DeepSound 刘百云

Speaker-Aware State Space Modeling for Streaming Target Speaker Extraction

Abstract

Target speaker extraction (TSE) aims to recover the speech of a target speaker from complex mixtures containing interfering speakers, background noise, and reverberation by using an enrollment utterance from the target speaker. It is a key problem in personalized speech enhancement and real-time communication. To balance extraction capability and real-time inference efficiency, this paper proposes Speaker-Aware Mamba (SA-Mamba) for streaming target speaker extraction. Unlike external feature fusion after temporal modeling, SA-Mamba injects the target-speaker condition directly into the state dynamics of a selective state space model, making the input matrix B, output matrix C, and discretization step Δt jointly depend on the current acoustic features and the speaker representation derived from the enrollment utterance. This enables speaker-dependent writing, reading, and forgetting within the recurrent state. At the system level, we adopt a frequency-band partitioning strategy with fine low-frequency bands and merged high-frequency bands to reduce modeling complexity while preserving speaker-related details. A dual-path architecture is then used to model inter-band spectral structure and intra-band causal temporal dependencies, respectively. Furthermore, we design a progressive speaker distillation strategy: shallow layers use a precomputed global speaker condition for robust target localization, whereas deeper layers retrieve fine-grained context from the enrollment utterance through frame-level attention to improve reconstruction quality. Without relying on an additional speaker recognition model, the proposed framework fully exploits enrollment information while maintaining a fixed-size streaming state, providing a unified state-space modeling solution for low-complexity and high-fidelity target speaker extraction. In the IEEE SLT 2026 REAL-TSE Challenge, our method achieved strong performance in WER, SpkSim, DNSMOS, and F1, and ranked second in the final score.

Index Terms: target speaker extraction; personalized speech enhancement; streaming inference; state space model; Mamba; speaker conditioning; band splitting

I. Introduction

In realistic acoustic environments where multiple speakers, background noise, and room reverberation coexist, human listeners can selectively attend to a target voice according to speaker identity, a capability commonly known as the cocktail-party effect. Target speaker extraction (TSE) aims to equip machine listening systems with a similar capability: given an enrollment utterance from a target speaker, the system is required to recover the target-speaker signal from a speech mixture [1], [2]. Unlike blind source separation, which estimates multiple sources simultaneously and must address the permutation ambiguity, TSE outputs only the speech corresponding to the enrollment utterance. It is therefore well suited to personalized speech enhancement, real-time communication, meeting transcription, hearing aids, and smart devices [11], [16], [40].

Streaming TSE for online scenarios must simultaneously satisfy three requirements: target selectivity, causal modeling, and low complexity. First, the model must effectively exploit speaker cues in the enrollment utterance and maintain stable target discrimination when the target and interfering speakers have similar timbres, speech overlap is severe, or channel mismatch exists between enrollment and test utterances. Second, the model must capture long-term dependencies without accessing future frames, while avoiding explicit context caches that grow with sequence length. Finally, the system must control both computational cost and state size and reduce algorithmic latency to satisfy real-time deployment constraints on memory, power consumption, and response speed.

Existing TSE methods mainly focus on how target-speaker information is represented and injected. A common strategy is to use a pretrained speaker recognition model to extract a global embedding, such as a d-vector, x-vector, or ECAPA-TDNN embedding, from the enrollment utterance [5], [38], [42], and then fuse it into the separation network through concatenation, element-wise modulation, additive bias, FiLM, or related mechanisms [2], [18]. Such methods are simple to implement, and the enrollment embedding can be precomputed before inference, which makes them attractive for online deployment. However, a global embedding compresses the entire enrollment utterance into a single vector and therefore cannot fully preserve phoneme-level spectral patterns, short-term pronunciation variations, speaking style, and local context. When the target and interfering speakers have similar timbres or pronunciation patterns, this information bottleneck can weaken the model's discriminative ability.

To alleviate the information compression caused by fixed embeddings, recent studies have begun to use the time-frequency representation of the enrollment utterance directly. USEF-TSE retrieves frame-level target-speaker information from enrollment STFT features using multi-head cross-attention, demonstrating the importance of preserving detailed enrollment information for TSE [35]. DSINet further highlights the value of dynamic speaker information fusion for real-time TSE

[14], while TEA-PSE and related sub-band networks show that appropriate band partitioning and sub-band modeling help balance enhancement performance and computational efficiency [30], [31]. Nevertheless, continuously applying fine-grained cross-attention across multiple network layers introduces a per-frame cost that increases with enrollment length; reverting to a single global embedding reintroduces the information-compression problem. Therefore, preserving enrollment details while maintaining fixed states and low-cost streaming inference remains a central challenge in TSE system design.

State space models (SSMs), especially the selective state space model Mamba, provide a new modeling perspective for this problem [6], [7]. Unlike self-attention, which explicitly accesses context sequences, Mamba models long-range dependencies through a fixed-size recurrent state and uses input-dependent selective parameters to control information writing, reading, and forgetting. Recent work on speech separation and speech enhancement has demonstrated the potential of Mamba as an efficient sequence-modeling backbone [8], [9]. However, most existing methods mainly use it as an alternative to RNN or Transformer modules; the selective parameters are usually determined only by the current acoustic input, and the conditionable nature of the state dynamics has not been fully exploited to represent target-speaker information.

We argue that the target-speaker selectivity required by TSE is naturally compatible with Mamba's input-selection mechanism. If the selective parameters depend not only on the current mixture features but also on a target-speaker condition derived from the enrollment utterance, the state recurrence itself can be transformed into a target-speaker-oriented dynamic filtering process. Based on this idea, we propose Speaker-Aware Mamba (SA-Mamba), which injects speaker conditions directly into the parameter generation process of the SSM, so that the input matrix B, output matrix C, and discretization step Δt are jointly determined by the current acoustic features and the target-speaker condition. Unlike external feature fusion after temporal modeling, this mechanism changes what the model writes, what it reads, and at what time scale it forgets during recurrence, thereby forming target-speaker-dependent dynamic memory in a fixed state. We build a time-frequency-domain system for streaming TSE and use non-uniform band partitioning to reduce high-frequency modeling cost while preserving low-frequency speaker details [4], [13]. A dual-path backbone then models inter-band spectral structure and intra-band causal temporal dependencies. We further design a progressive speaker distillation strategy: shallow layers use a precomputed global speaker condition for target localization, whereas deeper layers retrieve enrollment details through frame-level attention, balancing real-time efficiency and reconstruction quality.

The main contributions of this paper are as follows. (1) We propose a speaker-conditioned state-space modeling method that directly incorporates target-speaker information into the selective parameters of Mamba, enabling speaker-dependent writing, reading, and forgetting in the state dynamics. (2) We design a non-uniform band-splitting and dual-path streaming

TSE architecture that reduces spectral modeling cost while preserving key speaker cues. (3) We propose a progressive speaker distillation strategy that hierarchically allocates global conditioning and frame-level enrollment retrieval to balance target localization and detail recovery. (4) We validate the proposed method under the IEEE SLT 2026 REAL-TSE Challenge setting. The results show that the system is competitive in WER, SpkSim, DNSMOS, and target activity F1, and achieves second place in the final score.

II. Proposed Method

A. System Overview

The proposed SA-Mamba-TSE performs target speaker extraction in the complex STFT domain. Given a mixture x and a target-speaker enrollment utterance a, the model first computes complex spectra using a 512-point STFT with a 256-sample hop and a 512-sample window, yielding F = 257 non-redundant frequency bins. The mixture and enrollment utterances are then passed through a shared non-uniform band-splitting frontend and mapped to D = 128 dimensional representations over K = 36 frequency bands. The enrollment branch further produces two types of speaker conditions: a global speaker condition obtained by pooling enrollment features over the band and time dimensions, and a frame-level retrieval memory composed of enrollment time-frequency features. The backbone consists of N = 36 dual-path blocks. Each block first models cross-band spectral structure along the band axis and then performs causal speaker-aware state-space recurrence along the time axis. Finally, the model estimates a complex mask for each band and reconstructs the target speech through iSTFT.

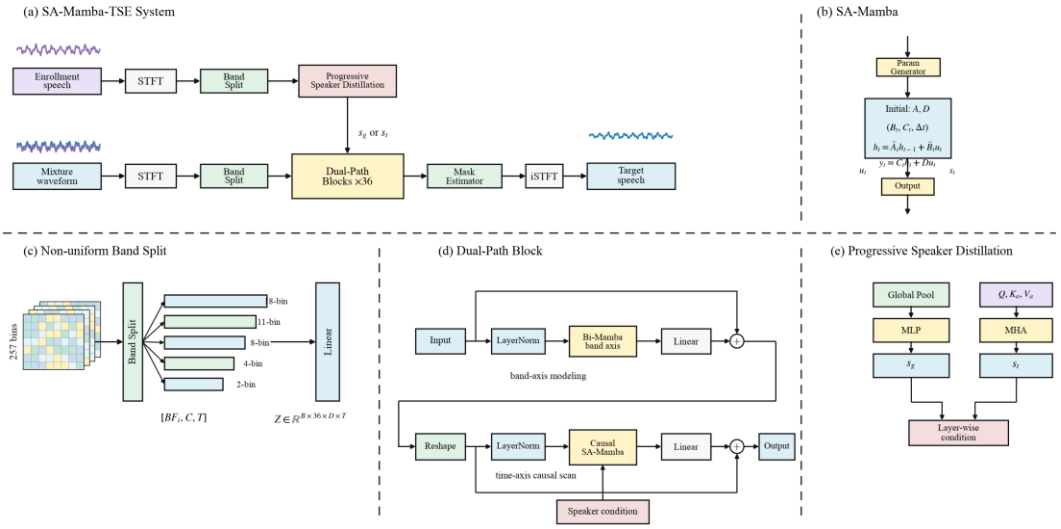

Fig. 1. Architecture of SA-Mamba-TSE. The mixture and enrollment utterances are first transformed into 257-dimensional complex spectra and partitioned into 36 non-uniform

frequency bands. Shallow dual-path blocks use the global speaker condition, whereas deeper dual-path blocks retrieve enrollment details through frame-level attention. SA-Mamba modulates the selective SSM parameters with the speaker condition during temporal recurrence.

B. Non-Uniform Band Splitting and Enrollment Encoding

Speech spectra exhibit different perceptual and statistical characteristics across frequency regions. The low-frequency region contains the fundamental frequency, low-order harmonics, and timbre-related speaker cues, and thus requires finer frequency resolution. The mid-frequency region carries formants and major intelligibility information. The high-frequency region mainly reflects consonant details, sibilants, and noise textures, and can be appropriately merged to reduce sequence length and parameter cost. Inspired by band-split modeling [4], [13], [23], we do not model all 257 frequency bins directly; instead, we partition the spectrum into 36 non-uniform frequency bands.

With a 16-kHz sampling rate and a 512-point STFT, the spacing between adjacent frequency bins is 31.25 Hz. The band widths used in this paper are

$$w = [2\times8, 4\times8, 8\times6, 11\times13, 18\times1], \quad \text{total}(w) = 257. \quad (1)$$

This partition corresponds to a fine-to-coarse frequency organization. The 0-0.5 kHz region is divided into eight 2-bin bands to preserve the fundamental frequency and low-order harmonics; the 0.5-1.5 kHz region is divided into eight 4-bin bands to model low-order formants and timbre contours; the 1.5-3 kHz region is divided into six 8-bin bands to cover the main speech intelligibility information; and the 3-8 kHz region is divided into fourteen wider bands to model high-frequency consonants, sibilants, and environmental noise components.

For the k-th band, the real and imaginary parts of the corresponding frequency bins in the mixture spectrum are concatenated along the channel dimension and then passed through LayerNorm and a linear projection to obtain a fixed-dimensional band representation:

$$z_k(t) = \text{Linear}_k(\text{LN}_k([\text{Re}(X_k(t)); \text{Im}(X_k(t))])) \in R^D. \quad (2)$$

The enrollment utterance uses the same band partitioning and projection structure, resulting in an enrollment feature E with shape B x K x D x T_a. The global speaker condition is obtained by average pooling the enrollment feature over the band and time dimensions followed by a two-layer MLP:

$$s\_g = \text{MLP}(\text{MeanPool}_{k,t}(E_{k,t})). \quad (3)$$

This global condition only needs to be computed once during enrollment, making it suitable for streaming inference. Meanwhile, the model retains the projected enrollment time-

frequency sequence as keys and values for deep frame-level attention, supplementing the local pronunciation and spectral details that are difficult to represent through global pooling.

C. Speaker-Aware Selective State Space Model

The core of Mamba is an input-dependent selective state-space recurrence [7]. Given a time-series input u_t, the state update and output can be written as

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t u_t, \quad y_t = C_t h_t + D u_t. \quad (4)$$

where $\bar{A}_t$ and $\bar{B}_t$ are obtained from the continuous-time parameters A, B_t and the discretization step Δt. Standard Mamba usually generates selective parameters only from the current input:

$$(B_t, C_t, \Delta t) = f\theta(u_t). \quad (5)$$

This design can adjust state writing, reading, and forgetting according to the current acoustic content, but it does not explicitly exploit the target-speaker condition provided by the enrollment utterance. For TSE, the model needs to determine not only what acoustic content is present in the current frame, but also whether that content belongs to the target speaker. Therefore, we extend the generation of selective parameters so that it depends on both the current acoustic input and the speaker condition:

$$(B_t, C_t, \Delta t) = f\theta([u_t; \varphi(s_t)]). \quad (6)$$

where s_t can be either the global speaker condition s_g or a dynamic condition obtained by frame-level attention, and phi(.) denotes a speaker projection layer with SiLU activation. This mechanism endows B_t, C_t, and Δt with speaker-dependent input writing, state reading, and time-scale control, respectively. Intuitively, when harmonics, formants, or short-term pronunciation patterns consistent with the target speaker appear in the current band, the model can strengthen the writing and reading of the corresponding information. When an interfering speaker dominates, the model can adjust the time scale to reduce contamination of the recurrent state by non-target speech.

Unlike concatenation, multiplication, or attention fusion after temporal modeling, SA-Mamba injects the speaker condition into the state dynamics themselves. During streaming inference, each frequency band maintains only a fixed-size convolution cache and SSM state. The per-frame recurrence complexity is independent of sequence length, and the main cost remains O(d_inner x d_state).

D. Progressive Speaker Distillation

Separation networks at different depths require different types of speaker information. Shallow layers mainly perform target localization and coarse source discrimination, requiring stable and low-cost identity conditions. Deeper layers are responsible for residual-interference

suppression and detail reconstruction, and therefore depend more on local pronunciation and spectral-shape information in the enrollment utterance. To this end, we adopt a progressive speaker distillation strategy: the first 30 dual-path blocks use the global speaker condition s_g, while the last six dual-path blocks use the frame-level enrollment retrieval condition.

In the deeper blocks, the current mixture feature is first projected into a query Q, and the enrollment features are flattened into cross-band and cross-time keys and values K_a and V_a. Multi-head attention is used to retrieve the target-speaker context most relevant to the current mixture feature from the enrollment utterance [12]:

$$S = \mathrm{MHA}(Q, K_a, V_a). \quad (7)$$

During training, the attention output is aggregated along the band dimension to obtain a frame-level speaker condition s_t with shape B x T x d_s, which is used for frame-wise state recurrence in the deep SA-Mamba layers. During streaming inference, K_a and V_a on the enrollment side can be precomputed and cached during initialization. For each arriving STFT frame, the model only computes the query for the current frame and retrieves the enrollment memory. In this way, PSD restricts the higher-cost frame-level attention to the deeper layers where detail reconstruction is most needed, while preserving the low-latency property of shallow recurrence.

E. Dual-Path Block Architecture

Each of the N = 36 dual-path blocks contains two submodules:

1) Intra-band path: bidirectional Mamba across frequency bands (spectral modeling)

Within each time frame, all 36 frequency bands are processed simultaneously. Since the frequency axis is not subject to causal constraints, bidirectional Mamba (forward + backward + linear combination) is used to capture full spectral context, including harmonic relationships and formant patterns. This is a frame-wise operation with no temporal state: each time step observes the complete spectral snapshot.

2) Inter-band path: causal SA-Mamba per band (temporal modeling)

Within each frequency band, the temporal sequence is processed independently while carrying hidden states across frames. This is where SA-Mamba operates: speaker-conditioned selectivity enables the temporal model of each band to maintain speaker-related information in its state, and causal constraints ensure streaming compatibility. The 36 frequency bands are processed in parallel as a batch, each with an independent SA-Mamba state.

F. Mask Estimation and Signal Reconstruction

After N dual-path blocks, the band-wise features are projected by a two-layer MLP with Tanh activation to obtain a complex ratio mask:

$$M_k = \sigma(G_k) \odot V_k \quad \text{(gated mask)} \quad (9)$$

The complex mask is applied to the input STFT as follows:

$$\hat{S}_{real} = X_{real} \cdot M_{real} - X_{imag} \cdot M_{imag} \quad (10)$$

$$\hat{S}_{imag} = X_{real} \cdot M_{imag} + X_{imag} \cdot M_{real} \quad (11)$$

The enhanced signal is obtained by inverse STFT with overlap-add.

III. Experimental Setup

A. Datasets

We construct two-speaker target speaker extraction experiments using the CN-Celeb and VoxCeleb1 datasets. CN-Celeb contains real recordings from approximately 3,000 Chinese speakers, while VoxCeleb1 covers diverse speakers and recording conditions. All clean speech is resampled to 16 kHz. Room impulse responses (RIRs) are used to simulate reverberation, and real environmental noise is added to generate mixtures closer to practical scenarios. The reverberation time T60 is set to 0.2-1.0 s, the SNR range is -6 to 3 dB, the relative energy difference between the two speakers is -5 to 5 dB, and the speaker-to-microphone distance is 0.66-2.0 m. The data are split into training, validation, and test sets at a ratio of 7:2:1. The target-speaker enrollment utterance is a 3-s clean speech segment with different linguistic content from the test utterance, so as to evaluate target selectivity under text-independent enrollment.

B. Model Configuration and Training Details

The model input is represented by the short-time Fourier transform (STFT). The window length is set to 32 ms, corresponding to 512 samples; the hop size is set to 16 ms, corresponding to 256 samples; and the frequency dimension contains 257 frequency bins. The band partition uses 36 non-uniform sub-bands to reduce high-frequency modeling cost while preserving low-frequency speaker cues. The backbone feature dimension is set to 128, and the speaker-condition dimension is set to 128. In the SA-Mamba module, the state dimension is $d_{state} = 16$, the local convolution kernel size is $d_{conv} = 4$, and the expansion factor is $expand = 2$, corresponding to an internal dimension of $d_{inner} = 256$. The network stacks 36 dual-path blocks in total. Progressive speaker distillation is applied to the last six layers, and the deep frame-level enrollment retrieval uses four attention heads. The model has 15.89 M parameters. Training uses the Adam optimizer with an initial learning rate of 1×10^{-4}; the learning rate is halved when the validation loss does not decrease for five consecutive epochs. The training objective combines SI-SNR loss and multi-resolution time-frequency-domain loss. The multi-resolution loss uses FFT sizes of {512, 1024, 2048} with corresponding hop sizes of {50, 120, 240} and window lengths of {240, 600, 1200}. The batch size is set to 16, and the model is trained for 200 epochs on eight GPUs.

C. Baselines

To validate the effectiveness of the proposed method, we select two causal target speaker extraction systems as baselines: BSRNN_EMB_CAUSAL and BSRNN_TFMAP_CAUSAL. BSRNN_EMB_CAUSAL uses causal BSRNN as the separation backbone [4], extracts a global speaker embedding from the enrollment utterance using ECAPA-TDNN [5], and fuses the embedding into the intermediate representation of the separator through element-wise multiplication. BSRNN_TFMAP_CAUSAL also adopts a causal BSRNN backbone, but further introduces time-frequency mapping (TF-Map) and frame-level context cross-attention (Context) to enhance speaker-condition modeling [14], [39]. Specifically, TF-Map generates target-related features according to the spectral correlation between enrollment and mixture speech, while the Context module retrieves frame-level speaker information from the enrollment utterance through cross-attention and injects it into the separation representation. In contrast to these external-fusion conditioning methods, our method directly uses the speaker condition to generate parameters of the selective state space model, so that target-speaker selection occurs during recurrent state updates.

D. Evaluation Metrics

Model performance is evaluated from three aspects: signal fidelity, perceptual quality, and speech intelligibility. Scale-invariant signal-to-noise ratio improvement (SI-SNRi) measures the SNR improvement of the output speech relative to the input mixture, in dB; higher values indicate more complete target-speech recovery. Perceptual Evaluation of Speech Quality (PESQ) estimates the subjective perceptual quality of enhanced speech and typically ranges from -0.5 to 4.5; higher values indicate better listening quality. Short-Time Objective Intelligibility (STOI) measures speech intelligibility and ranges from 0 to 1; higher values indicate that the speech content is easier to understand.

For the official evaluation of the IEEE SLT 2026 REAL-TSE Challenge, we further report word error rate (WER), speaker similarity (SpkSim), DNSMOS, and target-speaker activity detection F1. WER reflects the automatic speech recognition accuracy of the extracted speech, where lower values indicate better preservation of linguistic content. SpkSim computes the cosine similarity between the speaker embeddings of the extracted speech and the target enrollment utterance, where higher values indicate better speaker identity preservation. DNSMOS is used for non-intrusive evaluation of the overall perceptual quality of enhanced speech. The F1 score combines temporal precision and recall for target-speaker activity detection and evaluates whether the system outputs speech accurately during target-speaker active regions. Before official scoring, all audio is amplitude-normalized using a unified procedure to ensure comparability across systems.

E. Experimental Results

A. Ablation Study

To systematically analyze the contribution of key components in the proposed model, we conduct ablation experiments on the speaker-conditioning strategy and the causal temporal modeling module. All systems use the same data split, frequency-domain representation, training objective, and evaluation protocol, with only the corresponding module replaced. This setup ensures that performance differences mainly arise from the conditioning-injection strategy or the sequence-modeling mechanism itself. Table 1 reports the SI-SNRi, PESQ, and STOI results under different settings.

Table 1. Ablation results of speaker-conditioning strategies and temporal modeling modules.

Ablation category	Strategy or module	SI-SNRi↑	PESQ↑	STOI↑
Speaker-conditioning strategy	Embedding multiplication	9.23	1.98	0.83
	Layer-wise cross-attention	11.25	2.03	0.88
Temporal module	Causal TCN	8.26	2.11	0.81
	Causal GRU	10.52	2.34	0.83
	SA-Mamba (Ours)	13.63	2.87	0.89

For speaker-conditioning strategies, layer-wise cross-attention improves all three metrics over embedding multiplication: SI-SNRi increases from 9.23 dB to 11.25 dB, corresponding to an absolute gain of 2.02 dB; PESQ increases from 1.98 to 2.03; and STOI increases from 0.83 to 0.88. These results indicate that simple embedding multiplication can only provide coarse modulation of intermediate representations, whereas layer-wise cross-attention explicitly retrieves target-speaker-related information from the enrollment utterance and is therefore more effective in target-speaker selectivity and speech intelligibility.

For temporal modeling modules, SA-Mamba outperforms both causal TCN and causal GRU across all evaluation metrics. Compared with causal TCN, SA-Mamba improves SI-SNRi, PESQ, and STOI by 5.37 dB, 0.76, and 0.08, respectively. Compared with causal GRU, the corresponding gains are 3.11 dB, 0.53, and 0.06. This shows that convolutional structures are limited by fixed receptive fields, while gated recurrent structures, despite their state-memory capability, remain insufficient for jointly modeling long-range dependencies and target-speaker conditions. SA-Mamba maintains long-range context through selective state recurrence and directly incorporates the speaker condition into the state-update process, thereby achieving better speech recovery quality and target consistency under causal inference constraints.

Overall, Table 1 shows that the performance improvement does not come from replacing a single module, but from the synergy between the speaker-conditioning mechanism and selective

state-space temporal modeling. Layer-wise retrieval-based conditioning provides more effective target-identity cues, while SA-Mamba further internalizes these cues into state writing, reading, and time-scale control. As a result, the model can more stably preserve target-speaker information and suppress interfering speech in streaming scenarios.

B. IEEE SLT 2026 REAL-TSE Challenge

We participated in the IEEE SLT 2026 REAL-TSE Challenge. According to the official blind test results shown in Table 2, the proposed SA-Mamba system achieved second place in the overall score. Compared with the challenge baseline systems, our method performs better in word error rate, target-speaker activity detection F1, speaker similarity, and DNSMOS, while substantially reducing computational complexity. These results demonstrate that speaker-conditioned state-space modeling can effectively improve streaming target speaker extraction in realistic scenarios. The real conversational mixture setting emphasized by REAL-TSE further highlights the importance of evaluating target speaker extraction under practical acoustic conditions [25].

Table 2: Official blind test results on the IEEE SLT 2026 REAL-TSE Challenge.

Methods	#Params(M)	MACs(G/s)	TER↓	F1↑	SIM↑	OVRL↑	Score↓
BSRNN_EMB_CAUSAL	25.05	38.91	0.809	0.821	0.422	1.714	18.2500
BSRNN_TFMAP_CAUSAL	27.24	38.57	0.805	0.836	0.456	1.670	15.0000
SA-Mamba (Ours)	15.89	30.01	0.713	0.849	0.524	2.151	4.7500

The latency of this algorithm primarily stems from STFT processing; with a window length of 32 ms and a hop size of 16 ms, a latency of 48 ms is incurred. This approximates the result calculated using the l Latency calculating script. The latency results checked with the Latency calculating script as follow:

Perturbation @ 1.000s | Response @ 0.952s | Future dependency = +47.7 ms

Perturbation @ 2.000s | Response @ 1.956s | Future dependency = +44.5 ms

Perturbation @ 1.234s | Response @ 1.188s | Future dependency = +45.7 ms

Perturbation @ 2.346s | Response @ 2.296s | Future dependency = +49.9 ms

Perturbation @ 3.459s | Response @ 3.413s | Future dependency = +45.7 ms

Summary

Min : +44.5 ms

Mean : +46.7 ms

Max : +49.9 ms

IV. Conclusion

This paper proposed a speaker-conditioned state-space modeling method for streaming target speaker extraction. The method directly introduces the target-speaker condition into the selective-

parameter generation process of Mamba, making the input matrix B, output matrix C, and discretization step Δt jointly depend on the current acoustic features and the enrollment representation. As a result, target-speaker-dependent information writing, reading, and time-scale control are achieved during state recurrence. Compared with conditioning methods that perform external feature fusion after temporal modeling, the proposed modeling strategy internalizes speaker selectivity into sequence-state updates and is better suited to low-latency target speech extraction in streaming scenarios. At the system level, we combine non-uniform band splitting, dual-path streaming modeling, and progressive speaker distillation to construct the SA-Mamba-TSE framework with 15.89 M parameters. Experimental results show that the proposed method achieves competitive target-speech recovery quality while maintaining causal inference and fixed-state cost. On the official blind test set of the IEEE SLT 2026 REAL-TSE Challenge, our system ranked second in the overall score, verifying the effectiveness of speaker-conditioned state-space models for realistic target speaker extraction. Future work will further explore learnable band partitioning, multi-speaker condition extension, and distillation from non-causal teacher models to improve robustness and generalization in complex acoustic environments.

References

[1] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in Proc. ICASSP, 2018, pp. 5554-5558.

[2] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in Proc. Interspeech, 2019, pp. 2728-2732.

[3] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in Proc. ICASSP, 2023.

[4] Y. Luo and J. Yu, "Music source separation with band-split RNN," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 31, pp. 3174-3187, 2023.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in Proc. Interspeech, 2020, pp. 3830-3834.

[6] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in Proc. ICLR, 2022.

[7] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in Proc. NeurIPS, 2023.

[8] K. Li, G. Chen, Y. Hu, and D. Wang, "SPMamba: State-space model is all you need in speech separation," arXiv:2404.02063, 2024.

[9] R. Chao, W.-Y. Tsai, H.-M. Wang, and Y. Tsao, "An investigation of incorporating Mamba for speech enhancement," arXiv:2405.06573, 2024.

[10] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in Proc. ICASSP, 2020, pp. 696-700.

[11] T. Serre et al., "A lightweight dual-stage framework for personalized speech enhancement based on DeepFilterNet2," arXiv:2404.08022, 2024.

[12] C. Subakan et al., "Attention is all you need in speech separation," in Proc. ICASSP, 2021, pp. 21-25.

[13] J. Yu et al., "High fidelity speech enhancement with band-split RNN," in Proc. ICASSP, 2023.

[14] F. Hao et al., "DSINet: Towards real-time target speaker extraction with dynamic speaker information fusion," in Proc. ICASSP, 2024.

[15] H. Ma et al., "Enhancing intelligibility for generative target speech extraction via joint optimization with target speaker ASR," IEEE/ACM Trans. Audio, Speech, Lang. Process., 2025.

[16] M. Thakker et al., "Fast real-time personalized speech enhancement: End-to-end enhancement network (E3Net) and knowledge distillation," in Proc. Interspeech, 2022.

[17] A. Navon et al., "FlowTSE: Target speaker extraction with flow matching," arXiv:2505.14465, 2025.

[18] M. Delcroix et al., "Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam," in Proc. ICASSP, 2020, pp. 691-695.

[19] B. Tang, B. Zeng, and M. Li, "LauraTSE: Target speaker extraction using auto-regressive decoder-only language models," in Proc. ICASSP, 2025.

[20] Y. Wang et al., "Metis: A foundation speech generation model with masked generative pre-training," arXiv:2502.03128, 2025.

[21] K. Zhang et al., "Multi-level speaker representation for target speaker extraction," in Proc. ICASSP, 2024.

[22] W. Wang and M. Li, "Online neural speaker diarization with target speaker tracking," IEEE/ACM Trans. Audio, Speech, Lang. Process., 2024.

[23] X. Le et al., "Personalized speech enhancement combining band-split RNN and speaker attentive module," in Proc. ICASSP, 2023.

[24] T. Parnamaa and A. Saabas, "Personalized speech enhancement without a separate speaker embedding model," in Proc. ICASSP, 2024.

[25] S. Li et al., "REAL-T: Real conversational mixtures for target speaker extraction," in Proc. ICASSP, 2025.

[26] C. Deng et al., "Robust speaker extraction network based on iterative refined adaptation," in Proc. ICASSP, 2021.

[27] B. Zeng and M. Li, "SEF-PNet: Speaker encoder-free personalized speech enhancement with local and global contexts aggregation," in Proc. ICASSP, 2024.

[28] K. Zhang et al., "Speaker extraction with detection of presence and absence of target speakers," in Proc. Interspeech, 2025.

[29] H. Yin et al., "SpeakerLM: End-to-end versatile speaker diarization and recognition with multimodal large language models," 2025.

[30] Y. Ju et al., "TEA-PSE 2.0: Sub-band network for real-time personalized speech enhancement," in Proc. IEEE Spoken Language Technology Workshop (SLT), 2023, pp. 472-479.

[31] Y. Ju et al., "TEA-PSE 3.0: Tencent-Ethereal-Audio-Lab personalized speech enhancement system for ICASSP 2023 DNS Challenge," in Proc. ICASSP, 2023, pp. 1-2.

[32] M. Xu et al., "TIGER: Time-frequency interleaved gain extraction and reconstruction for efficient speech separation," in Proc. ICLR, 2025.

[33] J. Yu et al., "TSpeech-AI system description to the 5th deep noise suppression (DNS) challenge," in Proc. ICASSP, 2023.

[34] X. Hao et al., "Typing to listen at the cocktail party: Text-guided target speaker extraction," IEEE/ACM Trans. Audio, Speech, Lang. Process., 2025.

[35] B. Zeng and M. Li, "USEF-TSE: Universal speaker embedding free target speaker extraction," IEEE/ACM Trans. Audio, Speech, Lang. Process., 2025.

[36] B. Zeng and M. Li, "Universal speaker embedding free target speaker extraction and personal voice activity detection," arXiv:2501.03612, 2025.

[37] S. Wang et al., "WeSep: A scalable and flexible toolkit towards generalizable target speaker extraction," in Proc. ICASSP, 2025.

[38] H. Wang et al., "Wespeaker: A research and production oriented speaker embedding learning toolkit," in Proc. ICASSP, 2023, pp. 1-5.

[39] C. Sun and B. Qin, "X-CrossNet: A complex spectral mapping approach to target speaker extraction with cross attention speaker embedding fusion," arXiv:2411.13811, 2024.

[40] Z. Wang et al., "A framework for unified real-time personalized and non-personalized speech enhancement," in Proc. ICASSP, 2023.

[41] A. Sivaraman, S. Kim, and M. Kim, "Personalized speech enhancement through self-supervised data augmentation and purification," in Proc. ICASSP, 2024.

[42] G. Gao, B. Zeng, X. Wang, W. Yin, and Y. Chen, "Speech enhancement based on speaker embedding," Comput. Eng., vol. 48, no. 3, pp. 688-692, Mar. 2022.