MambaVoiceCloning: Efficient and Expressive Text-to-Speech via State-Space Modeling and Diffusion Control

Sahil Kumar and Namrataben Patel, Ph.D. in Mathematics

FACULTY MENTORS: Honggang Wang, Ph.D., and Youshan Zhang, Ph.D.

Katz School of Science and Health

Motivation & Problem

Most TTS conditioning stacks rely on transformer attention (quadratic) or recurrent modules (long-range drift). Even existing Mamba-TTS systems remain **hybrid at inference**, retaining attention for duration/style modules.

- **Quadratic Complexity:** Attention scales $O(T^2)$ in text/prosody encoders
- **Recurrent Drift:** LSTM/GRU exhibit long-range gradient instability
- **Hybrid Gap:** Prior Mamba-TTS keeps attention for style/duration

Key Contributions

1. First diffusion TTS with fully SSM-only inference conditioning across text, rhythm & prosody
2. Gated Bi-Mamba fusion + AdaLN — improves long-range prosody stability, reduces drift
3. Protocol-matched baselines isolating architectural impact of removing attention
4. Deployment-oriented evaluation covering long-form robustness, finite look-ahead streaming, and predictable linear-time behavior

Methodology: SSM-Only Conditioning Stack

- **Gated Bi-Mamba Text Encoder**

$$h_f = Mamba_f(x), \qquad h_b = Mamba_b(x)$$
$$h_T = \left(\sigma\left(W_g[h_f; h_b]\right) \odot [h_f; h_b]\right)W_o$$

Replaces self-attention with bidirectional scans + gated forward/backward fusion. O(T) complexity, bounded activations.

- **Expressive Mamba + AdaLN**

$$h_{M,s} = AdaLN(h_M, e), \qquad h_E = Mamba(h_{M,s})$$

Injects speaker prosody from mel features in O(T). Largest CMOS drop when removed (−0.41). Captures slow prosodic dynamics over long inputs.

- **Temporal Bi-Mamba Encoder**

$$h_B = [h_f; h_b]W_f$$

Models rhythmic structure & phoneme-duration alignment. Linear fusion kept (no 2nd gating) — disentanglement handled upstream.

21M	1.6×	72%	O(T)
Encoder Parameters	Encoder Speedup	Peak Memory vs StyleTTS2	Conditioning Complexity

OOD Set (CMOS-N drop vs full MVC)

Removed Component	CMOS-N Drop
Expressive Mamba predictor	**−0.41 (largest)**
Bi-Mamba text encoder	−0.38
Temporal Bi-Mamba encoder	−0.36

Architecture Overview

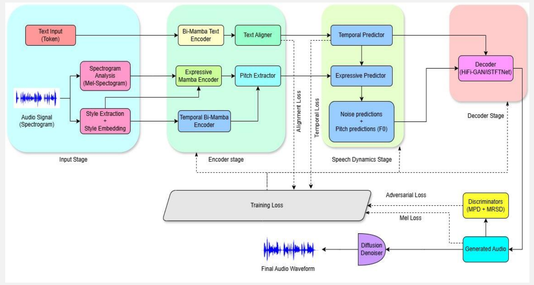
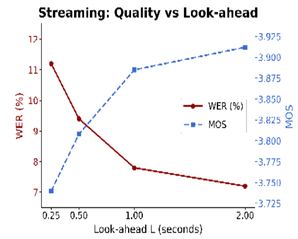
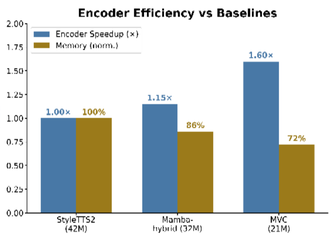


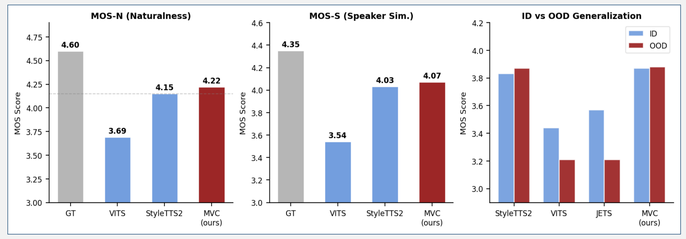
Fig 1: MVC framework. Bi-Mamba Text Encoder (phoneme modeling) + Temporal Bi-Mamba (rhythm/duration) + Expressive Mamba + AdaLN (prosody). Training-time aligner (dotted) discarded at inference — fully SSM-only deployment.

Streaming Performance & Encoder Efficiency

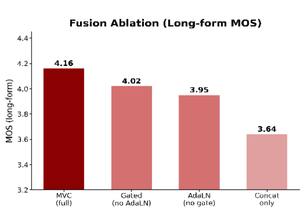


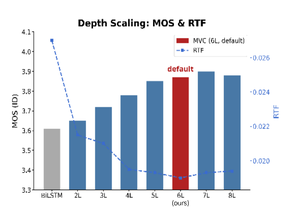


Main Results: Subjective Quality and ID/OOD Generalization



Ablation Studies: Fusion Design and Encoder Depth





LJSpeech Objective Results and Mamba Baseline Comparison

Model	F0 RMSE↓	MCD↓	WER↓	PESQ↑	RTF↓
VITS	0.667	4.97	7.23%	3.64	0.0211
StyleTTS2	0.651★	4.93	6.50%★	3.79	0.0174
Hybrid-Mamba	0.659	4.95	6.68%	3.79	0.0189
Bi-Mamba (concat)	0.656	4.93	6.58%	3.82	0.0181
MVC (ours)	**0.653**	**4.91★**	**6.52%**	**3.85★**	**0.0169★**

Conclusions

MVC is the first diffusion TTS with fully SSM-only conditioning over text, rhythm, and prosody, where Gated Bi-Mamba + AdaLN proves critical (MOS 4.16 vs 3.64). It delivers efficient high-quality synthesis with a 21M encoder (1.6× speedup, 72% memory) and consistent perceptual gains (+0.07 MOS over StyleTTS2). MVC further enables streaming TTS (≥0.5s) with strong quality (WER 9.4%, MOS 3.81) and robust generalization across VCTK and multilingual CSS10 (ES/DE/FR).

MVC vs Prior Systems at Inference

System	Attn @ Infer?	Rhythm	Prosody	SSM Only?
StyleTTS2	Yes	Attn/var	Attn ref	✗
Zhang'24	Hybrid	Mixed	Mixed	✗
MVC (ours)	None	Bi-Mamba	Mamba+LN	✓

CVTK Zero-Shot

Model	MOS-N	MOS-S
VITS	3.66	3.53
StyTTS2	4.12	4.01
MVC	**4.18★**	**4.09★**

References

References

[1] Gu, A. and Dao, T. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, 2024.
[2] Li, Y. A., Han, C., Raghavan, V. S., Mischler, G., and Mesgarani, N. *StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models*, 2023.
[3] Jiang, X., Li, Y. A., Florea, A. N., Han, C., and Mesgarani, N. *Speech Slytherin: Examining the Performance and Efficiency of Mamba for Speech Separation, Recognition, and Synthesis*, 2024.
[4] Zhang, X. et al. *Mamba in Speech: Towards an Alternative to Self-Attention*, 2024.
[5] Popov, V. et al. *Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech*, ICML, 2021.