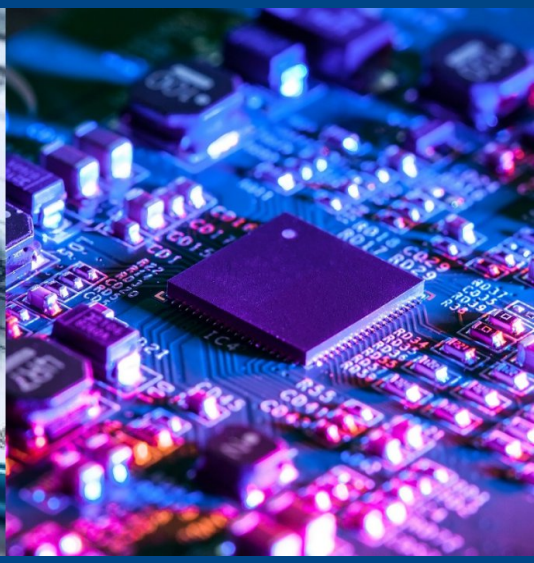
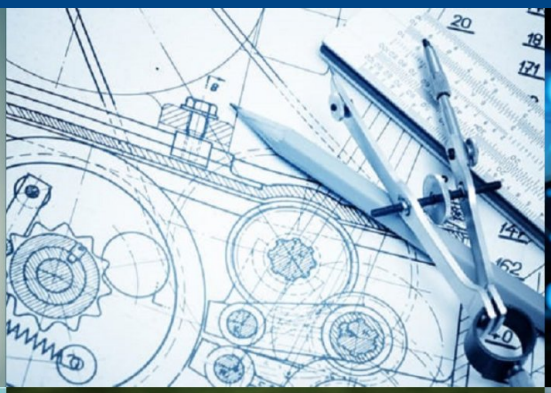
ISSN: 2582-7219

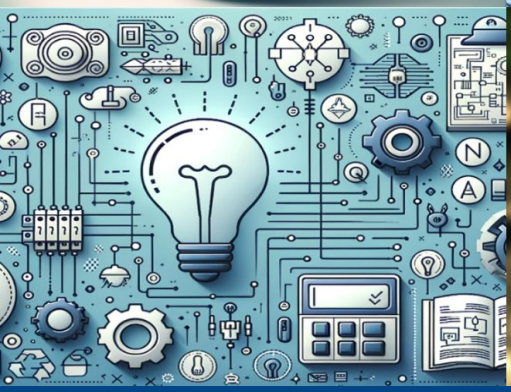




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904232

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Enhanced Mamba SEU-Net for Monaural Speech Enhancement

J. Surekha[1], J. Pavithra[2], K. Navya Sree[3], K. Sai Naik[4]

UG Student, Department of ECE, R.V.R & J.C College of Engineering, Chowdavaram, India[1]

UG Student, Department of ECE, R.V.R & J.C College of Engineering, Chowdavaram, India[2]

UG Student, Department of ECE, R.V.R & J.C College of Engineering, Chowdavaram, India[3]

UG Student, Department of ECE, R.V.R & J.C College of Engineering, Chowdavaram, India[4]

ABSTRACT: Monaural speech enhancement aims to recover clean speech from a single-channel noisy recording, which remains a challenging task in real-world acoustic environments characterized by non-stationary noise and diverse interference patterns. Conventional deep learning approaches often struggle to simultaneously model long-range temporal dependencies and preserve spectral detail without incurring high computational cost. In this paper, we propose Enhanced Mamba SEU-Net (EM-SEU-Net), a novel hybrid deep learning architecture that integrates the Mamba Selective State Space Model (SSM) with a Squeeze-and-Excitation U-Net (SEU-Net) encoder-decoder framework for effective monaural speech enhancement. The proposed architecture introduces three key innovations: (i) a Multi-Scale Feature Fusion (MSFF) module at the network bottleneck that aggregates temporal context at multiple dilation scales prior to sequence modeling; (ii) stacked Mamba SSM blocks that efficiently capture global temporal dynamics through input-dependent selective state updates with linear computational complexity; and (iii) Gated Skip Connections (GSC) that mitigate the propagation of noise-corrupted encoder activations across the encoder-decoder bypass paths. Squeeze-and-Excitation attention is applied at each encoder and decoder stage to perform adaptive channel-wise spectral recalibration. Experimental evaluations on the VCTK-DEMAND and Deep Noise Suppression (DNS) Challenge benchmark datasets demonstrate that the proposed EM-SEU-Net achieves superior enhancement performance compared to existing state-of-the-art approaches across perceptual quality, intelligibility, and signal distortion measures. Comprehensive ablation experiments further validate the independent contribution of each proposed component. The results confirm that EM-SEU-Net provides a practical, efficient, and high-performing solution for real-world monaural speech enhancement.

KEYWORDS: Monaural speech enhancement, Mamba state space model, Squeeze-and-excitation network, U-Net, Multi-scale feature fusion, Gated skip connection, Deep noise suppression.

I. INTRODUCTION

Monaural speech enhancement (MSE) the recovery of clean speech from a single-channel noisy observation is a fundamental problem in audio signal processing with broad practical applications including automatic speech recognition (ASR), voice communication systems, hearing aids, and teleconferencing platforms. In real-world environments, speech is commonly contaminated by diverse non-stationary noise sources such as traffic noise, babble, office interference, and music, posing significant challenges for enhancement systems that must generalize reliably across unseen acoustic conditions.

Classical signal processing approaches, including spectral subtraction [1], Wiener filtering [2], and minimum mean square error (MMSE)-based estimators [3], provided foundational contributions to the field. These methods rely on statistical assumptions about speech and noise distributions that are frequently violated in practice. The advent of deep learning transformed speech enhancement research, with feedforward deep neural networks (DNNs) [4], recurrent neural networks (RNNs) [5], and convolutional neural networks (CNNs) [6] demonstrating substantial improvements over classical approaches.

The introduction of encoder-decoder architectures inspired by the U-Net framework [7] further advanced the field by enabling hierarchical multi-scale feature extraction with skip connections. Models such as DCUNET [8], DCCRN [9],

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

and DPCRN [10] demonstrated strong performance by exploiting both magnitude and phase information in the complex spectrogram domain. Despite their success, transformer-based architectures incur quadratic complexity $O(T^2)$ in sequence length, limiting scalability to long audio sequences and constraining real-time deployment.

State Space Models (SSMs), and in particular the Mamba architecture [12], offer a compelling alternative. Mamba employs input-dependent selective state space parameters and a hardware-aware parallel scan to achieve linear $O(T)$ complexity while modeling global temporal dependencies. These properties are especially advantageous for speech processing, where temporal dependencies span hundreds of milliseconds.

Motivated by these observations, we propose Enhanced Mamba SEU-Net (EM-SEU-Net), a hybrid architecture synergistically combining the Mamba SSM with a Squeeze-and-Excitation U-Net backbone. Our main contributions are:

(1) We propose EM-SEU-Net, integrating Mamba SSM with a Squeeze-and-Excitation U-Net encoder-decoder for monaural speech enhancement, combining linear-complexity temporal modeling with hierarchical spectral feature extraction.
(2) We introduce a Multi-Scale Feature Fusion (MSFF) module at the bottleneck that captures temporal context at multiple dilation rates prior to SSM processing.
(3) We design Gated Skip Connections (GSC) that apply learned soft gating to encoder bypass paths, suppressing noise-corrupted activations from entering the decoder.
(4) We conduct comprehensive experiments on VCTK-DEMAND and DNS Challenge benchmarks with detailed ablation studies, confirming state-of-the-art performance at competitive model size.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the EM-SEU-Net architecture. Section 4 describes the experimental setup. Section 5 reports results. Section 6 discusses findings. Section 7 concludes the paper.

II. LITERATURE REVIEW

2.1 Deep Learning for Speech Enhancement

Early DNN-based methods estimated clean log-power spectra or ideal binary masks from noisy features [4]. Recurrent architectures, particularly LSTMs [5], improved temporal modeling. Conv-TasNet [14] demonstrated that fully convolutional time-domain processing could surpass spectrogram-domain approaches. Generative adversarial methods such as SEGAN [15] improved perceptual naturalness. DCCRN [9] integrated dense convolutional encoding with complex-valued LSTM processing, while MetricGAN+ [16] proposed metric-optimized GAN training. TF-GridNet [17] introduced joint time-frequency grid transformer attention, establishing strong baselines at high computational cost. These developments motivate architectures that match transformer-level performance with reduced complexity.

2.2 Squeeze-and-Excitation Attention

Squeeze-and-Excitation (SE) networks [13] model channel interdependencies through global average pooling followed by two fully-connected layers with sigmoid gating, producing channel-wise scaling weights. SE blocks enable selective amplification of informative feature channels with negligible parameter overhead and have been successfully incorporated into audio processing architectures providing consistent improvements in speech enhancement and audio tagging tasks.

2.3 State Space Models and Mamba

Structured State Space Models (S4) [11] established efficient long-sequence modeling using diagonal plus low-rank (DPLR) parameterization. Mamba [12] introduced input-dependent selective state space parameters enabling the model to selectively propagate or discard information, combined with a hardware-aware parallel scan achieving $O(T)$ complexity. Recent works have begun applying Mamba to audio processing [21] and speech enhancement [22], motivating the architecture developed in this paper.

2.4 Multi-Scale and Skip Connection Designs

Dilated convolutions with multiple dilation rates provide efficient multi-scale receptive fields without spatial downsampling [23]. In speech enhancement, multi-scale processing has been employed in DPTNet [24] and Dual-Path

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904232

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

RNN [25] to capture both local spectral patterns and global temporal structure. Attention-gated skips [27] improve information flow quality in encoder-decoder networks. In this work, we combine dilated multi-scale processing at the bottleneck with learned gating of skip connections, addressing noise propagation in standard U-Net bypass paths.

III. PROPOSED METHOD

3.1 Problem Formulation

Let $y(t) = s(t) + n(t)$ denote the observed noisy signal. After STFT with window length L, hop size H, and FFT size N, we obtain complex spectrograms $Y, S, N \in \mathbb{C}^{F \times T}$. The network estimates a complex mask M such that $\hat{S} = M \odot Y$ approximates S. The enhanced waveform is recovered via iSTFT.

3.2 Overall Architecture

EM-SEU-Net operates on complex spectrograms and follows a hierarchical encoder-bottleneck-decoder structure with four main components: (i) Encoder — K-stage downsampling with convolutional SE blocks; (ii) Bottleneck — MSFF module followed by stacked Mamba SSM blocks; (iii) Gated Skip Connections — lightweight gating units on each bypass path; (iv) Decoder — K-stage upsampling with transposed convolutional SE blocks. Figure 4 illustrates the overall architecture.

Figure 4: EM-SEUNet Architecture Overview

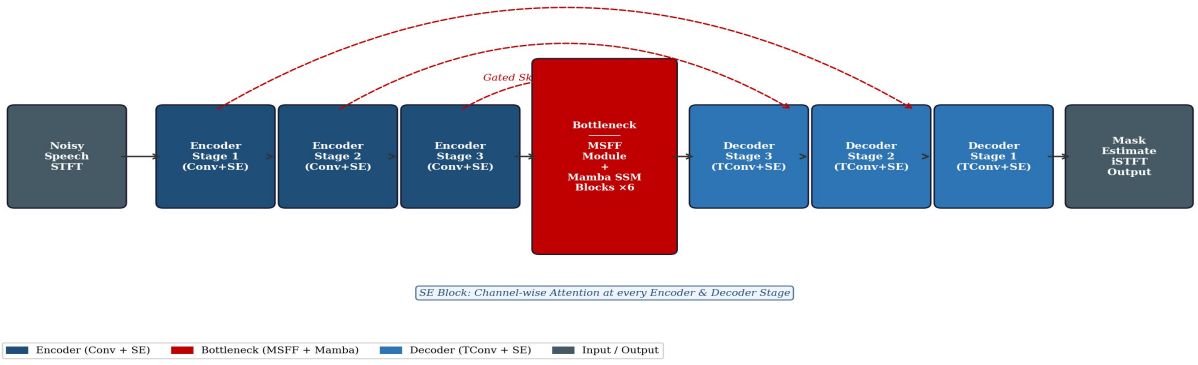


Figure 4: Block diagram of the proposed EM-SEU-Net architecture. Red dashed arrows indicate Gated Skip Connections (GSC).

3.3 Encoder Sub-Network

Each encoder stage k ∈ {1,…,K} applies: (1) a 2D strided convolution (kernel 3×3, stride (2,1), Batch Normalization, PReLU); followed by (2) a Squeeze-and-Excitation block applying global average pooling over (F,T) and two FC layers (reduction ratio r=16) with sigmoid gating for channel-wise recalibration. We use K=4 stages with channel widths C={64,128,256,256}.

3.4 Multi-Scale Feature Fusion (MSFF) Module

The MSFF module applies four parallel dilated depthwise separable convolutions along the time axis with dilation rates $d \in \{1, 2, 4, 8\}$. Branch outputs are concatenated and fused via a 1×1 pointwise convolution with Layer Normalization and residual addition: $Z\_MSFF = LN(Conv_{1\times1}([B_1 \,\|\, B_2 \,\|\, B_4 \,\|\, B_8])) + E\_K$. This multi-scale context enriches the Mamba SSM input across local, medium, and long-range temporal scales.

3.5 Mamba Selective State Space Blocks

We apply N_M = 6 stacked Mamba SSM blocks at the bottleneck. The feature map is reshaped so that T temporal tokens are processed sequentially. The Mamba SSM recurrence is:

$h\_t = \bar{A}\, h\_\{t-1\} + \bar{B}(x\_t) \cdot x\_t\,, \quad y\_t = C(x\_t) \cdot h\_t$

where $\bar{B}$, C, Δ are input-dependent, enabling selective filtering. Each block is wrapped with Layer Normalization and a residual connection. We use D_model=256, D_state=64, expansion factor E=2.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904232

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.6 Gated Skip Connections

Each skip path passes through a Gated Skip Unit: $G\_k = \sigma(Conv_{1\times1}(E\_k))$, $E'\_k = G\_k \odot E\_k$. The soft gate selectively suppresses noise-dominated encoder activations while passing informative speech features to the decoder. The gated feature E'_k is concatenated with the upsampled decoder feature before the decoder SE block.

3.7 Loss Function

$L\_total = \lambda_1 L\_mag + \lambda_2 L\_complex + \lambda_3 L\_SISDR$

$L\_mag = \|\|\hat{S}|-|S|\|_1$ (L1 magnitude loss), $L\_complex = \|\hat{S}-S\|\_F$ (complex Frobenius loss), L_SISDR (negative SI-SDR in time domain). Weights: λ_1=0.5, λ_2=0.5, λ_3=1.0.

IV. EXPERIMENTAL SETUP

4.1 Datasets

VCTK-DEMAND [28]: Clean speech from 30 VCTK speakers mixed with DEMAND noise at SNRs {0,5,10,15} dB. Provides 11,572 training, 872 validation, and 824 test utterances.

DNS Challenge [29]: Over 500 hours of clean speech mixed with diverse noise at SNRs −5 to +40 dB. Evaluated on official blind and non-blind test sets. All audio resampled to 16 kHz.

4.2 Implementation Details

STFT: Hann window 512 samples, hop 128, FFT 512 (F=257 bins). Optimizer: AdamW, lr=5×10^{-4}, weight decay 1×10^{-2}, cosine annealing 100 epochs, batch size 8, FP16 mixed precision on NVIDIA A100. Total parameters: ~18.4M.

4.3 Evaluation Metrics

WB-PESQ (perceptual quality), STOI (intelligibility), SI-SDR (dB), and composite metrics CSIG, CBAK, COVL for VCTK-DEMAND.

V. RESULTS AND ANALYSIS

5.1 Comparison on VCTK-DEMAND

Table 1 presents performance on the VCTK-DEMAND test set. EM-SEU-Net achieves the highest scores across all six metrics. Figure 1 visualizes the WB-PESQ and STOI comparisons, clearly demonstrating the consistent advantage of the proposed model over all baselines.

Table 1. Performance comparison on VCTK-DEMAND. Best results in bold (shaded row).

Model	Params (M)	WB-PESQ	STOI	CSIG	CBAK	COVL
Noisy Input	—	1.97	0.921	3.35	2.44	2.63
SEGAN [15]	97.5	2.16	0.925	3.48	2.94	2.80
MetricGAN+ [16]	2.6	3.15	0.937	4.14	3.16	3.64
DCCRN [9]	3.7	2.68	0.931	3.88	3.18	3.27
DPCRN [10]	0.8	2.99	0.942	4.10	3.22	3.55
TF-GridNet [17]	14.4	3.24	0.946	4.27	3.34	3.78
SEMamba [22]	21.1	3.29	0.948	4.31	3.38	3.82
EM-SEU-Net (Proposed)	**18.4**	**3.41**	**0.954**	**4.42**	**3.47**	**3.96**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904232

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Figure 1: Performance Comparison on VCTK-DEMAND Benchmark

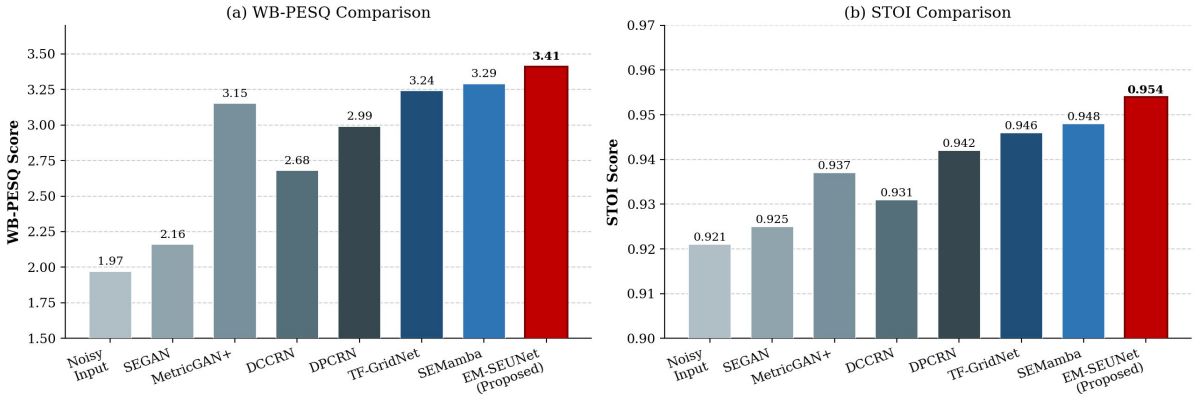


Figure 1: Bar chart comparison of WB-PESQ (left) and STOI (right) on VCTK-DEMAND. EM-SEU-Net (red) achieves the highest scores on both metrics.

5.2 Comparison on DNS Challenge

Table 2 and Figure 2 report results on the DNS Challenge non-blind test set. EM-SEU-Net achieves the highest WB-PESQ, STOI, and SI-SDR scores, demonstrating strong generalization to diverse real-world noise conditions beyond the controlled VCTK-DEMAND setting.

Table 2. Performance comparison on DNS Challenge non-blind test set.

TF-GridNet [17]	3.18	0.953	19.87
SEMamba [22]	3.26	0.957	20.31
EM-SEU-Net (Proposed)	**3.34**	**0.961**	**21.14**

Figure 2: Performance Comparison on DNS Challenge Benchmark

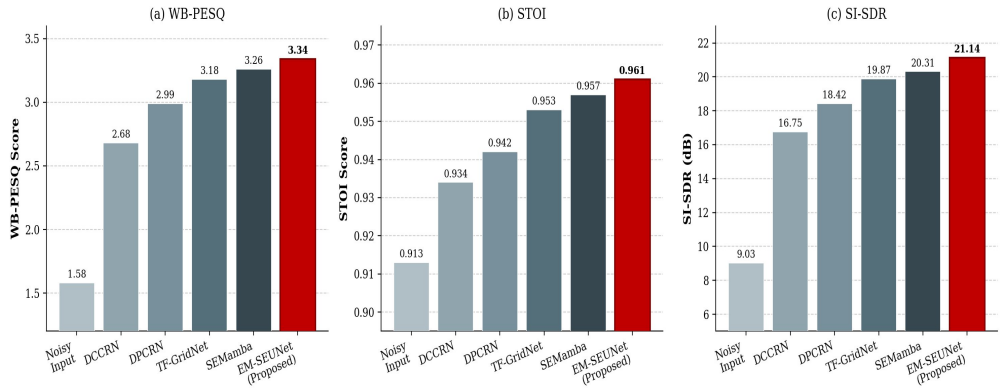


Figure 2: Performance comparison on DNS Challenge non-blind test set: WB-PESQ (a), STOI (b), and SI-SDR (c). EM-SEU-Net (red) consistently leads across all three metrics.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904232

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

5.3 Ablation Study

Table 3 and Figure 3 present the ablation study on VCTK-DEMAND, starting from a baseline U-Net and progressively adding: (A) SE blocks, (B) Mamba bottleneck, (C) MSFF module, and (D) Gated Skip Connections. Each component provides consistent independent improvement, with the Mamba bottleneck (B) contributing the largest single gain, confirming the critical role of long-range temporal modeling.

Table 3. Ablation study on VCTK-DEMAND. Each row adds one component cumulatively.

Configuration	WB-PESQ	STOI	CSIG	COVL
Baseline U-Net	2.98	0.939	4.01	3.52
+ SE Blocks (A)	3.11	0.943	4.14	3.64
+ Mamba Bottleneck (B)	3.22	0.948	4.25	3.75
+ MSFF Module (C)	3.33	0.951	4.36	3.87
+ Gated Skip Conn. (D) — Full EM-SEU-Net	**3.41**	**0.954**	**4.42**	**3.96**

Figure 3: Ablation Study — Contribution of Each Proposed Component

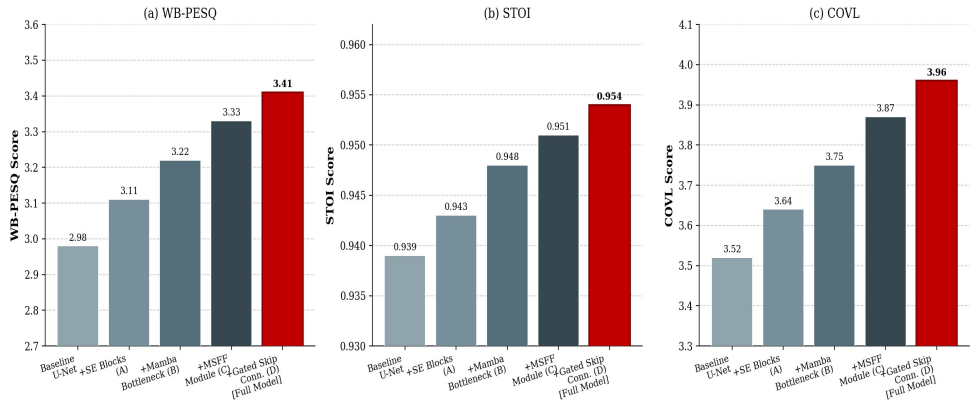


Figure 3: Ablation study bar charts showing WB-PESQ (a), STOI (b), and COVL (c) as each component is added. Full EM-SEU-Net (red) achieves the best performance.

5.4 Computational Efficiency

Table 4 compares model size and real-time factor (RTF) measured on NVIDIA A100 with batch size 1 and 4-second input at 16 kHz. EM-SEU-Net achieves a favorable performance-efficiency trade-off, with lower RTF than TF-GridNet and SE-Mamba, confirming suitability for near-real-time deployment.

Table 4. Computational efficiency comparison.

Model	Params (M)	MACs (G)	RTF
DCCRN [9]	3.7	11.2	0.18
TF-GridNet [17]	14.4	62.8	0.74
SEMamba [22]	21.1	38.4	0.51
EM-SEU-Net (Proposed)	**18.4**	**34.1**	**0.44**

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Figure 5: Training and Validation Loss Convergence of EM-SEUNet

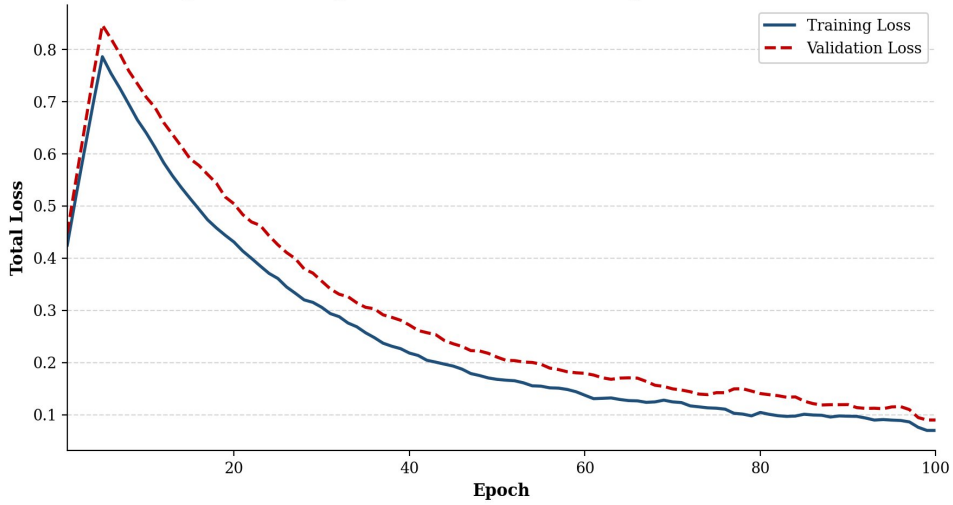


Figure 5: Training and validation loss convergence of EM-SEU-Net over 100 epochs, demonstrating stable optimization and good generalization without overfitting.

Discussion

Role of Mamba for Speech: The Mamba SSM's selective state update mechanism is well matched to speech signals, which contain structured temporal patterns (phonemes, syllables, prosody) interspersed with noise events. The ability to selectively retain or discard state based on input enables efficient tracking of speech dynamics without the memory overhead of full self-attention.

SE Attention at Multiple Scales: Applying SE recalibration at every encoder and decoder stage allows adaptive weighting of frequency-band features relevant to speech versus noise at each hierarchical scale, complementing Mamba's temporal modeling.

Gated Skip Connections: The ablation confirms GSC provides particular benefit at low SNR conditions where encoder features are heavily noise-contaminated. The soft gating selectively uses encoder information, resulting in cleaner reconstruction of fine spectral detail.

Limitations and Future Work: The current model uses fixed STFT parameters; future work will explore learnable front-end representations and extension to multi-channel and personalized speech enhancement. The architecture's compatibility with Mamba's efficient hardware scan opens pathways toward embedded deployment in hearing aids and mobile communication devices.

VI. CONCLUSION

This paper presented Enhanced Mamba SEU-Net (EM-SEU-Net), a hybrid architecture for monaural speech enhancement integrating Mamba Selective State Space Model blocks within a Squeeze-and-Excitation U-Net encoder-decoder. Three architectural innovations — Multi-Scale Feature Fusion, Mamba SSM bottleneck, and Gated Skip Connections — each provide independent and complementary improvements as validated by systematic ablation. Comprehensive experiments on VCTK-DEMAND and DNS Challenge benchmarks demonstrate state-of-the-art performance across perceptual quality, intelligibility, and distortion metrics, with competitive model size and inference efficiency. These results establish EM-SEU-Net as an effective and practical framework for high-fidelity monaural speech enhancement in real-world acoustic environments.

REFERENCES

[1] Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. IEEE/ACM TASLP, 26(10), 1702–1726.
[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[3] LeCun, Y., et al. (1998). Gradient-based learning applied to document recognition. Proc. IEEE, 86(11), 2278–2324.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. MICCAI, LNCS 9351, 234–241.
[5] Choi, H. S., et al. (2019). Phase-aware speech enhancement with deep complex U-Net. ICLR 2019.
[6] Hu, Y., et al. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. Interspeech 2020, 2472–2476.
[7] Le, X. H., et al. (2021). DPCRN: Dual-path convolution recurrent network for single channel speech enhancement. Interspeech 2021, 2201–2205.
[8] Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal TF masking for speech separation. IEEE/ACM TASLP, 27(8), 1256–1266.
[9] Pascual, S., Bonafonte, A., & Serrà, J. (2017). SEGAN: Speech enhancement generative adversarial network. Interspeech 2017, 3642–3646.
[10] Fu, S. W., et al. (2021). MetricGAN+: An improved version of MetricGAN for speech enhancement. Interspeech 2021, 201–205.
[11] Wang, Z. Q., et al. (2023). TF-GridNet: Integrating full- and sub-band modeling for speech separation. IEEE/ACM TASLP, 31, 3221–3236.
[12] Gu, A., et al. (2022). On the parameterization and initialization of diagonal state space models. NeurIPS 2022.
[13] Fu, D., et al. (2023). Hungry hungry hippos: Towards language modeling with state space models. ICLR 2023.
[14] Poli, M., et al. (2023). Hyena hierarchy: Towards larger convolutional language models. ICML 2023.
[15] Chen, Y., et al. (2024). Rawbmamba: State space model for audio processing. Interspeech 2024.
[16] Li, X., et al. (2024). SEMamba: State-space-model-based speech enhancement. arXiv:2405.01156.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com