Phoneme-Aware Mamba Watermark: An Active Defense System Against Purified Speech Deepfakes

Yanda Shao [1,*], *Mengke Zhang* [1,*], *Zhixin Lin* [2,*], *Tianyi Yang* [1,**]

[1]Beijing University of Posts and Telecommunications, China
[2]Beijing Institute of Technology, China

972359872@bupt.edu.cn, zmk1460997475@bupt.edu.cn, 1120233595@bit.edu.cn, 3662748738@qq.com

Abstract

With the rapid advancement of Text-to-Speech (TTS) and Voice Conversion (VC), forged speech has reached a highly realistic level, posing serious threats to voice biometrics. Traditional passive detection methods suffer from limited generalization against unseen synthesis models.

In this paper, we propose a phoneme-aware active watermarking defense built on the Mamba framework. Leveraging a Dual-column Bidirectional State Space Model, the system captures long-range dependencies and covertly embeds identity bitstrings into speech. To resist diffusion-based purification attacks such as PhonePuRe, we introduce a phoneme-guided embedding mechanism that tightly couples the watermark with speech semantics, ensuring transfer to forged speech during speaker adaptation. Experiments demonstrate effective source tracing with ultra-low latency (about 1 s) and high True Positive Rate under strong purification attacks. Source code is available at `https://github.com/Silence-ai423/phoneme-aware-mamba-watermark`.

Index Terms: speech watermarking, deepfake defense, state space models, adversarial purification, voice cloning

1. Introduction

The proliferation of zero-shot voice cloning technologies, exemplified by models like YourTTS [1] and sv2TTS [2], has drastically lowered the barrier to generating realistic speech deepfakes, leading to rising identity spoofing risks. The research community has primarily relied on passive Audio Deepfake Detection (ADD) systems. While ADD frameworks using pre-trained models such as wav2vec 2.0 [3], XLS-R [4], and Conformer [5] have shown success on benchmark datasets, they often fail to detect novel artifacts generated by emerging diffusion models such as DiffWave [6] and AudioPure [7].

Recent advancements introduce perturbation-based active defenses to disrupt deepfake generation, including AntiFake [8] and VoiceGuard [9]. However, studies such as De-AntiFake [10] show that surface-level adversarial perturbations are highly vulnerable. Algorithms like PhonePuRe [10] can effectively erase these signals, causing a collapse of the protection mechanism. This vulnerability highlights the need for a "shadow-like" defense paradigm—a signal that persists as an intrinsic timbre feature and transfers into forged samples even after purification and cloning.

To bridge this gap, we propose an active watermark defense system driven by a State Space Model (SSM) [11]. The main contributions of this work are three-fold: ♦ We pioneer the application of the Mamba architecture [11] and its dual-column bidirectional variant [12, 13] in active audio watermarking tasks, exploiting its linear computational complexity and bidirectional modeling capabilities. ♦ We design a phoneme-guided representation, utilizing stable phoneme regions obtained via Montreal Forced Aligner (MFA) [14] to protect the watermark bitstring from destructive diffusion-based purifications. ♦ We empirically validate that the proposed watermark successfully transfers from the original samples of targeted victims to the cloned audio, maintaining high retrievability even under state-of-the-art Voice Conversion pipelines [1, 2].

[*]These authors contributed equally.
[**]indicates the corresponding author.

2. Related Work

2.1. Audio Deepfake Detection (ADD)

Audio deepfake detection has progressed from handcrafted acoustic features (e.g., MFCCs and prosodic cues) with shallow classifiers to end-to-end deep models. Architectures such as RawNet2 [15] and AASIST [16] learn discriminative representations directly from raw waveforms, significantly improving performance. More recently, self-supervised models like XLS-R [4] and its Mamba-based variants, including XLSR-Mamba [12] and XLSR-MamBo [13], leverage large-scale pretraining to extract robust phonetic and timbral features. Meanwhile, the Mamba architecture [11] enables efficient long-sequence modeling with linear complexity, making it well suited for capturing subtle and distributed audio artifacts.

2.2. Active Speech Protection and Watermarking

Active speech protection, often framed as proactive audio defense, aims to prevent voice cloning or impersonation. Methods such as AntiFake [8], VoiceGuard [9], and timbre watermarking approaches [17] introduce imperceptible perturbations or adversarial noise to disrupt cloning models. While effective against voice synthesis attacks, these approaches generally do not embed retrievable information, limiting traceability. Recent timbre watermarking methods, such as AudioMarkNet [18], adopt a data-poisoning strategy: watermark patterns are learned during model adaptation, enabling retrieval of speaker-specific identifiers. However, these methods remain fragile under advanced purification techniques, particularly phoneme-conditioned denoising or spectral manipulation [10, 19].

2.3. Adversarial Purification

Adversarial purification techniques counteract active audio defenses. For example, De-AntiFake [10] introduces PhonePuRe, a two-stage purification pipeline combining wave-level diffusion denoising inspired by DiffWave [6] with phoneme-

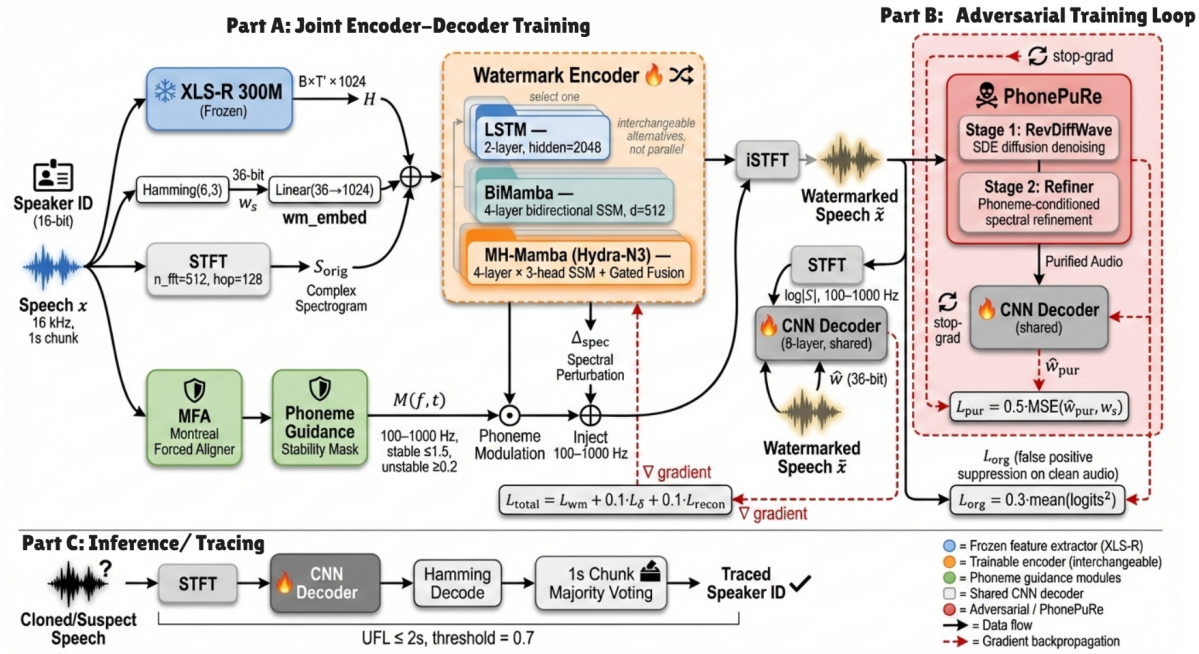


Figure 1: *Overview of the proposed* ***Phoneme-Aware Mamba Watermark*** *framework.* ***(A) Joint Encoder-Decoder Training:*** *The system features a dual-branch input where frozen XLS-R 300M features* $\mathbf{H}$ *and the speaker identity* $\mathbf{w}_s$ *condition the encoder. The generated spectral perturbation* Δ_{spec} *is modulated by the Phoneme Guidance mask* $M(f,t)$ *to target stable spectral regions before being injected into the STFT. The encoder and CNN decoder are trained end-to-end.* ***(B) Adversarial Training Loop:*** *To ensure robustness, the watermarked speech* $\tilde{\mathbf{x}}$ *undergoes differentiable purification attacks via PhonePuRe (RevDiffWave + Refiner). The encoder is optimized to resist this purification through the adversarial loss* $\mathcal{L}_{pur}$*.* ***(C) Inference & Tracing:*** *During deployment, the CNN decoder extracts the watermark from potentially cloned speech using 1-second chunk-wise majority voting to identify the source speaker. Legend: Blue modules are frozen; Red dashed lines indicate gradient/loss flow.*

conditioned spectral refinement. This framework effectively removes surface-level defense signals and restores perceptual audio quality, posing a significant threat to conventional watermarking schemes. Our work addresses this by embedding watermarks into phoneme stability regions, making phoneme information itself the defense carrier.

3. Proposed Methodology

3.1. Problem Formulation

Let $x \in \mathbb{R}^T$ denote an original speech waveform. Our objective is to learn a watermark embedding function

$$E_\theta : (x, w) \to \tilde{x} \tag{1}$$

where $w \in \{0,1\}^K$ represents a binary watermark code independent of speaker identity, $\tilde{x}$ is the watermarked speech, and θ denotes model parameters. A detector D_ϕ recovers the embedded watermark as

$$\hat{w} = D_\phi(\tilde{x}) \tag{2}$$

The system is designed to satisfy four properties. First, perceptual transparency requires

$$d(x, \tilde{x}) < \epsilon, \tag{3}$$

where $d(\cdot)$ measures perceptual distance. Second, detectability ensures $D_\phi(\tilde{x}) = w$ Third, robustness to voice cloning requires

$$D_\phi(G(\tilde{x})) = w \tag{4}$$

for any voice cloning model G. Finally, robustness to purification demands

$$D_\phi(P(G(\tilde{x}))) = w \tag{5}$$

where P denotes a purification operator such as diffusion- or spectral-based refinement [10, 6, 19].

3.2. Overall Architecture

As illustrated in Figure 1, we adopt a unified encode–inject–decode framework. The input waveform $\mathbf{x}$ is processed through two parallel branches: (1) a frozen pretrained XLS-R 300M model [4] extracts contextualized phonetic representations $\mathbf{H}$; (2) the STFT of $\mathbf{x}$ yields the original complex spectrogram $\mathbf{S}$. A watermark encoder, conditioned on the speaker-specific identity code $\mathbf{w}_s$ [19], maps $\mathbf{H}$ to a spectral perturbation Δ_{spec}, which is additively injected into the 100–1000 Hz band of $\mathbf{S}$. The watermarked waveform is reconstructed via inverse STFT:

$$\tilde{\mathbf{x}} = \text{iSTFT}\left(\mathbf{S} + \Delta_{\text{spec}}(\mathbf{H}, \mathbf{w}_s)\right) \tag{6}$$

The watermark decoder shares no parameters with the encoder. It extracts the embedded code directly from the STFT magnitude of the received audio:

$$\hat{\mathbf{w}} = D_\phi\left(\log|\text{STFT}(\tilde{\mathbf{x}})|\right) \tag{7}$$

The detector shares the same front-end feature extractor to ensure consistency in representation space.

3.3. Phoneme-Guided Watermark Embedding

To improve robustness against phoneme-conditioned purification, watermark perturbations are aligned with acoustically stable phoneme regions. Phoneme boundaries are obtained using Montreal Forced Aligner (MFA) [14], from which we construct a spectral stability mask $M(f,t)$ within the 100–1000 Hz band, corresponding to formant-dominated steady-state regions.

Watermark perturbations are applied as:

$$\Delta_{\text{spec}} = \Delta \odot M(f,t) \cdot \alpha_{wm}, \tag{8}$$

where α_{wm} controls watermark strength and $\odot$ denotes element-wise multiplication. By enhancing low-variance phoneme interiors and suppressing transition frames, watermark energy is concentrated in acoustic subspaces that are more likely to survive purification.

3.4. Watermark Encoder Architectures

Modeling long-range timbre dependencies is critical for robust watermark embedding. We investigate three encoder architectures.

The baseline Bi-LSTM updates hidden states sequentially:

$$h_t = \text{LSTM}(x_t, h_{t-1}), \tag{9}$$

which is straightforward but limited in capturing global dependencies.

The Dual-column Bidirectional Mamba (state space model, SSM) [12] encoder improves long-range modeling efficiency by updating

$$h_t = A(x_t)h_{t-1} + B(x_t), \tag{10}$$

and fusing forward and backward states:

$$H' = \text{Concat}(H_f, H_b). \tag{11}$$

This formulation provides linear complexity $\mathcal{O}(T)$ while maintaining stable long-sequence modeling.

Finally, we introduce a Multi-Head Mamba (MH-Mamba) encoder (Hydra-style SSM), which employs three parallel selective SSM heads with gated fusion across four stacked layers. Given the fused feature representation

$$X \in \mathbb{R}^{B\times T\times D}, \tag{12}$$

each layer performs residual multi-head state updates:

$$H^{(l)} = H^{(l-1)} + \text{Fusion}\left(\left\{\text{SSM}_i\left(H^{(l-1)}\right)\right\}_{i=1}^{3}\right), \tag{13}$$

where $\text{SSM}_i(\cdot)$ denotes the i-th parallel state space branch and $\text{Fusion}(\cdot)$ represents gated aggregation of the three heads.

To further incorporate phonetic structure, phoneme embeddings P_t modulate the state transition through a gating mechanism:

$$G_t = \sigma(W_p P_t), \tag{14}$$

$$h_t = G_t \odot \text{SSM}(h_{t-1}, x_t), \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid function. This phoneme-conditioned gating enables the encoder to adaptively emphasize stable acoustic regions, resulting in watermark perturbations that better survive purification.

3.5. Anti-Purification Adversarial Training

To ensure watermark survivability under purification, the training process incorporates purification perturbations [10] within the optimization loop. During training, purified samples are periodically introduced so that the encoder learns perturbations resilient to diffusion-based and spectral refinement operations. The overall objective function is defined as

$$\mathcal{L}_{\text{total}} = \lambda_{wm}\mathcal{L}_{wm} + \lambda_{\delta}\mathcal{L}_{\delta} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{org}}\mathcal{L}_{\text{org}} + \lambda_{\text{pur}}\mathcal{L}_{\text{pur}} \tag{16}$$

where $\mathcal{L}_{wm}$ enforces accurate watermark retrieval, $\mathcal{L}_{\delta}$ constrains perturbation magnitude, $\mathcal{L}_{\text{recon}}$ maintains waveform consistency, $\mathcal{L}_{\text{org}}$ penalizes false positives on clean speech, and $\mathcal{L}_{\text{pur}}$ explicitly encourages retrievability after purification. Joint optimization of these objectives yields watermark perturbations that are both imperceptible and robust.

3.6. Detection and Redundancy

The binary watermark is encoded using Hamming (6,3) error correction [20], producing a 36-bit sequence from an original 16-bit code. This sequence is repeatedly embedded over 1-second audio segments sampled at 16 kHz. During detection, segments are decoded independently and aggregated via majority voting:

$$\hat{w}^k = \text{MajorityVote}(\hat{w}^k_{1:T'}). \tag{17}$$

This redundancy ensures robustness against partial corruption. Under this configuration, the theoretical upper bound of the undetected forgery length (UFL) is at most 2 seconds.

4. Experiments

4.1. Experimental Setup

♦**Datasets.** Encoder training uses LibriSpeech train-clean-100 (253 speakers, $\sim$100 h, 16 kHz) [21]. Real-world threat simulation uses 11 selected VCTK speakers (downsampled to 16 kHz). ♦**Front-End and Phoneme Guidance.** The frozen XLS-R 300M extracts contextualized representations $\mathbf{H} \in \mathbb{R}^{B\times T'\times 1024}$ ($T'{=}\lfloor T/320 \rfloor$). Montreal Forced Aligner (MFA) provides phone-level boundaries (ARPAbet, $\sim$10 ms precision). A phoneme average spectrum dictionary is precomputed and used by the PhonemeGuidance module to derive per-phoneme stability masks $M(f,t)$ in the 100–1000 Hz band (stable bins $\leq$1.5, unstable bins $\geq$0.2). ♦**Watermark Codec.** STFT: $n_{\text{fft}}{=}512$, hop=128, Hann window. Embedding band: 100–1000 Hz ($|\mathcal{F}|{=}29$ bins). Speaker identity: 16-bit $\rightarrow$ 36-bit Hamming(6,3) codeword. Audio is processed in 1 s non-overlapping chunks. ♦**Decoder.** A shared 8-layer CNN decoder extracts watermarks from the log-magnitude spectrogram of the 100–1000 Hz region (channels $64{\rightarrow}128{\rightarrow}256$, BatchNorm, LeakyReLU, MaxPool2d), followed by adaptive pooling and a linear classifier ($256{\rightarrow}36$). ♦**Metrics.** *Imperceptibility*: SNR (dB) measures signal distortion (higher is better); PESQ [22] ranges from 1–5 and evaluates perceptual speech quality (higher is better). *Accuracy*: Bit-level accuracy (%) quantifies correct watermark bit recovery. *Defense*: TPR/Tracing Accuracy (%) measures successful source identification; UFL (%) reflects spoofing false acceptance under watermarking (lower is better); Retain. Rate (%) measures watermark robustness after purification (higher is better).

4.2. Watermark Encoder Training

Three encoders—LSTM, BiMamba, and MH-Mamba (Section 3.4)—are trained independently under identical settings. Each of 100 LibriSpeech speakers is assigned a unique watermark ID. The encoder generates a spectral perturbation Δ, optionally modulated by the phoneme stability mask, and injects it into the original complex spectrogram via additive embedding. The training loss is:

$$\mathcal{L}_{\text{enc}} = \|\Delta_{\text{spec}}\|^2 + \text{MSE}(\hat{\mathbf{w}}, \mathbf{t}) + 0.1 \cdot \text{MSE}(\tilde{\mathbf{x}}, \mathbf{x}) \quad (18)$$

where $\mathbf{t}=(2\mathbf{w}-1)\times 5$. We use AdamW (lr$=10^{-4}$), cosine annealing, batch size 16, for 200 epochs. The best checkpoint (highest validation bit accuracy) is selected.

4.3. Anti-Purification Adversarial Training

Starting from the encoder training checkpoint, each model is adversarially fine-tuned against the PhonePuRe two-stage purification (RevDiffWave + spectral Refiner), applied with probability $p_{\text{pur}}=0.3$ in a stop-gradient fashion. The joint loss uses weights $\lambda_{\text{wm}}=1.0$, $\lambda_{\delta}=0.1$, $\lambda_{\text{org}}=0.3$, $\lambda_{\text{pur}}=0.5$. Training uses AdamW (lr$=10^{-5}$), batch size 12, for 50 epochs, with model selection on purified accuracy. As shown in Table 1, adversarial fine-tuning markedly improves purified accuracy and retention rate across all encoders. MH-Mamba achieves the strongest robustness, indicating that phoneme-aware multi-head SSM modeling combined with purification-aware optimization yields a resilient and transferable watermark.

Table 1: *Watermark robustness before/after PhonePuRe purification across encoder types and with/without adversarial training stages.*

Encoder	Stage	Original Acc (%) ↑	Purified Acc (%) ↑	Retain. Rate (%) ↑
LSTM	Base	77.07	48.26	62.62
	Adv.	82.51	67.18	81.42
BiMamba	Base	84.77	75.53	89.10
	Adv.	89.68	80.25	89.48
MH-Mamba	Base	89.67	80.19	89.43
	Adv.	91.56	84.55	92.34

4.4. Real-World Threat Simulation

We simulate an end-to-end voice cloning attack. Eleven VCTK speakers are watermarked using the adversarially-trained model (1 s chunk-wise embedding). An attacker fine-tunes YourTTS and sv2TTS on the watermarked data, then generates 15 cloned utterances per speaker ($11\times15=165$ total). The defender performs chunk-wise decoding on the cloned speech, matches decoded bits against the watermark pool via Hamming similarity (threshold 0.7), and applies majority voting across chunks for source tracing. Table 2 demonstrates the system successfully traces the origin of forged audio sources with nearly 100% accuracy and extremely low latency of approximately 1.1 seconds when confronted with YourTTS and sv2TTS high-fidelity voice cloning attacks.

4.5. Speech Quality

Table 3 evaluates imperceptibility on 100 randomly sampled LibriSpeech utterance pairs per encoder. With all encoders maintaining PESQ scores above 3.5 and achieving SNR values exceeding 35 dB, the results demonstrate excellent acoustic transparency and perceptual quality.

Table 2: *Source tracing on TTS-cloned speech (11 speakers, 165 utterances).*

Encoder	TPR (%) ↑	UFL (s) ↑
LSTM	99.9 ± 0.2	1.2 ± 0.4
BiMamba	100.0 ± 0.1	1.1 ± 0.2
MH-Mamba	99.9 ± 0.1	1.0 ± 0.1

Table 3: *Speech quality of watermarked audio.*

Encoder	SNR (dB) ↑	PESQ ↑
LSTM	34.12 ± 4.23	3.53 ± 0.41
BiMamba	35.13 ± 3.48	3.55 ± 0.39
MH-Mamba	35.67 ± 4.54	3.54 ± 0.57

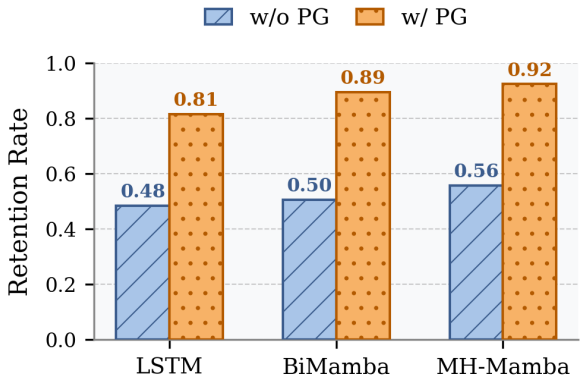


Figure 2: *Ablation on phoneme guidance. Watermark retain. rate (%) after PhonePuRe purification with and without phoneme guidance (PG) across three encoder architectures.*

4.6. Ablation: Phoneme Guidance

To isolate the contribution of the phoneme stability mask $M(f,t)$, we train all encoders with and without phoneme guidance. Figure 2 reports retain Rate after PhonePuRe purification, which reflects the phoneme-guided component plays a vital role in preventing the PhonePuRe purification attacks.

5. Conclusion

This paper presents a phoneme-aware active watermark defending system utilizing the Mamba architecture. By intertwining the watermark signal with phonetic steady states and employing an adversarial training paradigm against PhonePuRe, the system achieves highly robust, imperceptible, and ultra-low-latency traceability. Experimental results confirm that the phoneme-aware Mamba encoder successfully preserves speaker identity strings through complex TTS cloning pipelines, offering a pragmatic defense mechanism against emerging speech deepfakes.

6. Acknowledgments

The authors would like to warmly thank their friends for their unyielding support, constructive feedback, and invaluable companionship that made this research possible.

7. Generative AI Use Disclosure

During the preparation of this manuscript, the authors used [ChatGPT / Gemini] solely for the purpose of improving the English language phrasing and assisting with LaTeX formatting. After using this tool, the authors rigorously reviewed and edited the content as needed, and take full responsibility for the accuracy and originality of the final publication. All core scientific concepts, experimental designs, and data analyses were conducted entirely by the human authors.

8. References

[1] E. Casanova *et al.*, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, 2022, pp. 2709–2720.

[2] Y. Jia *et al.*, "Transfer learning from speaker verification to multi-speaker text-to-speech synthesis," in *Proc. NeurIPS*, vol. 31, 2018.

[3] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[4] A. Babu *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022, pp. 2278–2282.

[5] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[6] Z. Kong *et al.*, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, 2021.

[7] S. Wu, J. Wang *et al.*, "Defending against adversarial audio via diffusion model," in *Proc. 11th International Conference on Learning Representations (ICLR)*, 2023.

[8] Z. Yu, S. Zhai, and N. Zhang, "AntiFake: Using adversarial audio to prevent unauthorized speech synthesis," in *Proc. ACM CCS*, 2023, pp. 460–474.

[9] J. Li *et al.*, "Voice Guard: Protecting voice privacy with strong and imperceptible adversarial perturbation," in *Proc. IJCAI*, 2023.

[10] W. Fan, K. Chen, C. Liu, W. Zhang, and N. Yu, "De-AntiFake: Rethinking the protective perturbations against voice cloning attacks," in *Proc. 42nd International Conference on Machine Learning (ICML)*, 2025.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. Conference on Language Modeling (COLM)*, 2024.

[12] Y. Xiao and R. K. Das, "XLSR-Mamba: A dual-column bidirectional state space model for spoofing attack detection," *IEEE Signal Processing Letters*, 2025.

[13] K.-H. Ng, T. Song *et al.*, "XLSR-MamBo: Scaling the hybrid mamba-attention backbone for audio deepfake detection," *arXiv preprint arXiv:2601.02944*, 2025.

[14] M. McAuliffe *et al.*, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[15] H. Tak *et al.*, "End-to-end anti-spoofing with rawnet2," in *Proc. IEEE ICASSP*, 2021.

[16] J.-w. Jung *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE ICASSP*, 2022.

[17] C. Liu, J. Zhang *et al.*, "Detecting voice cloning attacks via timbre watermarking," *Proc. NDSS*, 2024.

[18] W. Zong, Y.-W. Chow, W. Susilo, J. Baek, and S. Camtepe, "AudioMarkNet: Audio watermarking for deepfake speech detection," *Preprint*, 2025.

[19] S. Welker *et al.*, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech*, 2022.

[20] S. Lin and D. J. Costello, *Error Control Coding*, 2nd ed. Pearson, 2004.

[21] V. Panayotov *et al.*, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015.

[22] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ)," in *Proc. IEEE ICASSP*, 2001.