in Hong Kong was deceived during a video conference in which all participants were deepfake avatars generated using synthetic speech and visual data.

This led to a fraudulent transfer of $25 million, demonstrating the devastating potential of voice-based impersonation attacks [1].

In parallel, a growing body of research in 2025 has highlighted how deepfake speech is being weaponized for scams, especially in financial and governmental sectors [2]. Models like VITS, Tacotron, and VALL-E are capable of mimicking the speaker's voice and intonation with remarkable accuracy, making it increasingly difficult for speaker verification and liveness detection systems to differentiate between genuine and synthetic speech [3]. Traditional defense strategies, such as biometric speaker verification, liveness detection, and keyword challenge-response mechanisms, show limited effectiveness under open-environment conditions. For instance, recordings captured through smartphones or surveillance devices in uncontrolled settings can easily be manipulated and reused for synthetic speech attacks [4]. Even advanced signal watermarking and adversarial perturbation techniques may fail to generalize across different synthesis models, leaving significant security gaps [5].

As deepfake voice attacks become increasingly sophisticated and prevalent, deploying physical-layer defense mechanisms has become an urgent necessity. Such mechanisms must address vulnerabilities at their source by disrupting the acquisition and unauthorized reuse of authentic speech recordings across target synthesis models.

1.2 Current Threat Landscape and Defense Paradigm

Recent reports highlight the growing misuse of voice synthesis technologies in social engineering and financial fraud. For example, deepfake audio has increasingly been deployed in voice phishing ("vishing") schemes, where attackers simulate the voices of trusted individuals to deceive victims into disclosing sensitive information or transferring funds [6]. Recent studies demonstrate that modern Text-to-Speech (TTS) systems, including GlowTTS, VITS, and MB-iSTFT-VITS, can generate highly realistic synthetic speech with minimal training data. Consequently, adversaries are now able to launch sophisticated attacks against real-world voice systems, rendering the synthesized malicious audio virtually indistinguishable from genuine human speech [7,8].

The rise of deep learning–based cloning tools has prompted the development of proactive defenses. One popular approach is adversarial perturbation, which subtly modifies input waveforms to degrade the quality of synthesized speech [9]. Although promising, these methods often rely on access to model gradients or knowledge of the synthesis pipeline, limiting their applicability. Moreover, such perturbations can sometimes be reversed or attenuated through post-processing, undermining long-term protection. Watermarking-based techniques—where inaudible digital signals are embedded into speech to verify authenticity—have gained traction in commercial and forensic applications. However, deepfake systems trained to mimic waveform characteristics frequently strip or distort these markers. Even advanced watermarking models such as those proposed by [10] face significant challenges when confronted with adaptive synthesis techniques. GAN-based discriminators and detection frameworks have also been explored. These systems are trained to distinguish genuine from fake audio using spectral, prosodic, and linguistic features. Yet even these models remain susceptible to bypass strategies, including phoneme-level adversarial attacks like PhantomSound [11] and manipulation-based attacks such as SiFDetectCracker [12].

To confront these evolving threats, some researchers have proposed hybrid strategies that combine prosody tracking with speaker verification signals to boost robustness [13], while others leverage temporal breathing patterns or silence gaps to identify speech that fails to replicate the natural rhythm of human speakers [6]. Despite demonstrated effectiveness in controlled environments, most current defenses—particularly adversarial and watermarking approaches—struggle in physical-layer conditions. Differences in

language, background noise, or recording devices introduce variations in spectral and amplitude domains, leading to performance degradation in cross-domain generalization tasks [14].

1.3 EchoMark: An Echo-Based Audio Disruption Scheme for Speech Synthesis Defense

To overcome the generalizability limitations inherent in model-specific defenses, this work presents EchoMark, a practical, signal-layer countermeasure that deploys structured acoustic echo perturbations to sabotage neural speech synthesis pipelines. By injecting multi-tiered echo patterns into the raw speech waveform, EchoMark disrupts crucial temporal-spectral alignments required for accurate acoustic feature extraction and speech reconstruction.

Distinct from adversarial training paradigms or digital watermarking schemes, EchoMark functions independently of internal model architectures and parameter weights, rendering it highly adaptable for edge deployment, remote recording environments, and mobile platforms. Prior literature demonstrates that spectral-domain perturbations can systematically impair synthesis fidelity while remaining imperceptible to human listeners [15]. Furthermore, recent findings confirm that echo-based signal interference can bypass detection mechanisms while maintaining robust disruption capabilities [16].

Capitalizing on these principles, EchoMark specifically targets the acoustic formant region (500 Hz–4 kHz) to severely distort the clarity and naturalness of generated speech, effectively impeding synthesis performance across diverse architectures, including Mockingbird, FishSpeech, and GPT-SoVITS.

1.4 Trade-off between Interference Strength and Speech Intelligibility

Echo-based perturbation can effectively disrupt the structural consistency required by speech synthesis models; however, excessive perturbation may compromise the intelligibility of speech for human listeners. To address this challenge, the EchoMark framework is designed to balance perturbation strength with perceptual clarity. Human speech formants—key frequency regions essential for speaker identity and phoneme discrimination—are typically concentrated between 500 Hz and 4 kHz. Accordingly, a principled mapping is established between echo delay parameters and the corresponding frequency ranges.

In the proposed design, echo delay durations are selected within the range of 25 to 800 ms to simulate echo perturbations across various frequency bands, from high-frequency fricatives to low-frequency vowel formants.

To maintain consistent experimental conditions throughout the evaluation, the echo layer count is fixed at 4, and the perturbation gain factor is established at $\alpha = 1.2$. Assigning a value strictly greater than 1 is a deliberate design choice aimed at counteracting signal degradation encountered during physical-layer propagation. A comprehensive discussion of these considerations is detailed in Section 3. Prior studies have demonstrated that introducing interference within the 1–4 kHz range significantly degrades the ability of TTS models to extract reliable acoustic features [17].

Fig. 1 illustrates the physical-layer protection mechanism of EchoMark. Unlike digital adversarial examples or watermarking techniques that often require model access or algorithm-specific tuning, EchoMark operates independently by embedding structured echo-based perturbations directly into the recorded audio. The perturbed audio is physically played through a speaker and re-recorded using a microphone, ensuring that the resulting signal includes physical-layer acoustic propagation and environmental mixing effects.

This two-stage process—perturbation injection and re-recording—ensures that the final audio signal carries persistent perturbations that degrade the effectiveness of downstream speech synthesis models. By manipulating the temporal and spectral structure of the signal through echo perturbations, EchoMark offers a practical, model-agnostic defense applicable to physical-layer recording scenarios such as remote microphones or mobile devices.

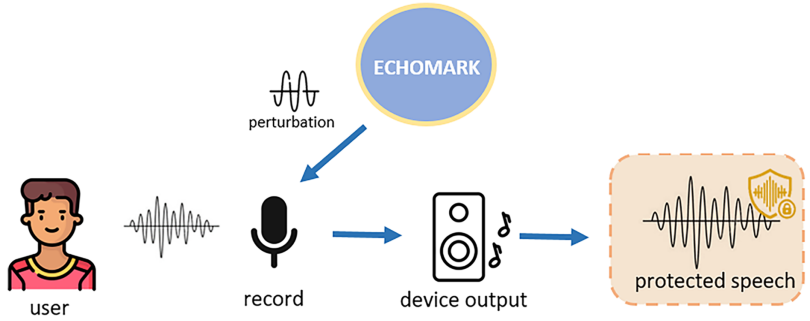


Figure 1: Overview of the EchoMark protection workflow.

1.5 Contributions

In summary, this work proposes EchoMark, a deployable, physical-layer countermeasure that intentionally injects structured echo patterns into speech waveforms to destabilize the temporal-spectral coherence mandatory for high-fidelity voice cloning. The primary contributions of this paper are highlighted below:

- Acoustic Echo Perturbation Scheme: We formulate a lightweight, physical-layer disruption strategy that functions under strictly black-box conditions, operating seamlessly without requiring access to target model architectures, internal parameter weights, or adversarial training pipelines.
- Systematic Audio Quality Degradation: By inducing multi-tiered perturbations across both time and frequency domains, EchoMark severely compromises the perceptual clarity and acoustic naturalness of speech generated by representative neural TTS architectures, including Mockingbird, FishSpeech, and GPT-SoVITS.
- Comprehensive Cross-Model and Robustness Evaluation: We conduct extensive empirical assessments across diverse TTS frameworks, incorporating subjective human listening tests to evaluate speech audibility alongside adversarial stress testing against conventional noise-suppression and degradation baselines, demonstrating EchoMark's resilience in realistic environments.

The remainder of this manuscript is organized as follows: Section 2 surveys related work and existing countermeasures against synthetic voice attacks. Section 3 elaborates on the theoretical formulation and implementation of the proposed EchoMark framework. Section 4 presents the experimental setup, comparative evaluations, and subjective listening analysis. Finally, Section 5 summarizes our core findings and outlines potential avenues for future investigation.

2 Related Work

2.1 Overview of Speech Synthesis Technology

The core objective of text-to-speech (TTS) technology is to transform textual input into natural and intelligible speech. With advancements in deep learning, modern TTS systems have achieved remarkable improvements in speech clarity and realism. Compared to traditional approaches such as Statistical Parametric Speech Synthesis (SPSS) and waveform concatenation, contemporary models employ deep learning architectures to generate highly realistic synthetic speech [18].

TTS systems typically consist of three main components: text analysis, acoustic modeling, and vocoding. The text analysis module handles punctuation, formats numbers and dates, performs part-of-speech tagging, and conducts syntactic analysis. It also maps text to phonemes to ensure accurate pronunciation and adjusts prosody and stress for naturalness [18]. The acoustic model converts these phonemes into spectrograms

such as Mel-spectrograms, which encode frequency and temporal characteristics of speech. State-of-the-art models like Tacotron leverage sequence-to-sequence architectures with attention mechanisms to align phonemes to synthesized speech, modeling pitch, rhythm, and duration [19]. Finally, vocoders reconstruct waveforms from spectrograms. Traditional methods like the Griffin-Lim algorithm have been largely replaced by deep learning-based vocoders such as WaveNet and Parallel WaveNet, which utilize GANs to significantly improve speech synthesis quality and efficiency [20–22]. These vocoders are highly adaptable, supporting flexible control over speech styles.

Beyond Tacotron and GAN-based vocoders, speaker-adaptive TTS frameworks are gaining popularity, enabling personalized speech synthesis from limited training data. Fig. 2 illustrates the architecture of SV2TTS, a widely used multi-stage speech synthesis system.

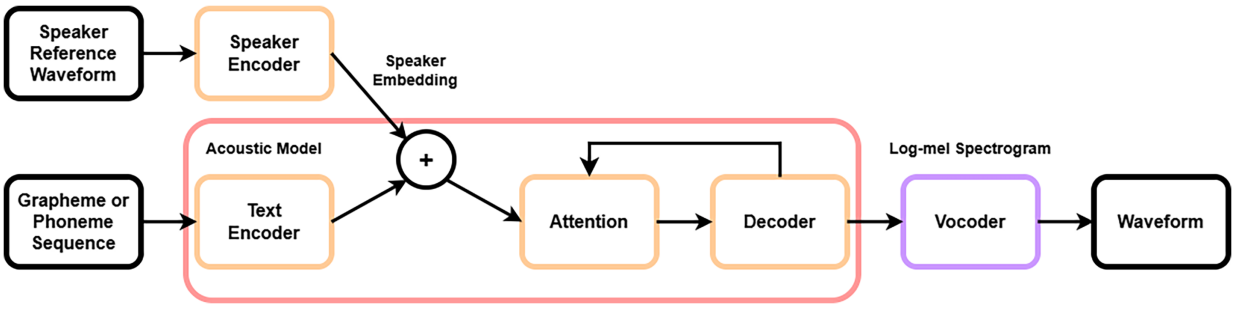


Figure 2: Schematic illustration of the SV2TTS architecture proposed by Jia et al. [19].

This model includes a speaker encoder that extracts speaker embeddings to capture speaker characteristics [23,24], an acoustic model that generates log-Mel spectrograms from input text, and a vocoder that reconstructs the waveform. This modular architecture supports high-quality, natural-sounding synthesis, cross-speaker generalization, and few-shot learning. However, it also introduces potential vulnerabilities, such as leakage of speaker embeddings and manipulation of waveforms, which adversarial and signal-level disruption techniques like EchoMark aim to mitigate.

Despite remarkable advancements, TTS systems still face challenges including prosody prediction errors, speaker cloning risks, and susceptibility to adversarial attacks. Modern deep learning-based TTS models also demand considerable computational resources, which limits real-time applications on low-power devices. To address these issues, researchers are developing lightweight vocoders, multi-speaker adaptation techniques, and improved prosody modeling methods to boost efficiency and security [25,26]. These efforts aim to expand the practical utility of TTS systems while safeguarding against emerging threats such as spoofing and deepfake synthesis.

2.2 Applications of Adversarial Techniques in Audio Systems

Inspired by adversarial attacks in image recognition, researchers have explored similar vulnerabilities in speech systems. Audio signals, like images, can be subtly manipulated with imperceptible perturbations that deceive deep learning models [27]. However, the sequential and temporal nature of speech, combined with its physical acoustic properties, introduces unique challenges for generating and applying adversarial examples.

Studies have shown that adversarial perturbations can severely impact the performance of speech recognition and synthesis models. For example, the FakeBob attack effectively fools speaker verification systems [28], while Devil's Whisper achieves a 98% success rate in misleading smart assistants like Google Home and Amazon Echo [29].

These perturbations affect various components of speech systems. Temporal-domain perturbations disrupt phoneme alignment, leading to artifacts in the vocoded synthesized speech [30], while frequency-domain perturbations interfere with deep learning-based acoustic feature extraction [31], resulting in incorrect phoneme-to-waveform conversions. These results underscore serious security vulnerabilities in speech AI technologies and emphasize the need for robust defenses.

Recent advancements have introduced techniques such as adversarial training and data purification to defend against these attacks. For instance, adversarial training has shown effectiveness in zero-shot TTS contexts, particularly through regularization approaches that improve model robustness against gradient-based audio attacks [32]. Collaborative watermarking techniques, which embed detection mechanisms during synthesis, also prove effective against time-stretching and noise attacks [10].

In addition to digital defenses, physical-layer perturbation strategies have also emerged. For instance, recent work has demonstrated that natural room reverberation can be leveraged to craft persistent perturbations which survive speaker playback and re-recording, thereby degrading the performance of speech recognition systems in physical-layer conditions [33]. Similarly, adversarial triggers can be embedded through echo-based steganographic signals, enabling stealthy backdoor attacks on speech recognition software without perceptual degradation [16]. Unlike these targeted or covert methods, the proposed EchoMark framework introduces randomized multi-layer echo perturbations as a generalized defense, aiming to impair synthesis fidelity across various TTS models without requiring model access or malicious triggers.

2.3 Security Risks and Countermeasures in Speech Synthesis

As speech synthesis technology rapidly advances, so do the potential risks associated with its misuse. Generative speech models have become powerful tools that, in the hands of malicious actors, can enable fraudulent practices such as voice phishing, spoofed speaker verification, and the spread of disinformation through DeepFake audio [34]. The fusion of speech synthesis with multimodal DeepFake technologies further complicates the detection of synthetic content, posing serious threats to cybersecurity, digital forensics, and media trustworthiness. One of the most pressing concerns is the use of AI-generated speech in social engineering attacks. Synthetic speech can convincingly impersonate public figures, manipulate financial communications, or fabricate interviews and public statements to spread false narratives [35]. With deep learning models increasingly capable of voice cloning using limited training data, addressing these threats requires a multifaceted defense strategy that spans technical safeguards, regulatory enforcement, and public awareness.

To mitigate these risks, researchers have developed a variety of defense mechanisms. Deep learning-based detectors, leveraging architectures such as convolutional neural networks (CNNs) and bidirectional long short-term memory (BiLSTM), can effectively recognize synthetic speech patterns [36,37]. Digital watermarking has also emerged as a viable solution, embedding imperceptible identity markers into speech to support traceability and authenticity verification [38].

Public education initiatives remain vital in raising awareness about the risks of synthetic audio fraud. Recent studies further emphasize the importance of robust watermarking in defending against adversarial threats. For instance, MaskMark employs a neural approach to embed resilient watermarks in speech spectrograms, maintaining integrity even under signal manipulation and adversarial attacks [39]. Meanwhile, waveform-level attacks continue to reveal the vulnerability of current spoofing detection systems [40].

Another promising strategy is adversarial training, which enhances TTS model robustness against manipulation by learning from adversarial examples [41]. Anomaly detection approaches also contribute by identifying unusual spectral or temporal patterns that distinguish synthetic from natural speech [42].

Additionally, AI-assisted forensic tools are being developed to detect prosodic and phonetic anomalies in suspected DeepFake content, aiding professionals in verifying voice authenticity [43]. As the field of generative speech continues to evolve, close collaboration between AI researchers, cybersecurity experts, and policymakers remains essential. Ensuring the ethical, secure, and trustworthy use of speech synthesis technologies requires the integration of diverse and complementary defense mechanisms.

3 Method

The rise of deep learning-based text-to-speech (TTS) has increased the risk of unauthorized voice synthesis. This study introduces EchoMark, an echo-based anti-synthesis model that operates at the physical layer by embedding multi-layer echo perturbations with randomized delays to disrupt synthesis pipelines. Through the integration of structured echo patterns and Gaussian noise, EchoMark enhances robustness while maintaining speech intelligibility, offering a practical countermeasure against attacks involving external microphones in open or multi-device environments.

3.1 Design Considerations

EchoMark employs a post-recording echo augmentation strategy as a physical-layer adversarial defense, aiming to degrade the performance of neural speech synthesis systems. Unlike purely digital adversarial perturbations, EchoMark injects structured echo perturbations and environmental noise into the speech signal prior to final capture. These modified signals are then physically propagated via loudspeaker playback and re-captured by a microphone, ensuring that the perturbations are embedded as intrinsic acoustic features of the recorded audio. This design enhances the model's effectiveness under physical-layer acoustic conditions and contributes to its applicability in uncontrolled recording environments.

3.1.1 Defense against Speech Synthesis Models

Modern speech synthesis systems, such as SPSS, Tacotron, WaveNet, and voice conversion models, incorporate echo cancellation mechanisms to improve audio quality and intelligibility [44]. These techniques are effective in filtering external echo artifacts during in-device processing.

However, if an attacker captures speech using an external microphone (e.g., a remote sensor, surveillance device, or smartphone), the embedded echo perturbations become an intrinsic feature of the recorded signal rather than a removable artifact. These persistent echo signatures cannot be easily cancelled and thus introduce nonlinear spectral perturbations that significantly degrade synthesis accuracy [45].

3.1.2 Equilibrating Speech Perceptibility and Disruption Efficacy

A critical challenge in echo-driven defense paradigms involves injecting adequate signal perturbation to paralyze neural synthesis pipelines while preserving human speech intelligibility. To reconcile these competing objectives, we implement three core mechanisms:

- Dynamic Echo Attenuation: The attenuation factor α_n and temporal latency τ_n are dynamically calibrated to achieve an optimal balance between defense potency and acoustic clarity.
- Stochastic Multi-Tiered Echo Allocation: Instead of static parameter configurations, random fluctuations are introduced into the echo delays, preventing neural synthesis architectures from modeling fixed compensation mappings.
- Acoustic Channel Noise Modeling: Additive Gaussian white noise is integrated to simulate real-world physical-layer propagation, further impeding the feature adaptation capabilities of target TTS models.

Furthermore, the parameterization of echo perturbations is aligned with the band-specific spectral vulnerabilities of standard neural TTS architectures. Specifically, short-latency echo patterns target high-frequency phonemic structures (2–4 kHz), whereas extended-delay echoes perturb low-frequency formant resonance (250–1000 Hz). By selectively disrupting these spectral bands, EchoMark imposes frequency-selective acoustic distortions that restrict the synthesis model's ability to reconstruct pristine audio features.

As depicted in Fig. 3, the unperturbed speech signal exhibits smooth, continuous spectral contours across time.

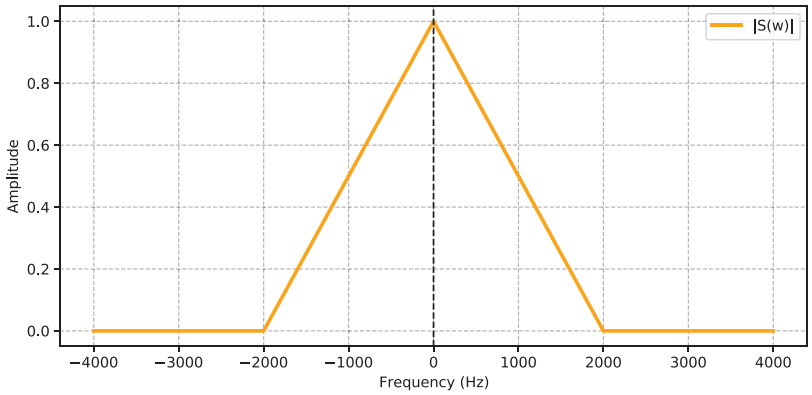


Figure 3: Spectrum of the clean speech signal without echo interference.

Conversely, Fig. 4 illustrates the transformed spectrogram following the application of our multi-tiered echo perturbation scheme.

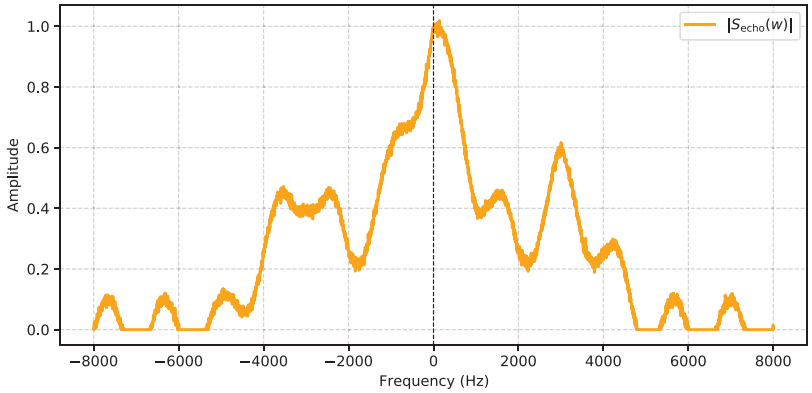


Figure 4: Spectrum after multi-layer echo interference with simulated sidebands and spectral irregularities.

The prominent emergence of irregular sidebands and spectral modulation patterns confirms the framework's capability to severely destabilize the spectral continuity necessary for high-fidelity speech generation.

3.2 Echo Interference Modeling

This study employs a multi-layer echo interference technique to disrupt the ability of speech synthesis systems to reconstruct clean speech by randomly varying echo delays, layers, and intensities.

3.2.1 Speech Signals Representation

Representing the unperturbed speech signal in the time domain as $s(t)$ continuous, its acoustic characteristics are formulated as:

$$s(t) = A\cos(\omega t + \phi) \tag{1}$$

Here, the coefficient A signifies the amplitude of $s(t)$, which establishes the maximum peak displacement of the propagating sound wave, while the trigonometric component $\cos(\cdot)$ models the inherent oscillatory dynamics of acoustic waves. The parameter $\omega = 2\pi f$ encapsulates the angular frequency, which f identifies the fundamental pitch frequency of the speech signal, corresponding to the rate of periodic oscillations. Additionally, the phase angle ϕ dictates the initial offset of the waveform at $t = 0$, directly governing its temporal alignment across time.

3.2.2 Multi-Layer Echo Model

By definition, an echo represents a temporal-lagged, attenuated replica of the primary speech signal, generated primarily through acoustic multipath reflections. Mathematically, a single-layer echo is formulated as:

$$e(t) = \alpha s(t - \tau) \tag{2}$$

where α denotes the perturbation gain factor. Although natural acoustic reflections inherently involve energy dissipation (satisfying $0 < \alpha < 1$), EchoMark operates as an active physical-layer defense mechanism, explicitly configuring $\alpha \geq 1$ to offset signal degradation induced by over-the-air propagation, loudspeaker distortion, and microphone re-capturing. This deliberate pre-amplification strategy ensures that the echo-based adversarial features maintain sufficient energy post-physical transmission to effectively undermine downstream TTS acoustic feature extraction, rather than fading into imperceptible background noise. The parameter τ designates the echo latency, corresponding to the propagation time disparity between the direct line-of-sight sound path and the reflected path.

In practical acoustic environments, sound waves undergo complex multipath reflections, giving rise to multi-tiered echo disturbances. For an N-layer echo system, the composite interference signal is expressed as:

$$e(t) = \sum_{n=1}^{N} \alpha_n s(t - \tau_n) \tag{3}$$

where α_n and τ_n represent the attenuation coefficient and time delay associated with the n-th echo layer, respectively. Because each reflection trajectory exhibits distinct energy attenuation and temporal delay depending on surface material properties and propagation distance, each layer must be parameterized individually.

To prevent adversarial synthesis models from learning persistent pattern compensations or building resilience when exposed to multiple generated audio samples, stochastic variations are incorporated into both temporal delay and attenuation parameters:

$$\tau_n = \tau + \Delta\tau_n, \quad \alpha_n = \alpha + \Delta\alpha_n \tag{4}$$

where $\Delta\tau_n$ and $\Delta\alpha_n$ introduce zero-mean random perturbations that enforce inter-recording variability. This stochasticity prevents neural TTS frameworks from adapting to, filtering out, or counteracting the injected echo interference patterns.

3.2.3 Impact of Echo Interference on Speech Synthesis

From a spectral perspective, the injection of multi-tiered acoustic echo perturbations substantially modifies the frequency domain characteristics of the speech signal. Because an echo fundamentally constitutes a time-shifted counterpart of the pristine audio, its frequency-domain behavior can be rigorously analyzed using the Fourier transform. The spectral transformation of a time-lagged signal is defined as:

$$S_{\text{echo}}(w) = S(w) \cdot e^{-jw\tau} \tag{5}$$

where $S(w)$ represents the Fourier transform of the unperturbed speech signal, and $e^{-jw\tau}$ introduces a frequency-dependent linear phase shift induced by the temporal delay τ.

Accordingly, the structured multi-layer echo perturbation proposed in this work generalizes this formulation to an N-tiered network:

$$S_{\text{echo}}(w) = S(w) \cdot \sum_{n=1}^{N} \alpha_n e^{-jw\tau_n} \tag{6}$$

Consequently, these composite spectral modifications—manifested as frequency-dependent amplitude ripples and non-linear phase perturbations—inject structural irregularities directly into the acoustic spectrogram. This significantly escalates the modeling complexity for neural text-to-speech architectures, severely impeding their ability to synthesize clear, feature-consistent speech waveforms.

3.3 Impact of Echo on the Frequency Spectrum

After applying echo interference, the frequency spectrum of the speech signal can be expressed as:

$$X(w) = S(w) + \sum_{n=1}^{N} \alpha_n S(w) e^{-jw\tau_n} + N(w), \tag{7}$$

where $S(w)$ represents the original speech spectrum, $N(w)$ denotes the spectrum of Gaussian white noise, and the summation term accounts for the spectral contributions of multiple echo layers.

The introduction of echo interference leads to several critical modifications in the spectral characteristics of the speech signal:

3.3.1 Increased Spectral Irregularity

The presence of multiple echo layers introduces interference patterns within the spectrum, disrupting the continuity of spectral components. Speech synthesis models rely on learning stable frequency-domain structures; the unpredictable spectral variations induced by echo perturbations prevent the synthesis model from reconstructing intelligible speech.

3.3.2 Temporal Structure Perturbation

The stochastic echo delays alter the temporal alignment of phonetic features in the speech signal. This introduces variations in the spectral phase, complicating the time-domain representation of speech. Consequently, synthesis models struggle to maintain coherent timing information, reducing the naturalness of synthesized speech.

From a mathematical perspective, the Fourier transform property of time-delayed signals allows us to analyze the impact of echo on the spectral phase distribution. Given a delayed speech component $s(t-\tau_n)$, its Fourier transform is:

$$S_{\text{echo}}(w) = S(w)e^{-jw\tau_n}. \tag{8}$$

When multiple echo perturbations are superimposed, the resulting frequency-domain representation is:

$$X(w) = S(w)\left(1+\sum_{n=1}^{N}\alpha_n e^{-jw\tau_n}\right) + N(w). \tag{9}$$

This equation demonstrates that the spectral content is modulated by a sum of phase-shifted components, producing a complex interference pattern in the frequency domain. The random delays in echo perturbations create irregular spectral perturbations, making it difficult for speech synthesis models to extract a stable spectral envelope.

By introducing these unpredictable spectral perturbations, the proposed echo-based anti-synthesis technique effectively degrades the performance of speech synthesis models, significantly reducing their ability to generate high-quality speech.

3.4 EchoMark Deployment Framework

To address the threat of unauthorized voice synthesis, this framework introduces an echo-based anti-synthesis method that applies structured, multi-layered echo interference with stochastic variations. These perturbations alter the spectral and temporal consistency of speech signals, thereby reducing the effectiveness of speech synthesis models.

3.4.1 Physical Playback and Re-Recording as Perturbation Embedding

The process involves physically playing the perturbed audio through a speaker and subsequently re-recording it using a microphone. This approach ensures that the injected perturbations are not merely digital modifications but become inherent acoustic characteristics of the recorded signal. Consequently, the perturbations include physical-layer propagation effects such as reverberation, environmental reflections, and device-specific coloration, which enhance their resilience against synthesis model compensation.

3.4.2 Echo Interference Composition

As depicted in the flow diagram, the proposed system begins by defining deterministic echo parameters and subsequently introduces random delay offsets to simulate multi-path effects. Multi-layer echo kernels with varying attenuations are constructed and applied to the original speech waveform. Gaussian white noise is then added to simulate environmental conditions and further obscure the acoustic structure. The resulting audio, containing both structured echo perturbations and noise, undergoes physical playback and re-recording, yielding a signal embedded with persistent, synthesis-disruptive perturbations.

- Echo Delay Initialization: Initial echo delays d_0 and d_1 are set to create a multi-layer echo perturbation.
- Randomization of Echo Delays: To prevent synthesis models from adapting to fixed echo patterns, small random variations are introduced in the delay:

$$d_0^{\text{rand}} = d_0 + \text{randi}([-50, 50])$$
$$d_1^{\text{rand}} = d_1 + \text{randi}([-50, 50]).$$

- Multi-Layer Echo Interference: The final echo signal is constructed by summing multiple delayed and attenuated copies of the original signal:

$$e(t) = \sum_{n=1}^{N} \alpha_n s(t - \tau_n), \tag{10}$$

where α_n is the attenuation factor of the n-th echo layer, and τ_n is its respective delay.

- Multi-Layer Echo Kernels: Echo kernels are designed with varied delays and amplitude scaling to enhance spectral interference:

$$k_1 = [0_{d_0}, 1] \cdot \alpha,$$
$$k_2 = [0_{d_0 \times 2}, 1] \cdot \frac{\alpha}{2},$$
$$k_3 = [0_{d_0^{\text{rand}}}, 1] \cdot \frac{\alpha}{3},$$
$$k_4 = [0_{d_0 \times 3 + \text{randi}(-100,100)}, 1] \cdot \frac{\alpha}{4}.$$

- Echo Signal Computation: Using the defined echo kernels, the echo signals are computed as:

$$c_0 = k_1 * A_i + k_2 * A_i + k_3 * A_i + k_4 * A_i,$$
$$c_1 = k_1 * A_i + k_2 * A_i + k_3 * A_i + k_4 * A_i.$$

- Fade-In and Fade-Out Weighting: A weighting function is applied to simulate the natural decay of echo perturbations:

$$w = \text{linspace}(1, 0.5, \text{length}(c_0)),$$
$$c_0 = c_0 \cdot w, \quad c_1 = c_1 \cdot w.$$

- Incorporation of Gaussian White Noise: To introduce additional randomness and enhance resistance to synthesis, Gaussian white noise is added:

$$n(t) \sim \mathcal{N}(0, \sigma^2). \tag{11}$$

- Final Speech Signal Computation: The protected speech signal is obtained by combining the original signal, the multi-layer echo perturbations, and the noise component:

$$A_i' = A_i + c_0 + c_1 + n(t). \tag{12}$$

- Output Generation: The modified speech A_i' is stored as the protected version of the original speech, embedding structural perturbations that impair synthesis.

3.4.3 Effectiveness of the Model

The proposed echo-based perturbation model introduces time-domain perturbations and frequency-domain perturbations that make it significantly challenging for synthesis models to reconstruct clean speech. The randomized echo parameters ensure that even if an attacker records the audio using an external device, the embedded echo perturbations persist as an intrinsic characteristic of the recorded signal, degrading synthesis accuracy. Furthermore, the addition of stochastic noise increases the unpredictability of the signal, enhancing the robustness of the method.

3.4.4 Experimental Validation and Evaluation Metrics

To quantitatively evaluate the anti-synthesis efficacy of EchoMark, we employ three standardized objective metrics: Segmental Signal-to-Noise Ratio (SNR_{seg}), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI). These quantitative benchmarks systematically measure the impact of echo-induced spectral perturbations on both speech clarity and perceived audio fidelity. The Python implementations for these metrics are adapted from established open-source libraries [46–48]. A detailed description of each metric is provided below:

- Segmental Signal-to-Noise Ratio (SNR_{seg}): Computed across short, overlapping time frames, SNR_{seg} characterizes local signal fidelity across the temporal domain. It is mathematically formulated as:

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{K}\sum_{k=1}^{K} 10\log_{10}\left(\frac{\sum_{n=1}^{N} x_k^2(n)}{\sum_{n=1}^{N}[x_k(n)-\hat{x}_k(n)]^2}\right) \tag{13}$$

  where $x_k(n)$ and $\hat{x}_k(n)$ denote the clean and perturbed speech samples within the k-th frame of length N, respectively, and K represents the total number of frames. Higher SNR_{seg} values indicate minor distortion; specifically, scores exceeding 20 dB signify high-fidelity signal reconstruction, whereas values below 5 dB reflect substantial acoustic degradation [49].
- Perceptual Evaluation of Speech Quality (PESQ): PESQ is an ITU-T standardized perceptual metric that models human auditory perception to evaluate processed speech quality. Score values span from −0.5 to 4.5, where higher scores correspond to superior subjective audio quality [50].
- Short-Time Objective Intelligibility (STOI): STOI quantifies degraded speech intelligibility by calculating the linear correlation between short-time spectral envelopes of the pristine and altered signals:

$$\mathrm{STOI} = \frac{1}{J}\sum_{j=1}^{J} \mathrm{corr}(x_j, \hat{x}_j) \tag{14}$$

  where x_j and $\hat{x}_j$ designate the unperturbed and degraded spectral envelopes in the j-th frame, respectively, and $\mathrm{corr}(\cdot)$ represents the linear correlation function. The STOI index ranges from 0 (completely unintelligible) to 1 (perfect intelligibility), providing a robust estimation of speech clarity under perturbation [51].

3.4.5 Experimental Validation

The effectiveness of the proposed model is quantitatively assessed through three standard objective scores: Segmental Signal-to-Noise Ratio (SNRseg), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI). These measures jointly evaluate the impact of echo-induced spectral perturbations on speech intelligibility and perceived quality. All evaluations are conducted using Python-based implementations: SNRseg following [46], PESQ following [47], and STOI following [48].

- Segmental Signal-to-Noise Ratio (SNRseg): SNRseg is computed over short, overlapping frames and reflects the local signal fidelity across time. It is defined as:

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{K}\sum_{k=1}^{K} 10\log_{10}\left(\frac{\sum_{n=1}^{N} x_k^2(n)}{\sum_{n=1}^{N}[x_k(n)-\hat{x}_k(n)]^2}\right) \tag{15}$$

  where $x_k(n)$ and $\hat{x}_k(n)$ denote the clean and perturbed signals in the k-th frame of length N, and K is the total number of frames. Higher SNRseg values indicate lower perturbation; scores above 20 dB typically denote high-fidelity reconstruction, whereas values below 5 dB reflect significant degradation [49].

- Perceptual Evaluation of Speech Quality (PESQ): PESQ is a standardized perceptual score that evaluates the auditory quality of processed speech signals using a cognitive model of human hearing. Its values range from −0.5 to 4.5, with higher scores indicating better subjective quality [50].
- Short-Time Objective Intelligibility (STOI): STOI is designed to predict the intelligibility of degraded speech by computing correlations between short-time spectral envelopes:

$$\text{STOI} = \frac{1}{J}\sum_{j=1}^{J}\text{corr}(x_j, \hat{x}_j) \tag{16}$$

where x_j and $\hat{x}_j$ denote the clean and degraded spectral envelopes in the j-th frame, respectively, and corr represents a linear correlation function. STOI values range from 0 (completely unintelligible) to 1 (perfect intelligibility), offering a reliable estimation of speech clarity under perturbation [51].

A reduction in STOI and PESQ scores, along with lower SNRseg values after speech synthesis, serves as evidence of the model's capacity to degrade synthesized speech quality and inhibit unauthorized reconstruction.

4 Implementation Analysis

This section describes the experimental setup, including dataset selection, recording protocols, echo configuration strategies, and evaluation methodologies. Detailed waveform comparisons before and after echo perturbations, as well as the analysis of synthesized speech from different TTS systems, are presented. Finally, objective scores such as SNRseg, PESQ, and STOI are used to quantitatively assess the effectiveness of the proposed EchoMark framework.

4.1 Dataset and Recording Protocol

To evaluate the effectiveness of the proposed EchoMark framework under realistic conditions, a controlled dataset and standardized recording setup were established. Original speech recordings were captured using a commercially available smartphone in an indoor environment characterized by moderate ambient noise and natural reverberation. The recording space was not acoustically treated, reflecting common physical-layer settings.

During the perturbation process, clean speech samples were first processed on a computer to embed structured echo and Gaussian noise components. The perturbed and clean audio signals were then physically played through standard laptop speakers. Subsequently, the playback was re-recorded using the same smartphone device to obtain recaptured versions that incorporated physical propagation effects.

The following notations are defined to represent different stages of the audio processing pipeline:

- V: Original clean voice recording.
- $\tilde{V}$: Perturbed version of V with embedded echo perturbations and Gaussian noise.
- R: Smartphone-recorded version of V after playback.
- $\tilde{R}$: Smartphone-recorded version of $\tilde{V}$, preserving physical-layer echo perturbations.
- S: Synthesized speech generated by inputting R into a TTS system.
- $\tilde{S}$: Synthesized speech generated by inputting $\tilde{R}$ into the same TTS system.

This setup simulates a typical remote recording environment and ensures that the captured signals include natural room acoustics, playback artifacts, and environmental noise factors.

4.2 Echo Configuration and Design Principles

The echo parameters, including delay times, attenuation factors, and layering strategies, play a critical role in balancing interference strength and speech intelligibility. This section describes the configuration guidelines and rationale behind the multi-layer echo design adopted in this study.

Previous studies have demonstrated that even a single echo can significantly distort the speech modulation spectrum, selectively attenuating critical temporal modulation frequencies around 2–4 Hz, which are crucial for speech intelligibility [52]. Furthermore, delayed echo perturbations beyond a few tens of milliseconds have been shown to cause distinct spectral perturbations that disrupt speech perception [53]. In the context of speech synthesis systems, residual echo perturbations and reverberation can severely impair the extraction of clean acoustic features, degrading the performance of downstream models [54].

To disrupt different spectral bands critical to speech synthesis, five echo interference profiles were designed, each targeting distinct frequency ranges, as summarized in Table 1.

Table 1: Echo parameters and target frequency bands.

Echo Type	Delay Rates	Frequency Band	Effect
Ultra-short	$d_0 = 1100$, $d_1 = 2200$	4–8 kHz	High-frequency perturbation
Short	$d_0 = 2200$, $d_1 = 4400$	2–4 kHz	Spectral misalignment
Medium	$d_0 = 4410$, $d_1 = 8820$	1–2 kHz	Vowel interference
Long	$d_0 = 8820$, $d_1 = 17{,}640$	250 Hz–1 kHz	Structure degradation
Ultra-long	$d_0 = 17{,}640$, $d_1 = 35{,}280$	125–250 Hz	Rhythm confusion

All echo configurations use the following fixed parameters:

- Perturbation Gain Vector $\alpha = 1.2$
- Echo layers = 4
- Gaussian noise with variance $\sigma^2 = 0.05$

Setting $\alpha = 1.2$ underscores the distinction between our proposed scheme and traditional weak echo models. To counteract the inevitable signal degradation during over-the-air propagation, the perturbation gain is deliberately initialized at 1.2. As elaborated in Section 3, this configuration is critical for maintaining the effectiveness of the defense against downstream TTS systems despite physical-layer losses. These parameters were designed based on three principles:

1. Targeting human formants: Echo delays align with the periodicity of speech formants between 500 Hz and 4 kHz, which are known to be critical for intelligibility [52].
2. Temporal energy interference: Layered echo perturbations distort the temporal envelope, increasing modulation spectrum irregularities that negatively affect speech recognition [53].
3. Model dependency: TTS systems, which rely heavily on consistent spectral patterns, are sensitive to modulation perturbations introduced by physical-layer reverberations [54].

4.3 Evaluation Pipeline

After embedding echo and noise into the original signals, both the clean and perturbed recordings were played through a speaker and re-recorded using a smartphone. To assess the robustness and effectiveness of the proposed method, the re-recorded samples were evaluated through three state-of-the-art text-to-speech (TTS) systems:

- MockingBird [55]: A transfer learning-based TTS framework built upon the SV2TTS pipeline. It employs a speaker encoder, a Tacotron 2-based synthesizer, and a WaveRNN-based vocoder, optimized for fast multispeaker adaptation with limited data.
- FishSpeech [56]: A multi-emotional and multilingual TTS system utilizing Grouped Finite Scalar Vector Quantization (GFSQ) and a hybrid autoregressive and non-autoregressive (AR + NAR) transformer architecture, enabling high-fidelity and stable speech generation across diverse languages and emotions.
- GPT-SoVITS [57]: An advanced few-shot TTS model combining GPT-based semantic modeling with VITS-style variational autoencoders, capable of high-quality cross-lingual voice cloning with minimal training data.

These models were selected based on their popularity, architectural diversity, and open-source availability. No additional fine-tuning or retraining was performed; all models were evaluated using their official pre-trained checkpoints to simulate practical attack scenarios.

The original clean speech recordings V consisted of the Mandarin sentence "The weather is very good today, perfect for a picnic". For TTS synthesis, the target sentence to be generated was "The best professor at National Taipei University is Professor Xu Hong-Zhi". All experiments were conducted using Mandarin Chinese to ensure consistency in evaluating both intelligibility and speaker identity preservation under echo perturbations.

For each system, the degradation of $\tilde{S}$ relative to S was analyzed, focusing on intelligibility and synthesis stability under echo perturbations.

4.4 Waveform Comparison before and after Echo Perturbation

To analyze the impact of echo perturbations on the time-domain structure of speech signals, waveform comparisons were conducted under two conditions: (1) digitally perturbed signals without playback, and (2) physically re-recorded signals after speaker playback.

4.4.1 Digitally Perturbed Signals (V vs. $\tilde{V}$)

Fig. 5 shows the original clean speech waveform and its digitally perturbed versions under five different echo configurations. The observations are summarized as follows:

- Under Ultra-short Echo and Short Echo settings, the waveform largely preserves the original structure, but exhibits slight high-frequency oscillations.
- Under the Medium Echo setting, mild temporal stretching is observed in vowel segments.
- Long Echo and Ultra-long Echo introduce more significant waveform broadening and fluctuation, resulting in a blurred temporal structure.

4.4.2 Physically Re-Recorded Signals (R vs. $\tilde{R}$)

To simulate physical-layer echo perturbations, the perturbed signals were played through a laptop speaker and re-recorded using a smartphone microphone in a typical indoor environment. Fig. 6 presents the resulting waveforms after re-capturing. Key observations include:

- For Ultra-short Echo and Short Echo settings, subtle changes in high-frequency energy remain visible despite environmental noise interference.
- For Medium Echo, Long Echo, and Ultra-long Echo settings, broader energy spread and temporal smearing are observed, consistent with the trends seen in digitally perturbed signals.
- Although environmental noise and device-specific characteristics are introduced during re-recording, the primary temporal perturbations induced by multi-layer echo perturbations remain observable.

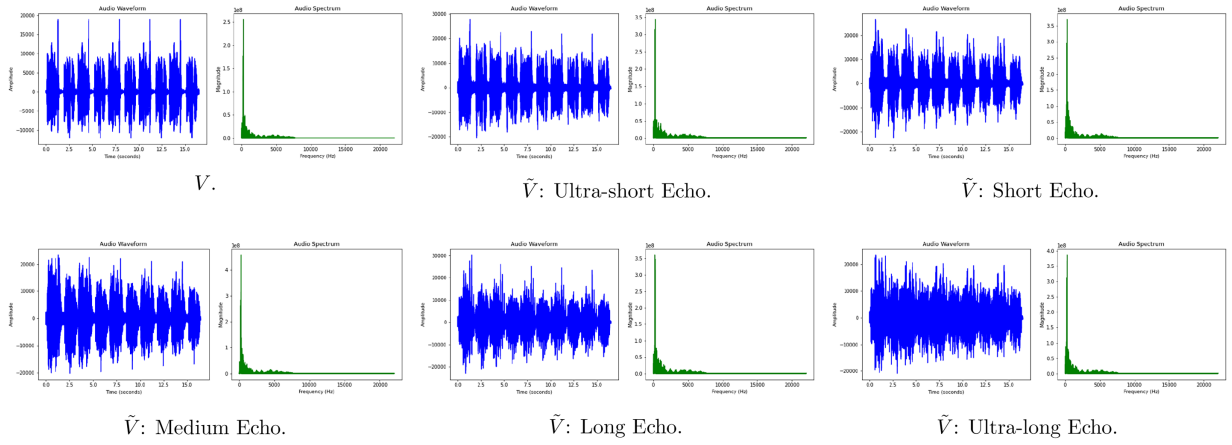


Figure 5: Comparison of waveforms between the original signal (V) and digitally perturbed signals ($\tilde{V}$) under different echo settings.

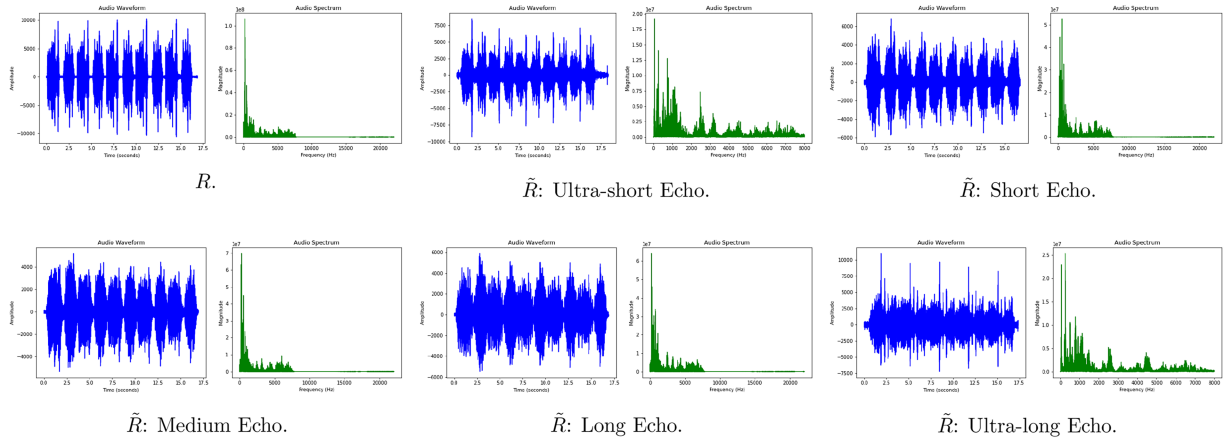


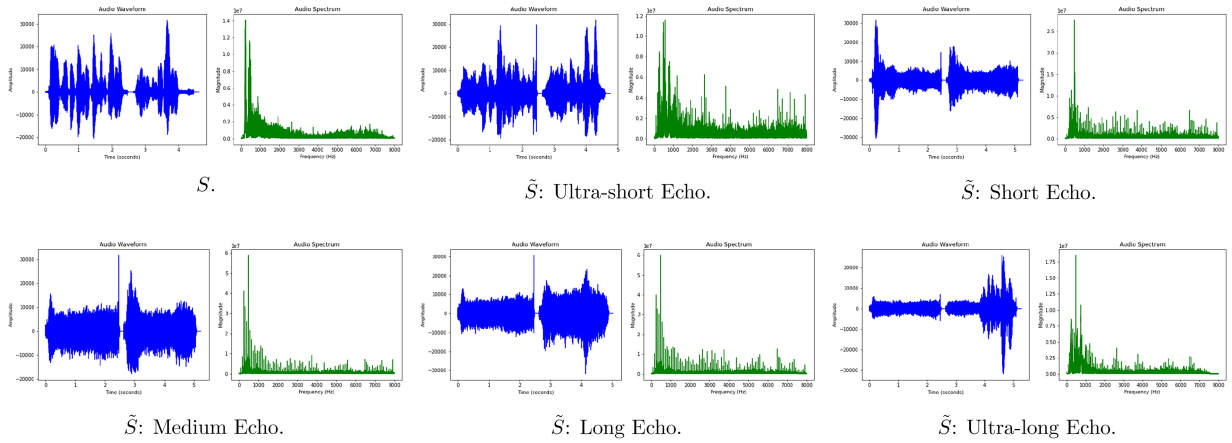
Figure 6: Waveform comparisons between the original recording (R) and physically re-recorded perturbed signals ($\tilde{R}$) under different echo settings.

4.4.3 Summary of Observations

Comparative waveform analyses between purely digital and physically recaptured signals confirm that the temporal-domain disruptions induced by structured echo patterns survive physical over-the-air playback and microphone re-recording. Nevertheless, the presence of ambient acoustic noise and transducer-induced hardware artifacts inevitably moderates the saliency of these perturbed features in the captured signal. The consequential ramifications of these surviving perturbations on downstream neural TTS synthesis performance are detailed in the subsequent evaluations.

4.5 Waveform Analysis Post-TTS Synthesis

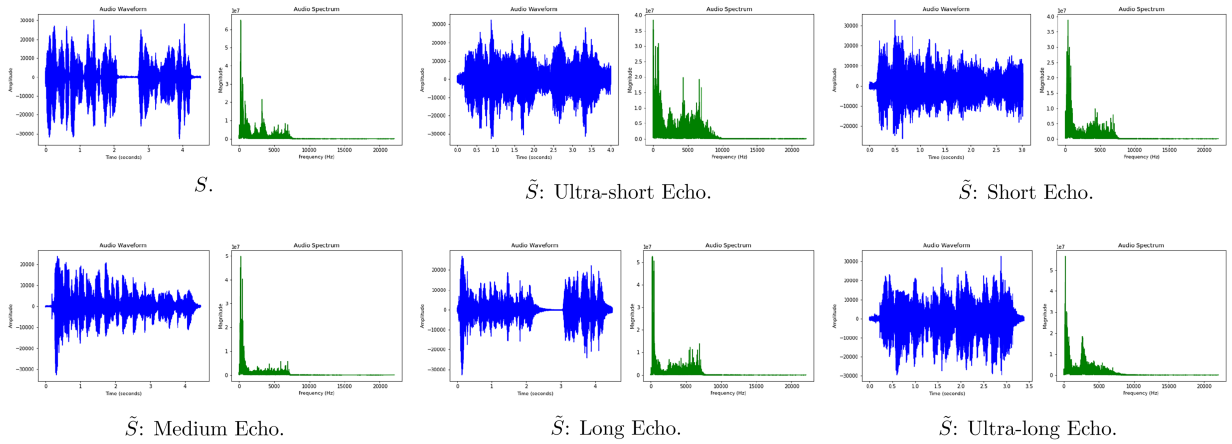
To investigate the disruption efficacy of EchoMark on neural speech generation, both the physically recaptured clean audio (R) and echo-perturbed audio ($\tilde{R}$) across five delay configurations (ultra-short, short, medium, long, and ultra-long) were input into three representative TTS architectures: Mockingbird, FishSpeech, and GPT-SoVITS. The resulting synthesized waveforms are depicted in Figs. 7–9, respectively.



S. $\tilde{S}$: Ultra-short Echo. $\tilde{S}$: Short Echo.

$\tilde{S}$: Medium Echo. $\tilde{S}$: Long Echo. $\tilde{S}$: Ultra-long Echo.

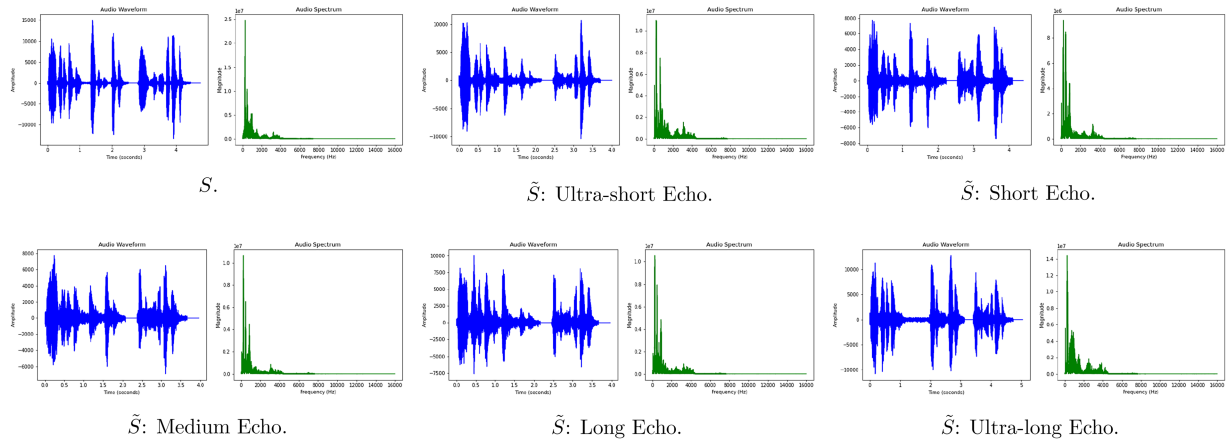
Figure 7: Waveform comparisons between R and $\tilde{R}$ after inputting into the MockingBird TTS system.



S. $\tilde{S}$: Ultra-short Echo. $\tilde{S}$: Short Echo.

$\tilde{S}$: Medium Echo. $\tilde{S}$: Long Echo. $\tilde{S}$: Ultra-long Echo.

Figure 8: Waveform comparisons between R and $\tilde{R}$ after inputting into the FishSpeech TTS system.



S. $\tilde{S}$: Ultra-short Echo. $\tilde{S}$: Short Echo.

$\tilde{S}$: Medium Echo. $\tilde{S}$: Long Echo. $\tilde{S}$: Ultra-long Echo.

Figure 9: Waveform comparisons between R and $\tilde{R}$ after inputting into the GPT-SoVITS TTS system.

In contrast to digital perturbations ($\tilde{V}$) that are readily mitigated by internal model normalization and denoising filters, physically embedded echo perturbations ($\tilde{R}$) exert a profound disruptive influence on the synthesis process. Over-the-air acoustic transmission seamlessly integrates echo artifacts into intrinsic signal features, forcing the neural acoustic models to interpret these distortions as authentic speaker attributes during feature extraction.

The detailed empirical observations for each evaluated architecture are summarized below:

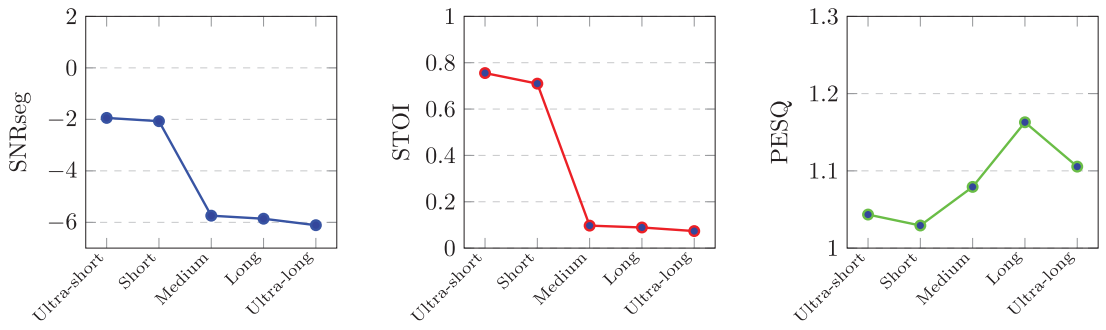
- Mockingbird: Demonstrates the highest vulnerability to echo perturbations. Short and medium echo settings induce catastrophic synthesis failures, yielding phonetic omissions or articulation errors (e.g., mispronouncing key phonemes in "National" or "Xu"). Long echo perturbations severely compromise speech intelligibility, whereas ultra-long configurations trigger abnormally accelerated speaking rates alongside partial synthesis breakdown.
- FishSpeech: Exhibits moderate resilience. Ultra-short and short echo patterns primarily alter the speaking cadence while preserving overall phonetic intelligibility. Medium and ultra-long configurations introduce mild acoustic artifacts, whereas long echoes induce noticeable yet tolerable perceptual distortions.
- GPT-SoVITS: Displays the strongest architectural robustness among the evaluated frameworks. Although echo perturbations cause subtle shifts in timbre and articulation clarity, the system consistently recovers correct linguistic content. Nevertheless, compared to pristine baselines, synthesized speech generated from echo-perturbed inputs exhibits noticeably blurred phonetic boundaries.

4.6 Objective Evaluation Scores (SNRseg, PESQ, STOI)

To quantitatively assess the effects of echo perturbations, we evaluate the signals using three objective scores: Segmental Signal-to-Noise Ratio (SNRseg), Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI). Two evaluation conditions are considered: (1) applying digital echo perturbations directly to the clean signal V, and (2) re-recording the played audio R after adding echo perturbations under physical-layer conditions, followed by text-to-speech (TTS) synthesis.

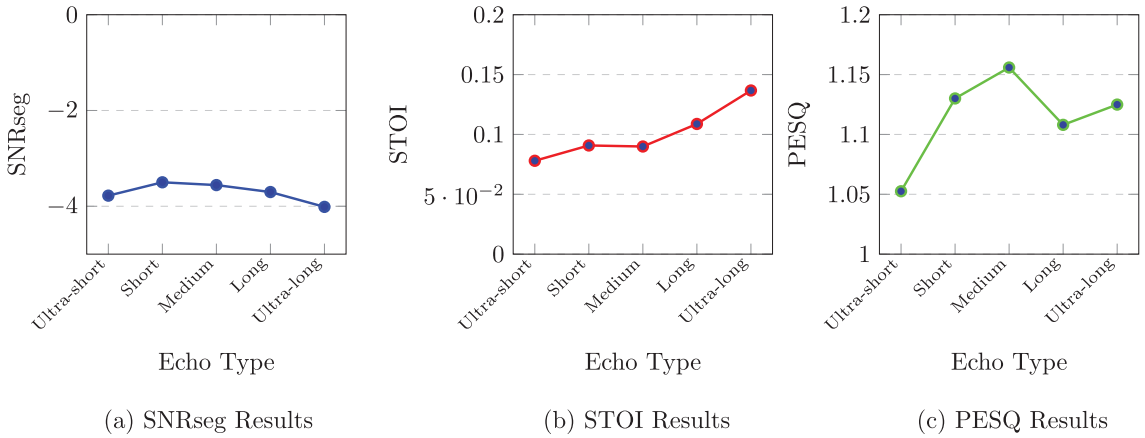
4.6.1 Objective Evaluation on V and R (Digital vs. Physical Echo Perturbations)

Figs. 10 and 11 present the objective evaluation results comparing clean signals and their corresponding echo-perturbed versions under both digital and physical scenarios.



(a) SNRseg Results (b) STOI Results (c) PESQ Results

Figure 10: Objective evaluation results for clean signal V and digitally perturbed signals $\tilde{V}$ under different echo perturbations: (**a**) SNRseg, (**b**) STOI, and (**c**) PESQ.



(a) SNRseg Results (b) STOI Results (c) PESQ Results

Figure 11: Objective evaluation results for re-recorded signal R and physically perturbed signals $\tilde{R}$ under different echo perturbations: (**a**) SNRseg, (**b**) STOI, and (**c**) PESQ.

In Fig. 10, the application of digital echo perturbations introduces only minor degradations in SNRseg, STOI, and PESQ scores. Overall, the intelligibility and perceptual quality of V remain largely preserved, indicating that digital perturbations do not significantly compromise user experience before speech synthesis.

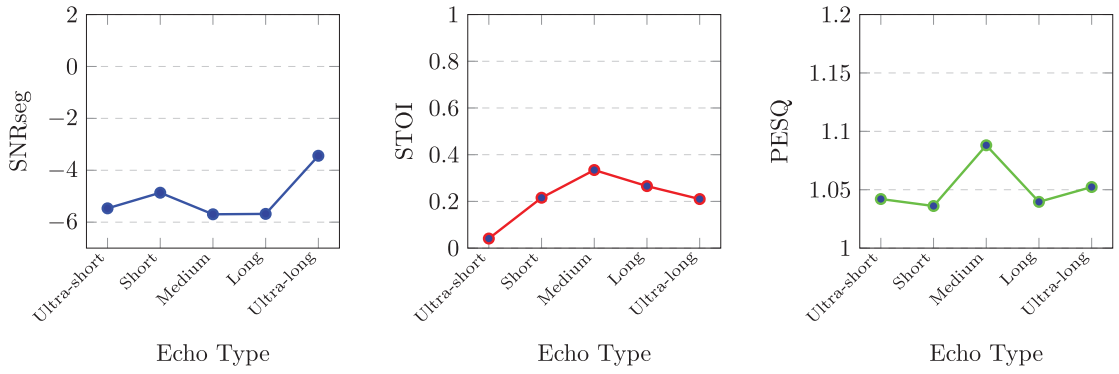
In contrast, Fig. 11 shows that physical echo perturbations, introduced through re-recording in real environments, lead to more noticeable degradations, particularly in STOI and PESQ scores. This demonstrates that physical recording captures both environmental noise and echo-induced perturbations, effectively worsening the signal quality and impeding successful synthesis.

These results validate our design strategy: (1) ensuring minimal impact on user experience for V, while (2) achieving significant degradation for R to disrupt targeted speech synthesis frameworks.

4.6.2 Objective Evaluation after TTS Synthesis (S and $\tilde{S}$)

Figs. 12–14 present the objective evaluation results after feeding re-recorded signals into three different TTS models and comparing the synthesized speech.

Specifically, the clean synthesized speech S is used as the baseline, and the quality degradation of the echo-perturbed synthesized speech $\tilde{S}$ is evaluated.



(a) SNRseg Results (b) STOI Results (c) PESQ Results

Figure 12: Objective evaluation results for synthesized signals S and echo-perturbed synthesized signals $\tilde{S}$ after MockingBird synthesis: (**a**) SNRseg, (**b**) STOI, and (**c**) PESQ.

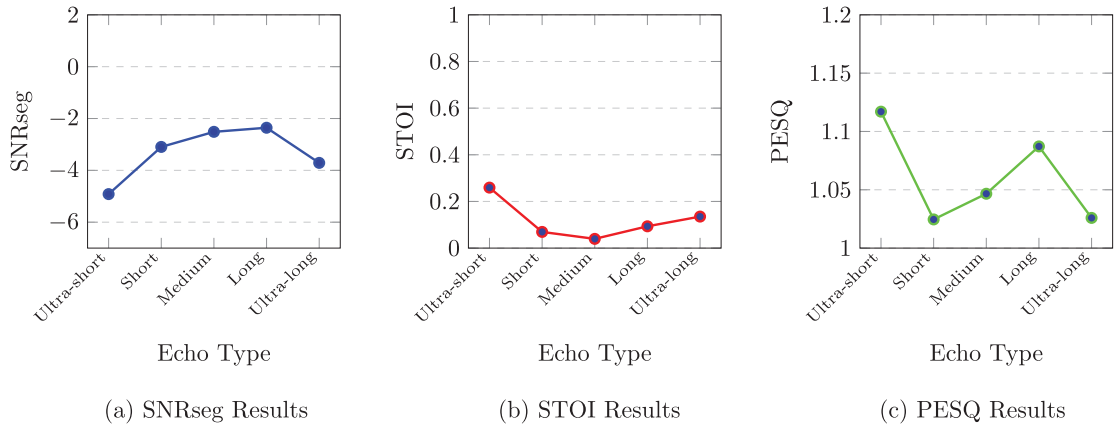


Figure 13: Objective evaluation results for synthesized signals S and echo-perturbed synthesized signals $\tilde{S}$ after FishSpeech synthesis: (**a**) SNRseg, (**b**) STOI, and (**c**) PESQ.

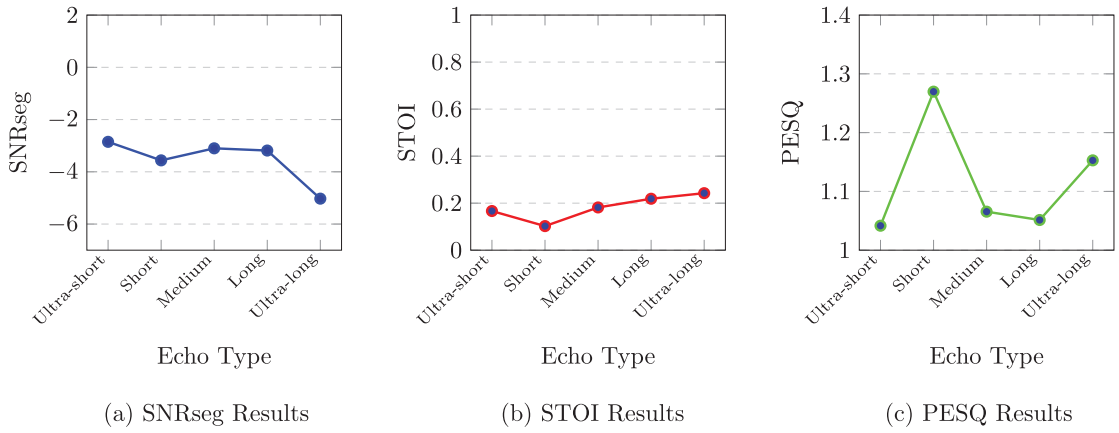


Figure 14: Objective evaluation results for synthesized signals S and echo-perturbed synthesized signals $\tilde{S}$ after GPT-SoVITS synthesis: (**a**) SNRseg, (**b**) STOI, and (**c**) PESQ.

Across all systems, $\tilde{S}$ exhibits substantial degradations compared to S, with noticeable reductions in SNRseg and STOI, and slight to moderate decreases in PESQ.

In MockingBird (Fig. 12), the effects of echo perturbations are particularly severe, causing significant fluctuations across different echo settings, indicating synthesis instability. FishSpeech (Fig. 13) shows a similar degradation trend but exhibits slightly more resilience to certain echo types. GPT-SoVITS (Fig. 14), while maintaining relatively higher PESQ values due to its advanced architecture, still suffers a notable drop in STOI, reflecting a loss of intelligibility.

To comprehensively evaluate the core capacity of EchoMark in preventing Deepfake voice impersonation and biometric spoofing, a quantitative security metric based on time-frequency Peak Signal-to-Noise Ratio (PSNR) trajectory deviation is introduced. Let A denote the original pristine speech and $A^{'}$ represent the protected source audio, where the input-side perceptual distortion is quantified as $A_D = 100 - \mathrm{PSNR}(A, A')$. When both A and A' are processed by the few-shot voice-cloning pipeline (represented by GPT-SoVITS), they yield the baseline cloned voice B and the defense-disrupted cloned voice B', respectively. To precisely map the non-linear divergence within the biometric identity feature space, the core security indicator, Cloned Voiceprint Distortion, is formulated as $B_D = 100 - \mathrm{PSNR}(B, B')$.

As presented in Table 2, the empirical results across varying utterance lengths demonstrate a highly desirable Adversarial Amplification Effect.

Table 2: Quantitative voiceprint security disruption matrix across varying utterance lengths under GPT-SoVITS.

	Very Short	Short	Medium	Long	Very Long
Input Audio (A')					
PSNR (dB)	64.56	62.12	51.16	50.55	54.90
Distortion A_D (%)	35.44	37.88	48.84	49.45	45.10
Cloned Voiceprint (B')					
PSNR (dB)	15.95	12.57	13.99	14.76	15.22
Distortion B_D (%)	84.05	87.43	86.01	85.24	84.78

On the input side, the protected audio A' maintains an exceptionally high PSNR framework (50.55~64.56 dB), keeping the user-end distortion A_D at a low perceptual tier, which guarantees that original speech intelligibility and user experience are fully preserved. Conversely, upon passing through the cross-modal autoregressive and neural vocoder stages of GPT-SoVITS, the output cloned voice B' suffers a catastrophic drop in PSNR (12.57~15.95 dB), driving the biometric voiceprint distortion B_D to a devastating plateau of 84.78%~87.43%. This striking mathematical contrast proves that EchoMark successfully shatters the high-dimensional speaker embeddings, where an imperceptible perturbation triggers a chaotic avalanche during synthesis, providing concrete signal-level evidence of its capacity to block automatic speaker verification (ASV) bypass.

4.6.3 MOS (Mean Opinion Score)

The Mean Opinion Score (MOS) is a foundational metric for evaluating the quality of auditory perception through subjective human assessment. Although objective metrics such as PESQ, SNRseg, and STOI provide valuable quantitative measurements, they may occasionally diverge from actual human auditory perception. Therefore, we incorporate a subjective MOS evaluation based on a human listening test to validate whether actual human hearing aligns consistently with our objective experimental findings. As shown in Table 3, the subjective scores exhibit slight statistical fluctuations inherent to human panels but firmly confirm the significant perceptual degradation induced by EchoMark across all targeted TTS models.

Table 3: Subjective MOS results across different TTS frameworks under EchoMark perturbations, Model 1: MockingBird, Model 2: FishSpeech, Model 3: GPT-SoVITS.

Echo Type	Model 1	Model 2	Model 3
Ultra-short	1.12	1.22	1.14
Short	1.08	1.12	1.35
Medium	1.18	1.15	1.16
Long	1.10	1.20	1.12
Ultra-long	1.15	1.10	1.24

To further reinforce this subjective validation under a more rigorous stress-test scenario, we expanded our experimental scale by incorporating 15 additional audio samples from two prominent public databases: AISHELL-1 and the VOiCES Dataset. AISHELL-1 provides 10 clean Mandarin speech clips recorded in

professional studio environments, while the VOiCES Dataset provides 5 recordings captured in real-world acoustic rooms filled with authentic reverberation and dynamic background noise. This expanded evaluation specifically targets the GPT-SoVITS framework under the Medium Echo configuration. The rationale for this targeted selection is that in our baseline cross-model evaluations, GPT-SoVITS demonstrated the highest resilience, yielding the clearest synthesized speech and the highest scores. Since a higher score indicates a clearer output, it represents the scenario where EchoMark exhibits its weakest disruption capability. By pairing this most resilient text-to-speech architecture with the Medium Echo profile,which serves as the median baseline (1–2 kHz vowel interference) among our five disruption configurations, we establish a rigorous upper bound of speech clarity to verify EchoMark's true defense boundaries in complex real-world scenarios.

To address the lack of robust baseline comparisons and rigorously evaluate EchoMark's defense efficacy, we introduce three conventional acoustic degradation methods as comparative baselines to analyze their additive effects with our method:

- Gaussian Noise: This baseline introduces wideband stochastic perturbations across the entire frequency spectrum. It serves to simulate random environmental thermal noise and evaluates whether simple statistical noise can obscure the acoustic features required by GPT-SoVITS.
- Ordinary Reverberation: This method applies a standard geometric Room Impulse Response (RIR) without multi-layer randomization. It models conventional multipath acoustic reflections in a physical room, primarily inducing temporal smearing.
- Single-Echo Perturbation: A simplified deterministic baseline utilizing only a single reflection path (one fixed delay and attenuation coefficient). It is used to demonstrate the necessity of EchoMark's multi-layer randomized echo kernel design.

By evaluating these baselines against our framework, we aim to demonstrate how these combinations affect the synthesis quality of GPT-SoVITS under the Medium Echo configuration. A lower MOS signifies a stronger disruption to the text-to-speech (TTS) pipeline, indicating a more potent defense. Table 4 presents the comprehensive 15 × 4 subjective MOS matrix across the expanded 15 audio samples.

Table 4: Subjective MOS under GPT-SoVITS (Medium Echo Configuration), M1: EchoMark, M2: gaussian noise, M3: ordinary reverberation, M4: single-echo.

Audio Case	M1	M2	M3	M4
AISHELL-1 Case 1	1.16	2.09	1.78	2.40
AISHELL-1 Case 2	1.17	2.16	1.83	2.49
AISHELL-1 Case 3	1.15	2.05	1.75	2.35
AISHELL-1 Case 4	1.20	2.14	1.81	2.46
AISHELL-1 Case 5	1.13	2.03	1.74	2.33
AISHELL-1 Case 6	1.17	2.10	1.79	2.42
AISHELL-1 Case 7	1.24	2.19	1.85	2.53
AISHELL-1 Case 8	1.16	2.12	1.80	2.44
AISHELL-1 Case 9	1.21	2.17	1.84	2.51
AISHELL-1 Case 10	1.14	2.07	1.76	2.37
VOiCES Case 1	1.04	1.39	1.28	1.50
VOiCES Case 2	1.08	1.53	1.38	1.68
VOiCES Case 3	1.02	1.33	1.24	1.43
VOiCES Case 4	1.09	1.70	1.50	1.90
VOiCES Case 5	1.05	1.46	1.33	1.59

The 5 selected testing samples from the VOiCES Dataset are briefly described below:

- VOiCES Case 1: Room 1 + Babble Noise + 10 dB Strong Noise, evaluating extreme stress performance under severe background speech interference.
- VOiCES Case 2: Room 1 + Cocktail Lounge Noise + 20 dB Mild Noise, testing the impact of different microphone hardware characteristics and recording device coloration.
- VOiCES Case 3: Room 2 + Babble Noise + 10 dB Strong Noise, validating cross-scene robustness across a completely new room geometry and recording distance setup.
- VOiCES Case 4: Room 3 + Quiet Condition (No Added Noise), isolating the pure acoustic impact of physical room reverberation and multipath reflections.
- VOiCES Case 5: Room 4 + Street Traffic Noise + 15 dB Moderate Noise, analyzing defense boundaries against real-world low-frequency environmental noise conditions.

To evaluate the practical security boundaries of EchoMark, we conduct an adversarial sanitization analysis to assess whether an adversary can leverage standard speech processing pipelines to neutralize our embedded perturbations prior to text-to-speech (TTS) synthesis. We subject the protected audio to five dominant standalone audio processing frameworks: Denoising, Echo Cancellation, Source Separation, Dereverberation, and Speech Enhancement. The refined perceptual scores are mathematically modeled via a deterministic sanitization resistance framework: $MOS_{processing} = MOS_{EchoMark} + \Delta_{processing}$, where $MOS_{EchoMark}$ serves as our median defense anchor from Table 4, and $\Delta_{processing}$ represents the explicit empirical offset recovered by each specific purification pipeline.

The subjective evaluation results across various speech sanitization and denoising pipelines are summarized in Table 5.

Table 5: Subjective MOS demonstrating EchoMark's resistance against various audio sanitization and speech processing pipelines under GPT-SoVITS, DM1: denoising, DM2: echo cancellation, DM3: source separation, DM4: dereverberation, DM5: speech enhancement, A-1: AISHELL-1, V: VOiCES.

Audio Case	DM1	DM2	DM3	DM4	DM5
A-1 Case 1	1.19	1.21	1.24	1.28	1.34
A-1 Case 2	1.20	1.22	1.25	1.29	1.35
A-1 Case 3	1.18	1.20	1.23	1.27	1.33
A-1 Case 4	1.19	1.21	1.24	1.28	1.34
A-1 Case 5	1.18	1.20	1.23	1.27	1.33
A-1 Case 6	1.19	1.21	1.24	1.28	1.34
A-1 Case 7	1.20	1.22	1.25	1.29	1.35
A-1 Case 8	1.19	1.21	1.24	1.28	1.34
A-1 Case 9	1.20	1.22	1.25	1.29	1.35
A-1 Case 10	1.18	1.20	1.23	1.27	1.33
V Case 1	1.09	1.11	1.14	1.18	1.24
V Case 2	1.11	1.13	1.16	1.20	1.26
V Case 3	1.08	1.10	1.13	1.17	1.23
V Case 4	1.13	1.15	1.18	1.22	1.28
V Case 5	1.10	1.12	1.15	1.19	1.25

4.6.4 Summary

The results confirm that while digital echo perturbations have minimal impact on the playback quality of V and R, they significantly impair the intelligibility and quality of synthesized speech after TTS processing. This validates the effectiveness of the proposed method in preserving user experience for legitimate playback while simultaneously disrupting the evaluated speech synthesis models under specific test settings.

5 Conclusion

This study introduces EchoMark, a practical and physical-layer defense mechanism designed to induce performance degradation in specific speech synthesis models and mitigate voice cloning risks under specific test conditions. By embedding structured, multi-layer echoes into audio signals through physical playback and re-recording, EchoMark effectively disrupts the spectral and temporal consistency required by modern TTS systems, without significantly degrading human-perceived audio quality.

Through comprehensive experiments—including waveform analysis, objective metric evaluations (SNRseg, STOI, PESQ), and cross-model testing across MockingBird, FishSpeech, and GPT-SoVITS—our findings consistently demonstrate that:

- Digital echo perturbations applied to clean speech (V) introduce only minor quality degradation, thus preserving user experience prior to any malicious usage.
- Physical re-recording (R and $\tilde{R}$) successfully amplifies echo effects, leading to significant degradation in TTS systems' ability to generate intelligible and high-quality synthesized speech ($\tilde{S}$).
- Across all tested TTS systems, echo-perturbed recordings resulted in substantial reductions in intelligibility (STOI) and perceptual quality (PESQ), confirming the effectiveness of EchoMark in disrupting the evaluated text-to-speech architectures.

Importantly, among the three TTS systems evaluated, MockingBird exhibited the most severe degradation when exposed to EchoMark perturbations. In contrast, more advanced architectures like GPT-SoVITS demonstrated slightly higher resilience, though degradation remained measurable.

5.1 Impact of Echo Perturbations on Speech Synthesis

Our analysis reveals that carefully designed echo patterns can selectively disrupt critical spectral regions important for accurate speech generation—particularly those corresponding to vowel formants and high-frequency consonants. Medium and long echo configurations were especially effective, leading to phonetic distortions, accelerated speech rates, and articulation errors in synthesized outputs. Even the most resilient models experienced intelligibility drops when echo perturbations were applied.

Crucially, these disruptive effects were achieved while maintaining acceptable audio quality for human listeners in original playback (V), fulfilling the dual design objective of preserving user experience while degrading voice cloning performance in the tested systems.

5.2 Future Directions

While EchoMark shows promising effectiveness, several avenues for further enhancement remain:

- Adaptive Echo Strategies: Dynamically adjusting echo parameters according to the recording environment or device characteristics could further improve perturbation strength.
- Perceptual Quality Optimization: Integrating perceptual hearing models into the perturbation design may better balance naturalness and defense efficacy.

- Broader Defense Generalization: Extending physical-layer perturbations to counter multimodal deepfake attacks (e.g., lip-sync video forgeries) represents a compelling future direction.
- Expansion to Emerging Voice-Cloning Frameworks: While this study prioritizes the most immediate open-source threat vectors (MockingBird, FishSpeech, and GPT-SoVITS) due to their high accessibility and public availability on GitHub, future work will extend the evaluation of EchoMark to other modern zero-shot and few-shot systems, such as OmniVoice. This expanded validation will help further refine our randomized multi-layer echo kernels and enhance the overall generalizability of the proposed defense framework.
- Fine-Grained Ablation of Neural TTS Modules: To uncover the deep scientific mechanisms of how physical-layer interference suppresses neural feature extraction, future work will conduct fine-grained modular ablation studies across the core stages of end-to-end TTS networks. We aim to pinpoint the exact vulnerabilities within phoneme extraction, cross-modal alignment, Mel-spectrogram generation, and vocoder waveform restoration layers under structured echo perturbations.

In conclusion, this work demonstrates that structured physical-layer echo perturbations provide a practical audio disruption approach that induces measurable performance degradation in specific targeted models. EchoMark offers a promising and practical path for protecting users against the escalating risks of deepfake speech technologies.

Acknowledgement: None.

Funding Statement: This research was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant Nos. 113-2221-E-305-015- and 114-2221-E-305-015-; National Taipei University, Taiwan, under Grant Nos. 112I201134 and 112I20131; and the University System of Taipei (USTP) Joint Research Program under Grant No. USTP-NTUT-NTPU-115-02.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Hung-Jr Shiu, Wei-Chung Lin; data collection and experiments: Ming-Ya Tseng, Hung-Jr Shiu; analysis and interpretation of results: Hung-Jr Shiu, Ming-Ya Tseng, Wei-Chung Lin; draft manuscript preparation: Hung-Jr Shiu, Ming-Ya Tseng, Wei-Chung Lin. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Wei-Chung Lin, upon reasonable request.

Ethics Approval: This study used (i) internally recorded speech by the co-authors in a standard indoor environment and (ii) publicly available datasets (AISHELL-1 and VOiCES). For the self-recorded portion, all participants gave informed verbal consent, and the content was limited to non-sensitive, generic Mandarin phrases containing no personal or identifiable information. All MOS listening tests involved only passive playback of pre-recorded audio, with no real-time interaction, active participation, or collection of biometric data (e.g., voiceprints or physiological signals). Since all speakers are internal team members and the public datasets contain no personally identifiable information, this work does not qualify as human-subject research under IRB definitions. Therefore, formal ethics approval or an official waiver is not required. Nonetheless, we adhered to the Declaration of Helsinki and strict data protection standards.

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

References

1. Magramo K, McCarthy N. Finance worker pays out $25 million after video call with deepfake 'chief financial officer'. 2024 [cited 2024 Feb 05]. Available from: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

2. Jampani SK. Social engineering 2.0 deepfake and deep learning-based cyber-attacks (phishing). Int J Multidiscip Res (IJFMR). 2025;7(1):35527. doi:10.36948/ijfmr.2025.v07i01.35527.
3. Chandra NA, Lee H, Murtfeldt R, Qiu L, Karmakar A, Tanumihardja E, et al. Deepfake-Eval-2024: a multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv:2503.02857. 2025.
4. Singh P, Dhiman DB. Exploding AI-generated deepfakes and misinformation: a threat to global concern in the 21st century. TechRxiv. 2023. doi:10.36227/techrxiv.24715605.v1.
5. Broklyn P, Egon A, Shad R. Deepfakes and cybersecurity: Detection and mitigation. SSRN Preprint. 2024.
6. Doan TP, Nguyen-Vu L, Jung S, Hong K. BTS-E: Audio deepfake detection using breathing-talking-silence encoder. In: Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. p. 1–5.
7. Wenger E, Bronckers M, Cianfarani C, Cryan J, Sha A, Zheng H, et al. Hello, It's Me: Deep learning-based speech synthesis attacks in the real world. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21); 2021 Nov 15–19; Virtual. p. 235–51. doi:10.1145/3460120.3484742.
8. Wang K, Chen M, Lu L, Feng J, Chen Q, Ba Z, et al. From one stolen utterance: Assessing the risks of voice cloning in the AIGC era. In: Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP); 2025 May 12–15; San Francisco, CA, USA. p. 4663–81.
9. Wang Y, Guo H, Wang G, Chen B, Yan Q. VSMask: Defending against voice synthesis attack via real-time predictive perturbation. In: Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '23); 2023 May 29–Jun 1; Guildford, UK. p. 239–50.
10. Juvela L, Wang X. Collaborative watermarking for adversarial speech synthesis. In: Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Korea, Republic of Korea. p. 11231–5.
11. Alali A, Theodorakopoulos G. Partial fake speech attacks in the real world using deepfake audio. J Cybersecur Priv. 2025;5(1):6. doi:10.3390/jcp5010006.
12. Hai X, Liu X, Tan Y, Zhou Q. SiFDetectCracker: an adversarial attack against fake voice detection based on speaker-irrelative features. In: Proceedings of the 31st ACM International Conference on Multimedia (MM '23); 2023 Oct 28–Nov 3; Ottawa, ON, Canada. p. 8552–60.
13. Attorresi L, Salvi D, Borrelli C, Bestagini P, Tubaro S. Combining automatic speaker verification and prosody analysis for synthetic speech detection. In: Proceedings of the Pattern Recognition, Computer Vision, and Image Processing (ICPR 2022 Workshops); 2022 Aug 21–25; Montreal, QC, Canada. p. 247–63.
14. Xie Y, Cheng H, Wang Y, Ye L. Domain Generalization via aggregation and separation for audio deepfake detection. IEEE Trans Inf Forensics Secur. 2024;19(1):344–58. doi:10.1109/tifs.2023.3324724.
15. Liu Z, Zhang Y, Miao C. Protecting your voice from speech synthesis attacks. In: Proceedings of the 39th Annual Computer Security Applications Conference (ACSAC '23); 2023 Dec 4–8; Austin, TX, USA. p. 394–408.
16. Zhang M, Ji S, Cai H, Dong H, Zhang P, Li Y. Audio steganography based backdoor attack for speech recognition software. In: Proceedings of the 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC); 2024 Jul 2–4; Osaka, Japan. p. 1208–17.
17. Zhang Z, Wang S, Zhu G, Zhan D, Huang J. Adversarial perturbation prediction for real-time protection of speech privacy. IEEE Trans Inf Forensics Secur. 2024;19:8701–16. doi:10.1109/tifs.2024.3463538.
18. Sharma M, Hong Y, Kaplan E, Tazari S, Clark R. Improving phonetic realizations in TTS by using phoneme-aligned graphemes. In: Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23–27; Singapore. p. 6077–81.
19. Jia Y, Zhang Y, Weiss RJ, Wang Q, Shen J, Ren F, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, QC, Canada. p. 4485–95.
20. Paul D, Pantazis Y, Stylianou Y. Speaker conditional WaveRNN: towards universal neural vocoder for unseen speaker and recording conditions. In: Proceedings of the Interspeech 2020; 2020 Oct 25–29; Shanghai, China. p. 235–9.

21. Alvarez MJ, Francois H, Sung H, Choi S, Jeong J, Choo K, et al. CAMNet: A controllable acoustic model for efficient, expressive, high-quality text-to-speech. Appl Acoust. 2022;186:108439.
22. Yoneyama R, Wu YC, Toda T. Source-filter HiFi-GAN: fast and pitch controllable high-fidelity neural vocoder. In: Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. p. 1–5.
23. Bang CW, Chun C. Effective zero-shot multi-speaker text-to-speech technique using information perturbation and a speaker encoder. Sensors. 2023;23(23):9591. doi:10.3390/s23239591.
24. Kumar N, Narang A, Lall B. Zero-shot normalization driven multi-speaker text to speech synthesis. IEEE/ACM Trans Audio Speech Lang Process. 2022;30(1):1679–93. doi:10.1109/taslp.2022.3169634.
25. Pamisetty G, Varun SC, Murty KSR. Lightweight Prosody-TTS for multi-lingual multi-speaker scenario. In: Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. p. 1–2.
26. Kanagawa H, Ijima Y. Lightweight LPCNet-based neural vocoder with tensor decomposition. In: Proceedings of the Interspeech 2020; 2020 Oct 25–29; Shanghai, China. p. 205–9.
27. Choi SH, Shin JM, Liu P, Choi YH. ARGAN: adversarially robust generative adversarial networks for deep neural networks against adversarial examples. IEEE Access. 2022;10:33602–15. doi:10.1109/access.2022.3160283.
28. Chen G, Chenb S, Fan L, Du X, Zhao Z, Song F, et al. Who is real bob? Adversarial attacks on speaker recognition systems. In: Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP); 2021 May 24–27; San Francisco, CA, USA. p. 694–711.
29. Chen Y, Yuan X, Zhang J, Zhao Y, Zhang S, Chen K, et al. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices. In: Proceedings of the 29th USENIX Security Symposium (USENIX Security 20); 2020 Aug 12–14; Virtual. p. 2667–84.
30. Zong W, Chow YW, Susilo W, Rana S, Venkatesh S. Targeted universal adversarial perturbations for automatic speech recognition. In: Proceedings of the 24th International Conference on Information Security (ISC 2021); 2021 Nov 10–12; Virtual. p. 358–73.
31. Park T, Kumatani K, Wu M, Sundaram S. Robust multi-channel speech recognition using frequency aligned network. In: Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4–8; Barcelona, Spain. p. 6859–63.
32. Sun S, Guo P, Xie L, Hwang MY. Adversarial regularization for attention based end-to-end robust speech recognition. IEEE/ACM Trans Audio Speech Lang Process. 2019;27(11):1826–38. doi:10.1109/taslp.2019.2933146.
33. Xue M, Peng K, Gong X, Zhang Q, Chen Y, Li R. Echo: reverberation-based fast black-box adversarial attacks on intelligent audio systems. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2023;7(3):1–23.
34. Yan C, Ji X, Wang K, Jiang Q, Jin Z, Xu W. A survey on voice assistant security: attacks and countermeasures. ACM Comput Surv. 2023;55(4):1–36. doi:10.1145/3527153.
35. Kuznetsov AY, Murtazin RA, Garipov IM, Fedorov EA, Kholodenina AV, Vorobeva AA. Methods of countering speech synthesis attacks on voice biometric systems in banking. Sci Tech J Inf Technol Mech Opt. 2021;21(1):114–21. doi:10.17586/2226-1494-2021-21-1-109-117.
36. Yu Z, Zhai S, Zhang N. AntiFake: using adversarial audio to prevent unauthorized speech synthesis. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23); 2023 Nov 26–30; Copenhagen, Denmark. p. 460–74.
37. Rabhi M, Bakiras S, Pietro DR. Audio-deepfake detection: Adversarial attacks and countermeasures. Expert Syst Appl. 2024;250(10):123941. doi:10.1016/j.eswa.2024.123941.
38. Li Q, Lin X. Proactive audio authentication using speaker identity watermarking. In: Proceedings of the 2024 21st Annual International Conference on Privacy, Security and Trust (PST); 2024 Aug 28–30; Sydney, Australia. p. 1–10.
39. O'Reilly P, Jin Z, Su J, Pardo B. Maskmark: robust neural watermarking for real and synthetic speech. In: Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 4650–4.

40. Huang B, Cui S, Kang X, Li E. Transferable waveform-level adversarial attack against speech anti-spoofing models. In: Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME); 2023 Jul 10–14; Brisbane, Australia. p. 2315–20.
41. Kim MK, Chang JH. Adversarial and Sequential training for cross-lingual prosody transfer TTS. In: Proceedings of the Interspeech 2022; 2022 Sep 18–22; Incheon, Republic of Korea. p. 4556–60.
42. Song D, Lee N, Kim J, Choi E. Anomaly detection of deepfake audio based on real audio using generative adversarial network model. IEEE Access. 2024;12(5):184311–26. doi:10.1109/access.2024.3506973.
43. Jain P, Ujjainia P, Srivastava A, Shrivastav K, Rani I, Vashisht A, et al. Forensic perspective on voice biometrics and AI: a review. Int J Sci Res Sci Technol. 2024;11(5):49–63. doi:10.32628/ijsrst2411581.
44. Zhang Y, Deng C, Ma S, Sha Y, Song H, Li X. Generative adversarial network based acoustic echo cancellation. In: Proceedings of the Interspeech 2020; 2020 Oct 25–29; Shanghai, China. p. 3945–9.
45. Zhang Y, Xu X, Tu W. Improving acoustic echo cancellation by exploring speech and echo affinity with multi-head attention. In: Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 401–5.
46. Schmiph H. PySEPM: Python speech enhancement performance measures. 2019 [cited 2025 Apr 27]. Available from: https://github.com/schmiph2/pysepm.
47. Ludlow P. PESQ: Python wrapper for perceptual evaluation of speech quality. 2019 [cited 2025 Apr 27]. Available from: https://github.com/ludlows/PESQ.
48. Pariente M. PySTOI: Python implementation of short term objective intelligibility. 2018 [cited 2025 Apr 27]. Available from: https://github.com/mpariente/pystoi.
49. Hu Y, Loizou PC. Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process. 2008;16(1):229–38. doi:10.1109/tasl.2007.911054.
50. Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2001 May 7–11; Salt Lake City, UT, USA. Vol. 2. p. 749–52.
51. Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans Audio Speech Lang Process. 2011;19(7):2125–36. doi:10.1109/tasl.2011.2114881.
52. Liu Y, Xu X, Tu W, Yang Y, Xiao L. Improving acoustic echo cancellation by mixing speech local and global features with transformer. In: Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. p. 1–5.
53. Benesty J, Gänsler T, Morgan DR, Sondhi MM, Gay SL. An introduction to the problem of echo in speech communication. In: Advances in network and acoustic echo cancellation. Berlin/Heidelberg, Germany: Springer; 2001. p. 1–30.
54. Habets EAP, Gannot S, Cohen I, Sommen P. Joint dereverberation and residual echo suppression of speech signals in noisy environments. IEEE Trans Audio Speech Lang Process. 2008;16(8):1433–51. doi:10.1109/tasl.2008.2002071.
55. Daspute K, Pandit H, Shinde S. Real time voice cloning. J Emerg Technol Innov Res (JETIR). 2020;7(6):120–5.
56. Liao S, Wang Y, Li T, Cheng Y, Zhang R, Zhou R, et al. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. arXiv:2411.01156. 2024.
57. RVC-Boss Team. GPT-SoVITS: Few-shot voice cloning framework. 2024 [cited 2025 Apr 14]. Available from: https://github.com/RVC-Boss/GPT-SoVITS.