Dual-Path Attention-Enhanced Mamba Network for Efficient Long-Sequence Speech Separation

Shasha Long[1], Junmei Yang[1*], Qianyu Yan[1], Delu Zeng[1]
[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
202321012845@mail.scut.edu.cn, *yjunmei@scut.edu.cn, 202420111685@mail.scut.edu.cn, dlzeng@scut.edu.cn

***Abstract*—This paper proposes a Dual-path Attention-enhanced Mamba Network (DAMamba) for efficient long-sequence speech separation. To address Mamba's limitations in local detail modelling and channel redundancy, we introduce three key components: a Multi-Scale Dilated Convolution module (MSDC) for local feature capture, a Segment-wise Channel Attention-enhanced Mamba Block (SAEMB) for channel calibration, and a dual-path enhancement architecture for global-local collaboration. Experiments on LRS2-2Mix and Libri2Mix show that DAMamba achieves SI-SNRi of 17.5 dB with only 26.9 percent parameters of SepFormer, confirming its efficiency and superiority.**

***Index Terms*—dual-path, speech separation, state space model, attention mechanism**

I. Introduction

Speech separation aims to extract independent speaker sources accurately from mixed speech signals containing multiple speakers. As a fundamental front-end task in speech processing, its effectiveness critically influences downstream applications such as speech recognition and speaker verification. Recently, the introduction of deep learning techniques has propelled rapid advancement in speech separation research, with time-domain speech separation methods achieving breakthrough progress. The Conv-TasNet proposed by Luo [1] established the end-to-end time-domain speech separation paradigm, whose encoder-separation network-decoder three-level architecture became the foundation for subsequent research. However, traditional CNNs struggle to capture long-range dependencies in audio signals while preserving local feature sensitivity, owing to their limited receptive fields. This limitation can result in inadequate global temporal context modelling for complex speech mixture [2], [3]. Recurrent Neural Networks (RNNs), with their sequence modelling capability, are theoretically suitable for speech signal processing [4]–[6]. The DPRNN [4] based on dual-path structure effectively alleviates the conflict between long-term dependency and computational efficiency through block-wise alternating processing strategies. However, RNNs still face gradient vanishing or explosion during training, as well as computational latency due to sequential processing. With the success of Transformer in sequence modelling tasks [7]–[9], speech separation models based on self-attention mechanism, such as SepFormer [10] and DPTNet [5], MossFormer [11], and MossFormer2 [12] achieve excellent performance through explicit computation of global interactions. However, the computational complexity of self-attention mechanism grows quadratically with sequence length, posing severe computational challenges when processing long speech sequences of thousands of frames. Moreover, existing dual-path studies based on attention mechanisms [13], [14] suffer from high model complexity and insufficient parameter lightweight.

To overcome these bottlenecks, novel sequence modelling techniques based on State Space Models (SSMs) [15] have emerged. Among them, the Mamba model [16] reduces computational complexity to linear while maintaining modelling capabilities comparable to Transformer by introducing selective scanning mechanisms, becoming a significant breakthrough in long-sequence modelling [17]–[20]. Researchers have adapted Mamba for speech separation, proposing SPMamba [21], Dual Path Mamba [22], Mamba-TasNet [23], which show promising results for long sequences. Nevertheless, current Mamba models still have limitations in speech separation applications: First, Speech signals contain multi-scale temporal structures such as fundamental frequency and harmonics, where Mamba models may be less sensitive than convolutional layers in modelling local fine-grained patterns (such as phonemes, short-term spectral variations); Second, Existing Mamba models primarily focus on temporal dependency modelling, lacking mechanisms to calibrate importance differences across feature channels, making them vulnerable to noise-related redundant channels; Third, within dual-path frameworks, block-wise processing may lead to insufficient capture of long-distance associations between block-level features.

To address these issues, this paper proposes a Dual-path Attention-enhanced Mamba Network (DAMamba) based on the classic encoder-separation network-decoder architecture. Our main contributions are as follows:

- **Multi-scale Gated Dilated Convolution module (MSDC)**: Using three parallel gated branches, it synchronously captures high-, medium-, and low-scale local temporal features, enhancing multiscale representation while avoiding gradient issues in sequential dilated convolutions.
- **Segment-wise Channel Attention-enhanced Mamba Block (SAEMB)**: This block integrates a Segment-

* is the corresponding author.

wise Cumulative Channel Attention (SCCA) module with Mamba to adaptively weight channel importance via segment-wise calibration, addressing Mamba's uniform channel treatment and improving feature discriminability.

- **Efficient Collaborative Dual-path Enhancement Architecture**: It constructs dedicated pathways for local enhancement and global interaction, enabling unified modeling of multiscale features, long-term dependencies, and channel calibration through coordinated optimization of MSDC, SAEMB, and lightweight inter-block attention.

Experiments on LRS2-2Mix and Libri2Mix datasets show that the proposed model achieves excellent performance of SI-SNRi 17.5 dB with only 26.9 percent of SepFormer's parameters, significantly outperforming existing baseline models, confirming the model's efficiency and superiority.

II. Proposed Method

A. Overall Architecture

The model follows an end-to-end architecture, as shown in Fig. 1. The single-channel mixed speech x is mapped to high-dimensional features h_x by the encoder. After block processing, these features enter stacked Enhanced Dual-Path Blocks (EDPB) for feature modelling, producing target speaker masks m_1, m_2. These masks are multiplied with encoder output features to obtain separated features, which the decoder transforms back to time-domain waveforms $\hat{s}_i$.

1) Encoder: The encoder consists of a 1D convolution with kernel size 16, stride 8, and 128 output channels, followed by ReLU activation function. Given mixed speech $x \in \mathbb{R}^{1\times L}$, it produces:

$$h_x = \text{ReLU}(Conv1D(x)) \tag{1}$$

2) Separation Network: The core separator consists of R serial Enhanced Dual-Path Blocks (EDPB). As shown in Fig. 2, each EDPB block contains Intra and Inter paths. Input features are first segmented into blocks and reshaped. The Intra path performs local feature enhancement via the Multi-Scale Dilated Convolution module (MSDC) and Segment-wise Attention-enhanced Mamba module (SAEMB) to achieve multiscale feature extraction and structured temporal modelling. The Inter path models global dependencies through SAEMB, then strengthens inter-block associations via a Block-level Feature Interaction Attention (SAB) module. Enhanced features are fused with residual connections.

3) Decoder: After encoding, features pass through the separation network to output masks $m_i, i \in \{1, 2\}$ for each source (corresponding to two speakers). Masks m_i are element-wise multiplied with encoder output h_x, and the decoder reconstructs the waveform of time-domain signals by applying transposed convolution to the product:

$$\hat{s}_i = \text{TransposeConv1D}(h_x \odot m_i) \in \mathbb{R}^{1\times L} \tag{2}$$

where $\odot$ represents element-wise multiplication. The decoder's stride and convolution kernel size are consistent with the encoder.

B. Multi-Scale Gated Dilated Convolution Module (MSDC)

Speech signals contain rich multiscale temporal structures, from short-term phonemes to long-term sentences, with significant differences in temporal resolution. Traditional CNNs [1]–[3] often use fixed receptive fields or serial dilated convolutions [24] to expand receptive fields, but single layers struggle to capture multiscale information simultaneously, and serial structures suffer from long gradient propagation paths. To address this, we design the Multi-Scale Gated Dilated Convolution module (MSDC), as shown in Fig. 2(b). It contains three parallel branches with dilation rates 1, 2, 4 and kernel sizes of 3, 5, and 7, respectively. Each branch includes a dilated convolution for feature extraction and a gated convolution (using $Sigmoid$ activation)for adaptive selection. The gating mechanism dynamically adjusts branch importance, enabling flexible multiscale fusion. Outputs are concatenated and fused via 1D convolution.

For branch i:

1) Dilated convolution feature extraction:

$$X_i = \text{DilatedConv1d}(X'_{\text{intra-in}}, k = k_i, d = d_i) \tag{3}$$

2) Gating weight generation: consistent parameters with feature extraction convolution:

$$G_i = \sigma(\text{DilatedConv1d}(X'_{\text{intra-in}}, k = k_i, d = d_i)) \tag{4}$$

3) Gated selection output:

$$X_{g_i} = X_i \odot G_i \tag{5}$$

Dilation rates d_i cover feature extraction from high-frequency fine-grained to low-frequency coarse-grained features. Gated convolution generates weight vectors to weight the output of filtered high-quality features. Here X_i is the output of feature extraction convolution for branch i, G_i is the weight vector generated by gated convolution for branch i, X_{g_i} is the gated-filtered feature for branch i.

4) Feature concatenation and fusion: Concatenate gated-filtered features X_{g1}, X_{g2}, X_{g3} from three branches along the channel dimension, then compress the channel number to C through 1×1 convolution, achieving lightweight fusion and dimension unification of multiscale features, outputting fused feature $X_{\text{ms}} \in \mathbb{R}^{B\times C\times L}$:

$$X_{\text{ms}} = \text{Conv1D}(\text{Concat}(X_{g1}, X_{g2}, X_{g3})) \tag{6}$$

The core advantages of the MSDC module are: First, the differentiated parameter design of the three branches precisely matches the temporal characteristics of high, medium, and low-frequency components of speech, achieving comprehensive coverage of multiscale features; Second, introducing the gating mechanism as a core innovation dynamically adjusts feature weights through per-branch gating units, strengthening effective information and suppressing noise redundancy, solving the problem of insufficient feature discriminability caused by average fusion in traditional multiscale convolutions; Third, targeted enhancement of intra-block local multiscale features provides high-quality input for subsequent long-term dependency modelling and channel calibration.

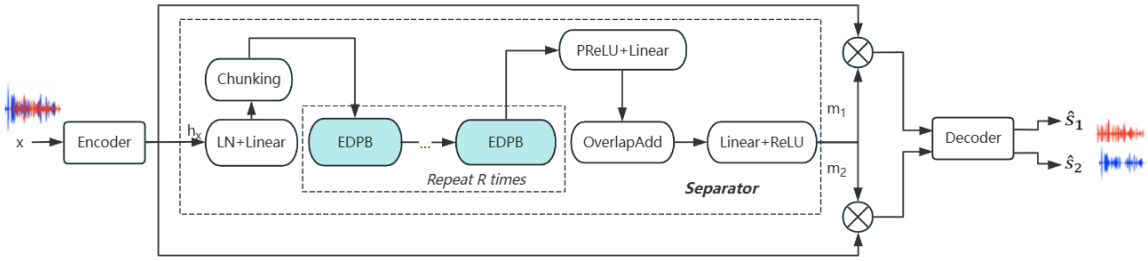

Fig. 1: Overall architecture of the proposed DAMamba model.

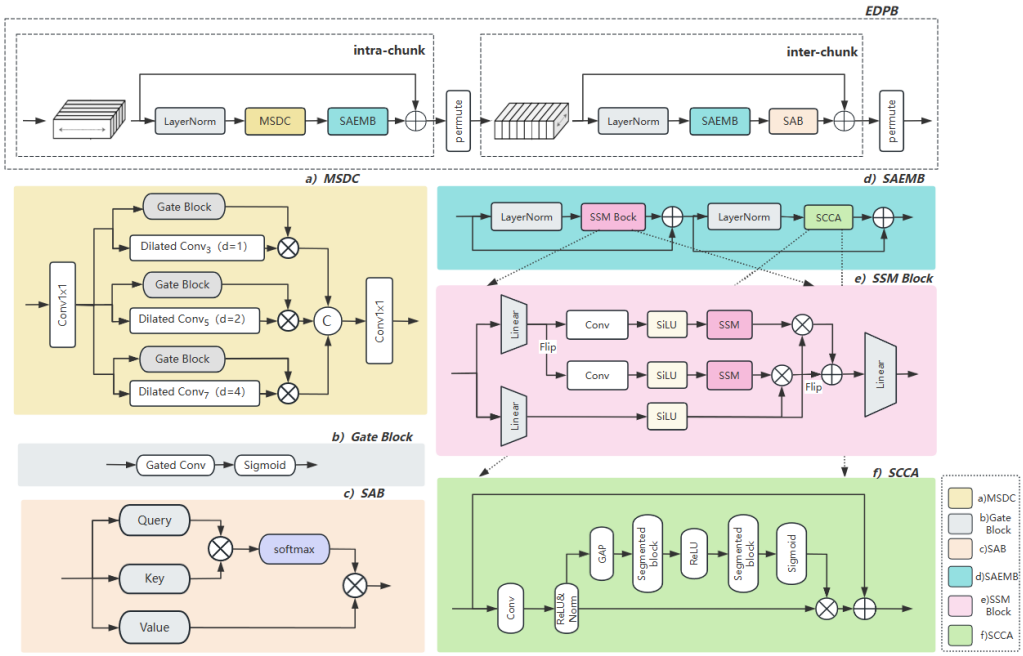

Fig. 2: Separator workflow diagram showing the dual-path processing within each Enhanced Dual-Path Block (EDPB). The input features are processed through intra-path for local feature enhancement and inter-path for global dependency modelling.

C. Segment-wise Attention-Enhanced Mamba Module

While Mamba can efficiently model long-term dependencies, it treats feature channel dimensions equally, making it difficult to distinguish importance differences between different channels, potentially leading to interference from redundant channel information such as noise-related channels on effective feature representation [25]. Squeeze-and-excitation mechanisms [26] typically use two fully connected layers to model all channel relationships, which may lead to computational inefficiency and fine-grained information loss. Inspired by SpinalNet [27], we propose the Segment-wise Cumulative Channel Attention module (SCCA), achieving fine-grained calibration through segments of channel dimensions.

1) SSM Block module: achieves global context modelling of sequence data through parallel processing of forward and reverse sequence scans. This module includes linear projection layers, convolution, SiLU activation function, the selective state space model (SSM), and sequence flip operations.

As shown in Fig. 2(c), input passes through a linear projection, then undergoes feature processing and weighting separately. The feature processing branch processes forward and reverse scans in parallel; their outputs are weighted separately, the reverse sequence is flipped back to the original sequence order, and both are averaged and fused.

Forward processing:

$$y^{+} = \text{SSM}(\text{SiLU}(\text{Conv}(\text{Linear}(x)))) \tag{7}$$

Reverse processing:

$$y^{-} = \text{SSM}(\text{SiLU}(\text{Conv}(\text{Flip}(\text{Linear}(x))))) \tag{8}$$

Weight:

$$a = \text{SiLU}(\text{Linear}(x)) \tag{9}$$

Weighted average fusion output:

$$y = \text{Linear}\left(\frac{a \otimes y^{+} + a \otimes \text{Flip}(y^{-})}{2}\right) \tag{10}$$

2) SCCA module: As shown in Fig. 2(f), for input $Z \in \mathbb{R}^{B\times L\times D}$, where B is batch size, L is sequence length, and D is channel number (feature dimension), SCCA module's core idea is hierarchical processing and cumulative interaction. The process is as follows: First, after aggregating spatial information through global average pooling of input Z, uniformly segment along the channel dimension into G segments, obtaining $z = [z_1, z_2, \ldots, z_G]$, each segment length $\frac{D}{G}$.

Segment-wise cumulative interaction: First segment input z_1 passes through fully connected layer f_1 to get output s_1; concatenate previous segment output with current segment original features as input to corresponding fully connected layer, subsequent segments follow similarly:

$$s_1 = f_1(z_1), \quad s_g = f_g(\text{Concat}[s_{g-1}; z_g]), \quad g = 2, 3, \ldots, G \tag{11}$$

Feature fusion and activation: Concatenate all segment outputs $s = [s_1, s_2, \ldots, s_G]$, after ReLU nonlinear activation, restore to original dimension D through symmetric fully connected layer, finally generate channel attention weights $a \in \mathbb{R}^{B\times 1\times D}$ through Sigmoid. Element-wise multiply attention weights a with original input features Z, achieving channel-level feature calibration:

$$\widetilde{Z} = Z \odot a \tag{12}$$

This design achieves hierarchical flow and cumulative interaction of channel information. Compared to standard SE's two-layer bottleneck structure, it can establish richer nonlinear channel dependencies while avoiding information loss from excessive compression. Mamba's temporal dependency modelling provides feature inputs rich in temporal information for channel calibration, while SCCA's channel filtering can strengthen the effectiveness of Mamba output features, forming dual-dimensional feature enhancement in both temporal and channel dimensions, enabling the model to have selectivity in both spatial and channel dimensions, thereby improving model performance. This module addresses Mamba's indiscriminate treatment of channels through channel-level optimization of Mamba output features, providing more discriminative input for subsequent dual-path feature enhancement.

D. Block-level Feature Interaction Attention Module

To further enhance global context modelling, we introduce a single-head standard self-attention module (SAB) after SAEMB. Its Query, Key, and Value are derived from SAEMB's output $H \in \mathbb{R}^{(B\times K)\times S\times D}$ of the Segment-wise Attention-enhanced Mamba module (where S is the number of blocks, D is the feature dimension), specifically calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{13}$$

where d_k is the dimension of Q/K, reducing computational overhead through dimension reduction.

SAB and Mamba form an implicit-explicit complementary pair: Mamba excels at processing continuous patterns, while self-attention explicitly models inter-block associations, particularly effective for capturing discrete yet semantically related fragment relationships.

III. Experiments

A. Experimental Setup and Parameters

Experiments were conducted on an NVIDIA RTX 4090 GPU. We used the Adam [28] optimizer with an initial learning rate of $1.5 \times 10^{-4}$and a cosine annealing scheduler with warm restarts. Training lasted 200 epochs with a batch size of 1. Gradient clipping was applied with a threshold of 5. The loss function was negative Scale-Invariant Signal-to-Noise Ratio (SI-SNR), combined with utterance-level Permutation Invariant Training (uPIT). Early stopping was employed based on validation performance.

For fair comparison with other baseline models, key hyperparameter configurations are as follows: encoder convolution kernel size 16, stride 8, output channels D set to 128; block size K (length of a single block) in dual-path processing is 250; number of stacked Enhanced Dual-Path Blocks (EDPB) in separator R is 8; core Mamba-related parameters: SSM hidden state dimension 16, expansion factor 2, 1D convolution kernel size $d_{conv} = 4$. To mitigate overfitting, the dropout rate for attention and fully connected layers was set to 0.1.

1) Evaluation Metrics: This paper uses widely accepted objective evaluation metrics in speech separation field, including Scale-Invariant Signal-to-Noise Ratio (SI-SNR) [29], Signal-to-Distortion Ratio (SDR) [30], and their corresponding improvements (SI-SNRi, SDRi [31]), with higher scores indicating better separated signal quality. Additionally, the number of parameters (Parameters) and Multiply-Accumulate Operations (MACs) measure model complexity, with lower values indicating a more efficient model. The SI-SNR for a target signal s and its estimate $\hat{s}$ is defined as:

$$\text{SI-SNR}(s, \hat{s}) = 10\log_{10}\left(\frac{\|s_{\text{target}}\|^2}{\|\hat{s} - s_{\text{target}}\|^2}\right) \tag{14}$$

$$s_{\text{target}} = \frac{\langle \hat{s}, s\rangle}{\|s\|^2}s \tag{15}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product.

The SI-SNR improvement (SI-SNRi) is computed as the difference between the SI-SNR of the separated signal and that of the mixed signal:

$$\text{SI-SNRi} = \text{SI-SNR}(s, \hat{s}) - \text{SI-SNR}(s, s_{\text{mix}}) \tag{16}$$

Similarly, the SDR and its improvement are defined as:

$$\text{SDR}(s, \hat{s}) = 10\log_{10}\left(\frac{\|s\|^2}{\|s - \hat{s}\|^2}\right) \tag{17}$$

$$\text{SDRi} = \text{SDR}(s, \hat{s}) - \text{SDR}(s, s_{\text{mix}}) \tag{18}$$

where $\|\cdot\|^2$ represents L2 norm, s_{target} is the true target signal, e_{noise} is residual noise, $\hat{s}$ is the model's estimated output signal, s_{mix} is the original mixed signal.

TABLE I: Separation performance comparison on LRS2-2Mix and Libri2Mix datasets

Model	LRS2-2Mix		Libri2Mix		Params (M)	MACs (G/s)
	SI-SNRi (dB)	SDRi (dB)	SI-SNRi (dB)	SDRi (dB)		
BLSTM-TasNet [29]	6.1	6.8	7.9	8.7	23.6	43.0
SuDoRM-RF 1.0x [2]	11.0	11.4	13.5	14.0	2.7	4.6
SuDoRM-RF 2.5x [2]	11.3	11.7	14.0	14.4	6.4	10.1
Conv-TasNet [1]	10.6	11.0	12.2	12.7	5.6	10.2
DPRNN [4]	12.7	13.0	16.1	16.6	2.7	85.3
DPTNet [5]	13.3	13.6	16.7	17.1	2.7	102.5
SepFormer [10]	13.5	13.8	16.5	17.0	26.0	86.9
TDANet [34]	13.2	13.5	17.4	17.9	2.3	4.7
S4M [36]	15.3	15.5	16.9	17.4	3.6	38.7
Ours	**15.7**	**16.1**	**17.5**	**17.8**	7.0	56.4

2) Datasets: Experiments were conducted on two standard speech separation datasets.

LRS2-2Mix [32]: Mixed speech dataset generated based on open-source dataset LRS2 [33]. Each mixture is 2 seconds long with a 16 kHz sampling rate and SNR between -5 dB and 5 dB. The training, validation, and test sets contain 20,000, 5,000, and 3,000 samples, respectively.

Libri2Mix [35]: Two-speaker mixed dataset derived from LibriSpeech corpus, this paper uses train-100 subset of training set. With a sampling rate 8kHz, its training, validation, and test sets contain 13,900, 3,000, and 3,000 samples, respectively.

B. Comparative Experiments

Table I shows the performance comparison of the proposed model against baseline models on the LRS2-2Mix and Libri2Mix datasets.

On the LRS2-2Mix dataset, the proposed model achieves SI-SNRi of 15.7 dB and SDRi of 16.1 dB, surpassing traditional models, including CNN-based Conv-TasNet, RNN-based DPRNN, Transformer-based SepFormer, while also outperforming SSM-based S4M (SI-SNRi improvement of 0.4 dB).

On the Libri2Mix dataset, the proposed model achieves SI-SNRi of 17.5 dB and SDRi of 17.8 dB, significantly outperforming all compared models, with parameter count (7.0 M) only 26.9% of SepFormer (26.0 M), MACs (56.4 G/s) much lower than DPRNN (85.3 G/s) and DPTNet (102.5 G/s), proving the proposed model has higher computational efficiency while ensuring separation performance. The proposed model demonstrates stronger robustness in noise environments through multi-scale feature extraction and structured channel calibration, validating the effectiveness of the dual-path attention enhancement architecture.

C. Ablation Studies

To validate the contribution of each core module, ablation experiments were conducted on the Libri2Mix dataset. Using the complete Dual path Mamba model as a baseline, key components were sequentially removed or replaced. Results are shown in Table II:

TABLE II: Component ablation study on Libri2Mix dataset

MSDC	SCCA	SAB	SI-SNRi (dB)	SDRi (dB)
✓	—	—	16.9	17.3
✓	✓	—	17.3	17.6
✓	—	✓	17.1	17.4
—	✓	—	16.4	16.9
—	✓	✓	16.9	17.3
—	—	✓	16.1	16.6
✓	✓	✓	**17.5**	**17.8**

Removing MSDC causes an SI-SNRi drop of 0.6 dB, confirming the importance of multiscale feature capture. Removing SCCA causes a drop of 0.4 dB, validating the effectiveness of fine-grained channel calibration. Removing SAB causes a drop of 0.2 dB, indicating the complementary role of explicit inter-block association modeling to Mamba's implicit modelling. Removal of any module causes certain performance degradation, indicating the importance of each module's design and collaborative operation.

IV. Conclusion

This paper proposes an efficient Dual-Path Attention-Enhanced Mamba Network (DAMamba) for long-sequence speech separation. Through the collaborative design of the Multi-Scale Dilated Convolution module, the Segment-wise Channel Attention-Enhanced Mamba Block, the model achieves unified modelling of multiscale temporal structures, long-range dependencies, and channel discriminability of speech signals. Experimental results demonstrate that the proposed model achieves leading separation performance on both LRS2-2Mix and Libri2Mix datasets while maintaining low model complexity. Ablation studies validate the effectiveness of each innovative module. Compared to existing methods, DAMamba achieves higher separation performance with only 26.9 percent of SepFormer's parameters, confirming its superiority in efficient speech separation tasks.

Future work will explore more flexible multiscale feature fusion strategies and extend the model to scenarios with more speakers and real-time processing applications, further enhancing its practical value.

REFERENCES

[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, Aug. 2019.

[2] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient Networks for Universal Audio Source Separation," in *IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1-6.

[3] X. Hu, K. Li, W. Zhang, Y. Luo, J.-M. Lemercier, and T. Gerkmann, "Speech separation using an asynchronous fully recurrent convolutional neural network," in *Advances in Neural Information Processing Systems*, 2021, pp. 22509-22522.

[4] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46-50.

[5] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Annual Conference of the International Speech Communication Association*, 2020, pp. 2642-2646.

[6] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221-3236, 2023.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.

[8] J. Achiam *et al.*, "Gpt-4 technical report," 2023.

[9] A. Liu *et al.*, "Deepseek-v3 technical report," 2024.

[10] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 21-25.

[11] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1-5.

[12] S. Zhao *et al.*, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10356-10360.

[13] J. Yang, B. Zhang, L. Yang, D. Zeng, "A Single-Channel Speech Separation Model Based on Time-Domain Comprehensive Attention Mechanism," *Journal of South China University of Technology (Natural Science Edition)*, vol. 54, no. 1, pp. 70-82, 2026.

[14] L. Yang, B. Zhang, J. Yang, D. Zeng, "TFA-Conformer Based Network for Short Utterance Speaker Recognition," *Acta Electronica Sinica*, vol. 53, no. 08, pp. 2658-2667, 2025.

[15] K. Goel, A. Gu, C. Donahue, and C. Ré, "It's raw! Audio generation with state-space models," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 7616-7633.

[16] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, arXiv:2312.00752.

[17] J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing long-range dependency for biomedical image segmentation," 2024, arXiv:2401.04722.

[18] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li, "Mamba UNet: UNet-like pure visual Mamba for medical image segmentation," 2024, arXiv:2402.05079.

[19] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation," in *Proc. 27th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2024, pp. 578-588.

[20] Y. Yang, Z. Xing, and L. Zhu, "Vivim: A video vision Mamba for medical video object segmentation," 2024, arXiv:2401.14168.

[21] K. Li, G. Chen, R. Yang, and X. Hu, "SPMamba: Leveraging Long-Sequence Modeling with State Space Models for Speech Separation," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 2025, pp. 1-6.

[22] X. Jiang, C. Han, and N. Mesgarani, "Dual-path Mamba: Short and Long-term Bidirectional Selective Structured State Space Models for Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1-5.

[23] X. Jiang, Y. A. Li, A. N. Florea, C. Han, and N. Mesgarani, "Speech slytherin: Examining the performance and efficiency of mamba for speech separation, recognition, and synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1-5.

[24] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *International Conference on Learning Representations*, 2016.

[25] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "MambaIR: A simple baseline for image restoration with state-space model," in *Proc. ECCV*, 2024, pp. 222-241.

[26] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011-2023, Aug. 2020.

[27] H. M. D. Kabir *et al.*, "SpinalNet: Deep Neural Network With Gradual Input," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 5, pp. 1165-1177, Oct. 2023.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv:1412.6980.

[29] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696-700.

[30] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462-1469, July 2006.

[31] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626-630.

[32] K. Li, F. Xie, H. Chen, K. Yuan, and X. Hu, "An Audio-Visual Speech Separation Model Inspired by Cortico-Thalamo-Cortical Circuits," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6637-6651, Oct. 2024.

[33] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2018.

[34] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," in *International Conference on Learning Representations*, 2023, pp. 1-10.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206-5210.

[36] C. Chen, C.-H. H. Yang, K. Li, Y. Hu, P.-J. Ku, and E. S. Chng, "A neural state-space modeling approach to efficient speech separation," in *Annual Conference of the International Speech Communication Association*, 2023, pp. 3784-3788.