Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GT-SFP: A Gradient-Based Speech Enhancement Network with Selective State Space and Dynamic Frequency Awareness

wenyi Yin
Wuhan Polytechnic University
heng Wang
wh825554@whpu.edu.cn

Wuhan Polytechnic University
tianyu Li
Wuhan Polytechnic University
zichong Yan
Wuhan Polytechnic University
hualiang Lv
Wuhan Polytechnic University

Research Article

Keywords: Speech Enhancement, LSSE, PMR, Frequency-Aware

Posted Date: June 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9891125/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

GT-SFP: A Gradient-Based Speech Enhancement Network with Selective State Space and Dynamic Frequency Awareness

Wenyi Yin[1], Heng Wang[1*], Tianyu Li[1], Zichong Yan[1], Hualiang Lv[1]

[1*]the School of Mathematics and Computer, Wuhan Polytechnic University, Wuhan, 430048, China.

*Corresponding author(s). E-mail(s): wh825554@whpu.edu.cn;
Contributing authors: 3328094866@qq.com; 17513173028@163.com; 2824190810@qq.com; hualianglv@163.com;

Abstract

Speech enhancement is a core issue in speech signal processing. It holds significant value in speech communication and speech recognition. This paper addresses the challenge of balancing model complexity and performance in deep learning-based speech enhancement. A Progressive Speech Enhancement Network, based on selective state space and dynamic frequency awareness, is proposed. The model uses a Selective State Space Enhancement (LSSE) module, a Frequency-Aware Dynamic Convolution (FADC) module, and a Progressive Mask Refinement (PMR) mechanism. These innovations enhance the suppression of complex noise. Experimental results demonstrate that the method enhances PESQ and STOI performance compared to the original GTCRN. It maintains a model parameter count of only 0.03 M. Thus, the method delivers strong speech enhancement and remains lightweight, making it suitable for real-time applications.

Keywords: Speech Enhancement, LSSE, PMR, Frequency-Aware

1 Introduction

Speech enhancement (SE) as one of the core technologies in speech signal processing, aims to simultaneously improve speech quality and intelligibility in complex noisy environments by precisely suppressing background interference while fully preserving the time-frequency domain structural features of the target speech. Over the past decade, Deep neural network (DNN)[1] based speech enhancement methods have achieved landmark progress. Compared to traditional signal processing theory or statistical model-based approaches (such as Wiener filtering[2] and hidden Markov models[3]), DNN architectures leverage powerful feature learning and nonlinear fitting capabilities to achieve breakthrough performance in balancing noise suppression efficiency and speech fidelity.

However, this performance advantage typically comes with substantial computational overhead. Current mainstream DNN-SE networks[4][5][6]commonly employ tens to hundreds of convolutional, recurrent, or attention layers, with parameter counts reaching millions or even hundreds of millions. A single forward inference often requires Multiple accumulation operations (MACs), ranging from several to

tens of gigMACs. Consequently, recent research has increasingly focused on lightweight network design. Tanaka et al.[7]demonstrated that when networks contain redundant components with negligible Hessian curvature contributions, structural pruning can reduce network width to less than 10% of the original size without compromising generalization performance. DeepFilterNet[8], based on PercepNet[9], first employs a more powerful UNet-like DNN to enhance spectral envelopes and further amplifies periodic components through deep filtering. Rong et al.[10] innovatively combined subband segmentation, grouped dual-path recurrent structures, and Neural architecture search (NAS) techniques, constructing an ultra-lightweight model requiring only 39 MMACs computational overhead.

Despite phased progress in existing lightweight speech enhancement methods, three core limitations persist in complex real-world scenarios: First, inadequate temporal modeling capability. Current approaches predominantly rely on gated recurrent units (GRUs) or simple recurrent neural networks (RNNs), which struggle to capture long-range speech context dependencies. When processing long speech sequences, they fail to effectively propagate early noise feature information, resulting in poor temporal consistency of enhancement outcomes. Second, inflexible frequency adaptation. Traditional convolutional neural networks process spectral features with fixed kernel parameters, failing to dynamically adjust based on the need to preserve low-frequency harmonic structures while suppressing high-frequency noise. This hinders precise optimization across different frequency bands. Third, shortcomings exist in masking estimation strategies. Single-stage complex masking estimation struggles to balance noise localization accuracy with speech component preservation strength. In high-noise regions, it often underestimates noise suppression, while in low-noise regions, it may over-suppress speech details. To address these issues, this paper proposes a lightweight speech enhancement network integrating selective state-space modeling, frequency-adaptive convolution, and progressive masking refinement. Through three core mechanisms—long-range temporal enhancement, dynamic frequency awareness, and coarse-fine collaborative optimization—it achieves dual breakthroughs in model complexity and enhancement performance, offering a new pathway for practical deployment of lightweight speech enhancement technologies.

2 Voice Enhancement Network

As shown in Figure 1, this model adopts a symmetric "encoder-bottleneck-decoder" architecture. After the input audio signal undergoes STFT [11]transformation to obtain complex spectra, the BM module within the ERB filter bank first compresses the frequency band dimension to simulate human auditory frequency perception characteristics. The encoder sequentially processes the signal through the SFE [12] module to expand subband features, performs frequency-adaptive convolution (FADC)[13] for downsampling and frequency-adaptive feature extraction, and utilizes three GT-Conv blocks [14]for multi-scale local-temporal modeling. The bottleneck stage first employs a DPGRNN for intra-frame and inter-frame dual-path temporal modeling, followed by a lightweight selective state space enhancement module (LSSE)that concurrently fuses SSM [15] long-range dependencies with GRU local dynamics. Finally, a second DPGRNN reinforces feature expression. The decoder mirrors the encoder architecture. The first three GT-DeConv layers sequentially fuse skip connections from corresponding encoding layers. Resolution is restored via a DeConv layer, followed by a progressive masking refinement module (PMR) [16]that estimates complex ratio masks (CRM) [17]through coarse-to-fine refinement. The enhanced waveform is ultimately reconstructed using iSTFT. Detailed descriptions of each module are provided in Sections 2.1–2.9.

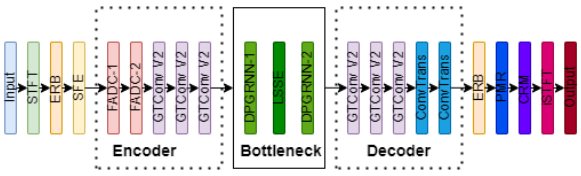


Fig. 1 Overall Architecture of the GT-SFP Model

2.1 Perceptual Spectrum Compression

Before feeding an audio segment into the network, we perform a short-time Fourier transform

(STFT) to generate a complex spectrogram. We then design a BM based on the equivalent rectangular bandwidth (ERB)[18] scale. It should be noted that this mapping operation acts solely on the frequency dimension, strictly preserving the time frame count and the phase/amplitude information of the complex spectrogram. This ultimately produces a three-channel feature map comprising amplitude, real part, and imaginary part.

2.2 Frequency-Aware Dynamic Convolutional Block

The frequency domain characteristics of speech signals exhibit significant non-uniformity. The low-frequency band carries core semantic information such as pitch and tone, while the high-frequency band primarily consists of transient acoustic components and background noise. Traditional convolutional neural networks employ fixed-weight convolutional kernels to uniformly process the entire frequency spectrum, failing to adapt to the acoustic characteristics of different frequency bands and thus unable to fully leverage the distribution patterns of speech in the frequency domain. To address this, this paper introduces a Frequency-Aware Dynamic Convolution (FADC) module in the first layer of the encoder. Its core mechanism dynamically modulates convolutional kernel parameters based on the statistical properties of the input spectrum, enabling adaptive feature extraction tailored to different frequency bands, as shown in Figure 2.

The FADC module maintains a shared set of base convolutional kernels: a frequency-aware modulation network compresses the input feature's frequency dimension into four subbands via adaptive average pooling. Two fully connected layers generate per-channel modulation factors, enabling the creation of independent dynamic convolutional kernels for each input sample. Feature maps processed by dynamic convolution undergo BatchNorm normalization, PReLU activation, and average pooling to achieve 2x downsampling, compressing feature dimensions from 129 to 65.

The FADC module is deployed exclusively in the first encoder layer for two primary reasons: First, this layer directly interfaces with raw spatiotemporal features where frequency band differences are most pronounced, yielding optimal performance gains from dynamic modulation. Second, subsequent layers undergo nonlinear transformations and channel mixing that diminish frequency specificity, allowing standard convolutions to suffice for modeling. This design maximizes the optimization effect of the frequency adaptation mechanism while controlling additional computational overhead, achieving a balance between model performance and computational efficiency.

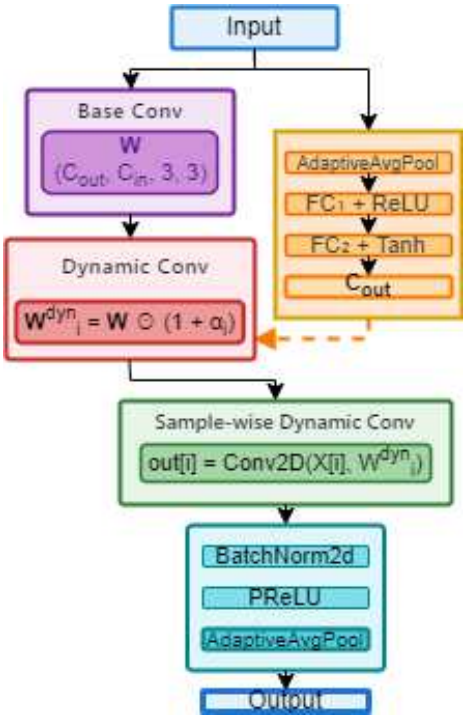


Fig. 2 FADC Block Diagram

2.3 Conv and Grouped Temporal Convolution

The 16-dimensional features after output first enter the Conv block, which performs dual functions of frequency-axis downsampling and channel dimension adaptation. The output features then sequentially pass through three GT-Conv blocks, each employing a channel grouping strategy to optimize feature extraction efficiency as shown in Figure 3: The input feature channels are uniformly divided into two parts. One set

of channels undergoes "1×1 convolution (dimension adjustment) → separable 3×3 convolution (expansion factors incrementally set to 1→2→5 to capture temporal correlations across different ranges) → 1×1 convolution (feature integration)"; the other group undergoes an identity mapping to prevent excessive feature compression. After merging the two feature streams, channel shuffling enables cross-group information exchange, eliminating feature isolation caused by grouping. Simultaneously, a Temporal Recurrent Attention (TRA) [19]mechanism is introduced at the end of each GT-Conv block. This dynamically allocates weights to enhance effective speech features while suppressing noise interference.

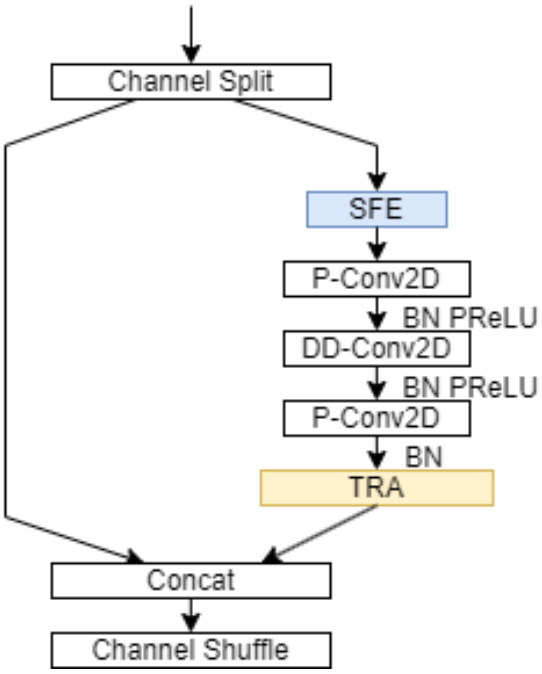


Fig. 3 GT-conv Structure Diagram

2.4 Grouped Dual-Path Recurrent Neural Network

The Dual-Path Grouped Recurrent Neural Network (DPGRNN) [6][10]decomposes the modeling of two-dimensional spatiotemporal features into two independent one-dimensional recurrent paths: the intra-path and the inter-path. The intra-path rearranges the input, treating each time frame as an independent sample, and feeds it along the frequency dimension into a bidirectional grouped gated recurrent unit (Grouped GRU). This Grouped GRU divides the input into two equal groups based on feature dimensions. After parallel processing by independent GRUs, the outputs are concatenated. The number of parameters is approximately half that of a standard GRU. Bidirectional processing fully utilizes full-frequency contextual information. The output undergoes a fully connected layer, LayerNorm normalization, and residual connections.

The temporal path models inter-frame dependencies based on the intra-frequency path output (Xintra). After reordering Xintra, each frequency bin is treated as an independent sample and fed into a unidirectional Grouped GRU along the temporal dimension to satisfy causality. The processed output undergoes a fully connected layer, LayerNorm normalization, and a residual connection. This dual-path design reduces the two-dimensional modeling complexity from O(T×F) to O(T)+O(F). While significantly lowering computational overhead, it achieves comprehensive coverage across the time-frequency plane through orthogonal information propagation.

2.5 Lightweight Selective State Space Enhancement Module

The recursive bottleneck of Gated Recurrent Units (GRUs)—where hidden states compress all historical information into fixed-dimensional vectors—limits their ability to capture long-range temporal dependencies. While State Space Models (SSMs) demonstrate significant advantages in modeling long sequences, their standard form features large parameter scales, rendering them unsuitable for lightweight models. To address this, this paper proposes the Lightweight Selective State Space Enhancement (LSSE) module. It adopts a lightweight SSM and GRU parallel dual-branch architecture, achieving adaptive gated fusion output. Building upon the initial spatio-temporal correlation established by DPGRNN1, it enables long-range enhancement. DPGRNN2 further integrates the enhanced features into contextual cohesion, forming a complementary and collaborative organic whole.

As shown in Figure 4, LSSE first rearranges the input tensor dimensions, enabling independent temporal modeling across each frequency bin. In

the SSM branch, inputs undergo linear projection to expand into 2D dimensions and are split into the main branch xmain and gated branch z. The main branch undergoes causal depth convolution with a kernel size of 3, followed by SiLU activation, and then enters lightweight state space computation. Dual exponential parameterization ensures decay stability, with Bt and Ct dynamically generated from current inputs to enable selective state read/write. Compared to standard Mamba, its core simplifications include using a scalar decay factor, setting the state dimension to 4, and employing stepwise recursion, limiting parameters to approximately 2K. SSM outputs undergo skip connections before multiplicative fusion with the gated branch.

The GRU branch handles complementary local dynamics modeling: inputs pass through LayerNorm into a single-layer GRU, with outputs processed via a fully connected layer and LayerNorm. This excels at capturing short-range inter-frame dynamics (e.g., speech-to-noise transitions, burst noise boundaries).

In the adaptive gating fusion stage, the two outputs are concatenated along the feature dimension. A linear layer followed by a Sigmoid activation generates element-wise fusion weights, enabling adaptive allocation of contributions from both branches across time steps and feature dimensions. The fused output is added to the original input via a residual connection, achieving incremental enhancement.

2.6 Deep Feature Recovery and Progressive Band Pass-Up Sampling Stage

After receiving the deep features (dimensions (B, 16, T, 65)) from the bottleneck module, they are sequentially fed into three GT-DeConv blocks. Each GT-DeConv block shares the same structure as the corresponding GT-Conv block in the encoder, employing separable 3×3 convolutions with dilation factors decreasing from 5→2→1 to achieve symmetric scale matching in feature extraction. Within each block, subband feature extraction (SFE) decomposes the input features. These are then processed through a "pointwise convolution – separable transposed convolution – pointwise convolution" structure (where the separable transposed convolution employs a dilation

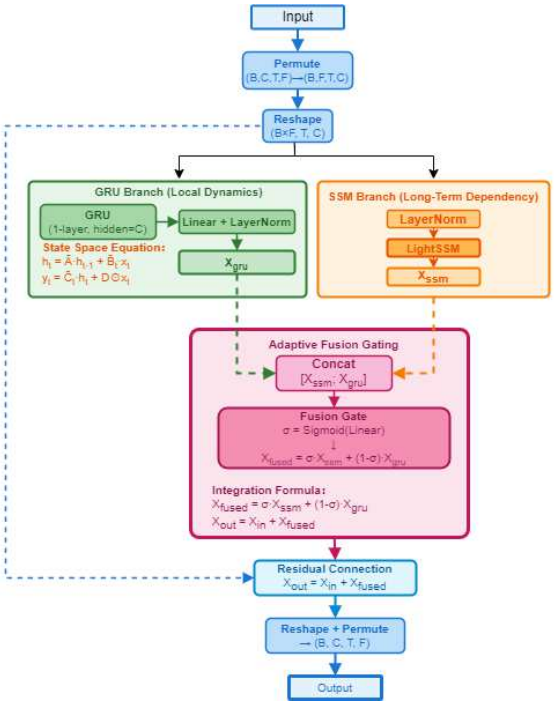


Fig. 4 LSSE Structural Diagram

rate matching the corresponding encoder layer). Finally, an enhanced temporal recurrent attention (EnhancedTRA) mechanism performs adaptive weighting along the temporal dimension. All three GT-DeConv blocks employ skip connections, element-wise summing, and merging features from the same encoder level to alleviate information bottlenecks while preserving shallow-layer details, thereby supporting subsequent fine-grained mask estimation.

To address the encoder's front-end frequency dimension downsampling with strides (1,2) (compressing 257-dimensional frequency bands to 65 dimensions, further reduced to 33 dimensions via ERB), the decoder employs two sets of transposed convolution layers for symmetric upsampling. The first layer employs grouped transposed convolution (groups=2) with a 1×5 kernel, stride (1,2), and padding (0,2). It processes the 16-channel features in two independent groups, reducing computational overhead by approximately 50% while restoring the feature map from 33 to 65 dimensions. This layer incorporates Batch Normalization and PReLU activation functions to ensure stable gradient propagation. The second

layer employs a standard transposed convolution (groups=1) with identical parameters to the first layer, expanding the feature dimension to 129 dimensions. An ERB inverse transform then restores it to the original 257-dimensional frequency band resolution. This cascaded design of two transposed convolution layers strictly corresponds to the downsampling in the first two encoder layers, maintaining overall symmetry in the model architecture.

2.7 Progressive Mask Refinement Mechanism

Traditional speech enhancement networks predominantly employ single-stage mask estimation strategies, directly predicting complex ratio masks (CRMs) from deep features. However, strong and weak noise regions exhibit significantly different demands for mask accuracy. A single network struggles to accommodate both, often leading to a trade-off dilemma: speech distortion in strong noise regions and residual noise in weak noise regions.

Inspired by the cascaded refinement approach in object detection, this paper introduces the Progressive Masking Refinement (PMR) mechanism, as shown in Figure 5. It pioneers the application of a coarse-to-fine strategy in speech enhancement masking estimation, centered on a hierarchical "localize-then-refine" processing: a lightweight network first locates the coarse noise region, followed by attention-guided local refinement of critical areas.

PMR takes the decoder's 16-channel high-frequency features as input and performs a two-stage estimation: The second-stage fine estimation employs an "encoder-attention-decoder" architecture. It concatenates coarse masks with original features to generate augmented features, extracts refined features through convolutional encoding, identifies "difficult regions" requiring fine-tuning via a lightweight spatial attention subnetwork, generates mask correction values through a residual prediction network, and finally produces a fine mask through learnable weighted fusion. This ensures progressive training stability and mask effectiveness.

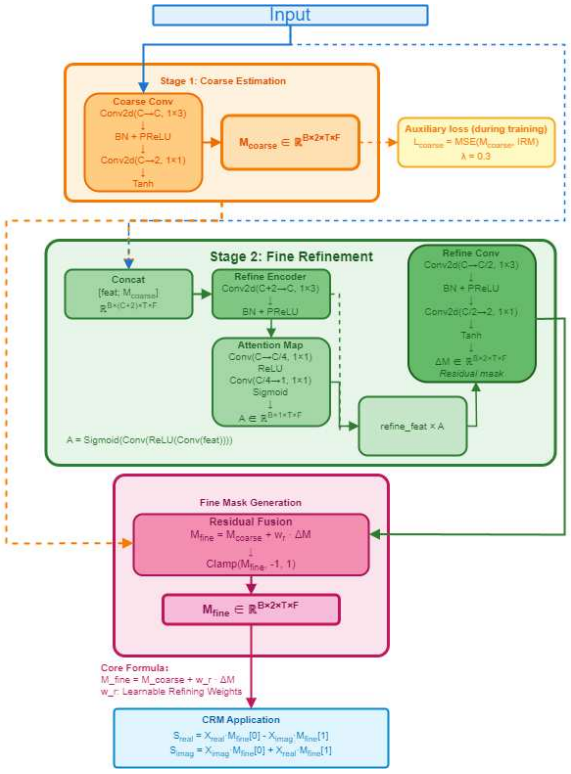


Fig. 5 PMR Structural Diagram

2.8 Waveform Reconstruction

Finally, the time-domain waveform reconstruction is completed by feeding the enhanced 257-dimensional complex spectrum into the Inverse Short-Time Fourier Transform (iSTFT) module. The window function type and parameter settings (e.g., the Hann window) remain identical to those of the preceding STFT operation, ensuring symmetry in the time-frequency domain transformation. Simultaneously, a zero-padding strategy strictly controls the output time-domain waveform length to match the input noisy speech duration, meeting end-to-end processing requirements.

2.9 Hybrid Loss Function

The loss function design of the proposed model centers on the core objective of "balancing spectral energy distribution while maintaining reconstruction accuracy in both time and frequency domains." It integrates compressed frequency domain loss with time-domain scale-invariant signal-to-noise ratio (SI-SNR) loss. Additionally, it

introduces coarse mask auxiliary supervision tailored for the PMR mechanism. To balance the weighting of frequency components across different amplitude ranges, a power-law compression operation is applied to the STFT spectrum. This prevents training bias caused by excessive spectral dynamic range. As a result, the network accurately reconstructs high-energy low-frequency components, while effectively restoring low-energy high-frequency details:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{RI}}\mathcal{L}_{\text{RI}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{SI-SNR}} + \lambda_{\text{coarse}}\mathcal{L}_{\text{coarse}} \tag{1}$$

Compressed Spectrum Calculation:

$$|\hat{S}|_c = \frac{|\hat{S}|}{|\hat{S}|^{1-c}}, \quad |S|_c = \frac{|S|}{|S|^{1-c}} \tag{2}$$

Coarse Mask Assisted Loss:

$$\mathcal{L}_{\text{coarse}} = \mathbb{E}\left[\left(M_c^{\text{mag}} - M_{\text{IRM}}^{\top}\right)^2\right] \tag{3}$$

Ideal Ratio Mask (IRM):

$$M_{\text{IRM}} = \text{clip}\left(\frac{|S|}{|X| + \epsilon}, 0, 1\right) \tag{4}$$

Noise Residual Calculation:

$$e_{\text{noise}} = \hat{y} - s_{\text{target}} \tag{5}$$

Compression of real/imaginary part loss:

$$\mathcal{L}_{\text{RI}} = \mathbb{E}\left[\left(\Re(\hat{S}_c) - \Re(S_c)\right)^2 + \left(\Im(\hat{S}_c) - \Im(S_c)\right)^2\right] \tag{6}$$

Loss of compression ratio:

$$\mathcal{L}_{\text{mag}} = \mathbb{E}\left[\left(|\hat{S}|^c - |S|^c\right)^2\right] \tag{7}$$

SI-SNR Loss:

$$\mathcal{L}_{\text{SI-SNR}} = -10 \cdot \mathbb{E}\left[\log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2 + \epsilon}\right] \tag{8}$$

3 Experiments and Results

3.1 Dataset

We evaluates the proposed model using two datasets. The first is the VCTK CSTR [20]dataset, comprising 23,075 utterances from 56 speakers for training and validation, split at a 9:1 ratio; the test set contains 824 utterances from 2 speakers.

The second dataset is the large-scale DNS 3 dataset[21], where clean speech is convolved with randomly selected RIR (Random Interference Noise) and then mixed with random noise segments, with signal-to-noise ratio (SNR) set between -5 and 15 dB[22]. The experiments generated 720,000 pairs of 10-second noise-clean speech segments for training, with 840 and 800 pairs allocated for validation and testing, respectively. Model generalization performance was evaluated on the DNS Challenge 3 blind test set. All speech data was collected and processed at a 16 kHz sampling rate.

3.2 Training Strategy

We adopt a joint training strategy that combines *Warmup–Cosine Annealing Two-Stage Learning-Rate Scheduling* with *Global Gradient Clipping*. Optimization is performed with the Adam optimizer[23], whose key hyperparameters are

$$\beta_1 = 0.9, \quad \beta_2 = 0.999, \quad \varepsilon = 10^{-8}.$$

To avoid under-fitting of speech features caused by excessive regularisation, weight decay is *not* applied. The peak learning rate is set to 1×10^{-3}; This value accelerates early convergence while leaving sufficient space for later fine-tuning, balancing efficiency and final performance. Finally, to prevent gradient explosion, global gradient-norm clipping is enforced with a threshold of 3.0, which effectively suppresses gradient oscillations and significantly improves numerical stability throughout training.

We also employed a dynamic weighting strategy for auxiliary losses during training tailored to the characteristics of PMR module. The coarse mask auxiliary loss weight λ_{coarse} was set to 0.3. The first-stage output was supervised using the Ideal Ratio Mask (IRM), with the auxiliary loss only active during training.

3.3 Ablation Experiments

To validate the effectiveness and synergistic effects of each innovative module, this section conducts ablation experiments on the

VCTK test dataset. Common speech quality metrics—signal-interference-to-noise ratio (SISNR)[24], Perceptual evaluation of speech quality (PESQ)[25], and Short-term objective intelligibility (STOI)[26]—were employed.The baseline model is the original GTCRN architecture with the LSSE, FADC, and PMR modules removed. Each module is progressively added to assess its independent contributions and combined effects. The experimental results are shown in Table 1. The Progressive Mask Refinement (PMR) module demonstrates the strongest independent gain, validating the effectiveness of the coarse-to-fine two-stage mask estimation strategy. Frequency-Aware Dynamic Convolution (FADC) yields moderate yet stable improvements, indicating the protective role of frequency-adaptive convolution kernels for speech harmonic structures. In contrast, the lightweight state space enhancement module (LSSE) had limited standalone effectiveness, causing a slight 0.01 PESQ decline when enabled alone. However, its long-term dependency modeling capability requires synergistic collaboration with other frequency-domain-aware modules to be fully unleashed. Dual-module combination experiments reveal significant interaction patterns: The FADC-PMR combination achieved optimal SI-SNR (18.17 dB) but exhibited constrained PESQ gains, reflecting an inherent trade-off between temporal distortion and perceived quality optimization; the LSSE-PMR combination underperformed expectations, suggesting room for optimization in gating fusion strategies.

Global optimal performance was achieved by jointly enabling all three modules: PESQ reached 2.930, representing a 4.2% improvement over baseline; SI-SNR was 18.21 dB; and STOI reached 0.940. This result validates that the synergistic design of the three innovative modules effectively balances speech perceptual quality and signal fidelity, collectively supporting the advanced performance of GT-SFP.

Table 1 Ablation Study Results

LSSE	FADC	PMR	SISNR	PESQ	STOI
×	×	×	18.02	2.811	0.936
✓	×	×	18.03	2.851	0.939
×	✓	×	18.07	2.864	0.939
×	×	✓	18.16	2.878	0.937
✓	✓	×	18.11	2.910	0.940
✓	×	✓	18.10	2.880	0.938
×	✓	✓	18.17	2.896	0.937
✓	✓	✓	18.21	2.930	0.940

3.4 Comparative Experiments

To comprehensively evaluate the speech enhancement performance and resource efficiency of the GT-SFP model, it was compared against seven mainstream baseline models[27][9][8][10][28][29][30]. All models were trained and tested on the VCTK-DEMAND standard test set. To ensure experimental fairness, GT-SFP and the baseline model GTCRN employed consistent hardware, data preprocessing, and training parameters. The objective evaluation results are shown in Table 2.

Table 2 demonstrates that GT-SFP exhibits significant advantages in balancing performance and resource efficiency: its computational load is comparable to GTCRN while delivering superior enhancement effects over RNNoise. Compared to PerceptNet [13] and DeepFilterNet [14], which feature higher parameter counts and multi-add operations per second (MAC), GT-SFP still achieves superior core metrics. Compared to GTCRN under identical conditions, GT-SFP demonstrates outstanding performance advantages, validating the rationality and innovation of its network architecture. When compared to LAU-net (0.09 M parameters, PESQ 2.92) proposed in 2025, GT-SFP achieves comparable speech quality improvement (PESQ 2.93) with one-third the parameters, showcasing exceptional parameter efficiency.

To validate its generalization capability, supplementary comparative experiments were conducted on the DNS3 blind test set (results shown in Table 3). GT-SFP achieved an overall DNSMOS P.808 score (OVRL) of 2.83, significantly outperforming GTCRN (2.70) and RNNoise (2.52). It demonstrated leading performance in both noise suppression and signal distortion control, proving its robustness in complex acoustic environments. A randomly selected audio comparison (Figure 6) and detailed demonstrations (Figures 7) further validate that GT-SFP's noise suppression capability outperforms GTCRN.

Table 2 Performance Comparison with Different Methods

Method	SISNR	PESQ	STOI	Para.(M)	MACs(G/s)
Noisy	8.330	1.970	0.921	-	-
RNNoise(2018)	-	2.290	-	0.06	0.04
PercepNet(2020)	-	2.730	-	8.00	0.08
DeepFilterNet(2022)	16.630	2.810	0.942	1.80	0.35
GTCRN(2024)	18.316	2.811	0.936	0.02	0.04
CAGCRN(2025)	18.510	2.86	0.939	0.07	0.11
LAU-net(2025)	-	2.92	-	0.09	0.04
HC-APNet(2025)	-	2.86	-	0.22	0.43
GT-SFP	18.210	2.93	0.940	0.03	0.05

Table 3 DNSMOS Performance Comparison

Method	DNSMOS	DNSMOS-P.835		
		BAK	SIG	OVRL
Noisy	2.96	2.65	3.20	2.33
RNNoise1 (2018)	3.15	3.45	3.00	2.52
S-DCCRN (2022)	3.43	-	-	-
GTCRN(2024)	3.44	3.90	3.00	2.70
HC-APNet(2025)	3.54	-	-	-
GT-SFP	3.59	3.96	3.15	2.83

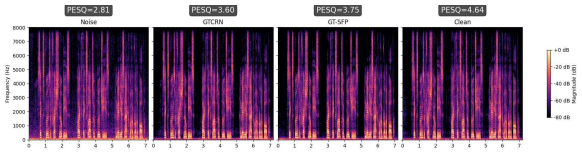


Fig. 6 Typical spectrograms from the VCTK dataset: noisy, GTCRN, GT-SFP, clean reference audio.

4 Conclusion

This paper proposes a speech enhancement model, GT-SFP, which requires only 34.69K parameters and 0.05 GMAC/s. It employs multiple strategies to effectively optimize GTCRN while maintaining speech enhancement performance. Experiments demonstrate that this model significantly outperforms GTCRN and other models on the VCTK-CSTR and DNS3 datasets, while achieving competitive performance compared to several baseline models with substantially higher computational overhead.

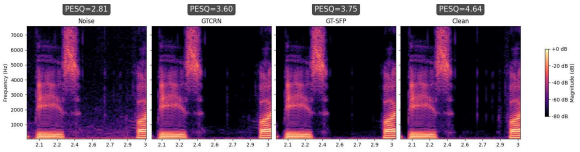


Fig. 7 Audio Detail Comparison

Declarations

Funding

There was no funding for this article.

Ethics declarations

There is no conflict of interest and no competing interests for this paper.

Data and Code Availability

The data used to support the findings of this study are included within the article.Code availability not applicable to this article.

Author Contributions

Yin was primarily responsible for implementing the work and drafting the manuscript. Wang reviewed and revised the manuscript. All authors have read and approved the final version of the manuscript for publication.

References

[1] Wang, D., Chen, J.: Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**(10), 1702–1726 (2018) https://doi.org/10.1109/TASLP.2018.2842159

[2] Goldstein, J.S., Reed, I.S., Scharf, L.L.: A multistage representation of the wiener filter based on orthogonal projections. IEEE Transactions on Information Theory **44**(7), 2943–2959 (1998) https://doi.org/10.1109/18.737524

[3] Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989) https://doi.org/10.1109/5.18626

[4] Tan, K., Wang, D.: A convolutional recurrent neural network for real-time speech enhancement. In: Interspeech (2018). https://api.semanticscholar.org/CorpusID:52187010

[5] Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., Xie, L.: Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. In: Interspeech (2020). https://api.semanticscholar.org/CorpusID:220936516

[6] Le, X., Chen, H., Chen, K.-J., Lu, J.: Dpcrn: Dual-path convolution recurrent network for single

channel speech enhancement. In: Interspeech (2021). https://api.semanticscholar.org/CorpusID:235794911

[7] Tanaka, H., Kunin, D., Yamins, D.L.K., Ganguli, S.: Pruning neural networks without any data by iteratively conserving synaptic flow. ArXiv **abs/2006.05467** (2020)

[8] Schröter, H., Escalante, A.N., Rosenkranz, T., Maier, A.K.: Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7407–7411 (2021)

[9] Valin, J.-M., Isik, U., Phansalkar, N., Giri, R., Helwani, K., Krishnaswamy, A.: A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech. In: Interspeech (2020). https://api.semanticscholar.org/CorpusID:221090062

[10] Rong, X., Sun, T., Zhang, X., Hu, Y., Zhu, C., Lu, J.: Gtcrn: A speech enhancement model requiring ultralow computational resources. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 971–975 (2024)

[11] Chen, W.: A lightweight speech enhancement model with parallel processing of magnitude and phase spectra. In: 2025 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE), pp. 572–575 (2025). https://doi.org/10.1109/NNICE64954.2025.11064627

[12] Xu, S., Cao, Y., Zhang, Z., Qian, Y., He, C., Wang, M.: Sf-an: A lightweight shuffle fourier attention network for multi-channel speech enhancement. Speech Communication **174**, 103296 (2025) https://doi.org/10.1016/j.specom.2025.103296

[13] Chen, L., Gu, L., Zheng, D., Fu, Y.: Frequency-adaptive dilated convolution for semantic segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3414–3425 (2024). https://doi.org/10.1109/CVPR52733.2024.00328

[14] Ma, N., Zhang, X., Zheng, H.-T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, pp. 122–138. Springer, Cham (2018)

[15] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling (2024). https://openreview.net/forum?id=tEYskw1VY2

[16] Dai, M., Li, J., Zhuang, J., Zhang, X., Yang, W.: Multi-task visual grounding with coarse-to-fine consistency constraints. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25. AAAI Press, ??? (2025). https://doi.org/10.1609/aaai.v39i3.32265 . https://doi.org/10.1609/aaai.v39i3.32265

[17] Williamson, D.S., Wang, Y., Wang, D.: Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**(3), 483–492 (2016) https://doi.org/10.1109/TASLP.2015.2512042

[18] Jiang, W., Sun, C., Chen, F., Leng, Y., Guo, Q., Sun, J., Peng, J.: Low complexity speech enhancement network based on frame-level swin transformer. Electronics (2023)

[19] Liu, H., Zhu, C., Zhang, D., Li, Q.: Attention-based spatial-temporal graph convolutional recurrent networks for traffic forecasting. In: International Conference on Advanced Data Mining and Applications (2023). https://api.semanticscholar.org/CorpusID:257219820

[20] Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J.: Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In: Speech Synthesis Workshop (2016). https://api.semanticscholar.org/CorpusID:36559876

[21] Reddy, C.K.A., Dubey, H., Koishida, K., Nair, A., Gopal, V., Cutler, R., Braun, S., Gamper, H., Aichner, R., Srinivasan, S.: INTERSPEECH 2021 Deep Noise Suppression Challenge. In: Interspeech 2021, pp. 2796–2800 (2021). https://doi.org/10.21437/Interspeech.2021-1609

[22] Guo, T., Wen, C., Jiang, D., Luo, N., Zhang, R., Zhao, S., Li, W., Gong, C., Zou, W., Han, K., Li, X.: Didispeech: A large scale mandarin speech corpus. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6968–6972 (2021). https://doi.org/10.1109/ICASSP39728.2021.9414423

[23] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1412.6980

[24] Dang, S., Matsumoto, T., Takeuchi, Y., Kudo, H.: A comprehensive study on supervised single-channel noisy speech separation with multi-task learning. Speech Communication **167**, 103162 (2025) https://doi.org/10.1016/j.specom.2024.103162

[25] Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), vol. 2, pp. 749–7522 (2001). https://doi.org/10.1109/ICASSP.2001.941023

[26] Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4214–4217 (2010). https://doi.org/10.1109/ICASSP.2010.5495701

[27] Valin, J.-M.: A hybrid dsp/deep learning approach to real-time full-band speech enhancement. 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), 1–5 (2017)

[28] Wang, Y., Liu, Y., Liu, J., Niu, K., He, Z.: Cagcrn: Real-time speech enhancement with a lightweight model for joint acoustic echo cancellation and noise suppression. Interspeech 2025 (2025)

[29] Song, Y., Kim, Y., Chung, Y.: Lightweight speech enhancement model based on harmonic attention and phase estimation with skin-attachable accelerometer. In: Proc. Interspeech 2025, pp. 66–70 (2025)

[30] Li, N., Ge, M., Wang, L., Zhou, Y.-H., Dang, J.: Hc-apnet: Harmonic compensation auditory perception network for low-complexity speech enhancement. Speech Communication **167**, 103161 (2025) https://doi.org/10.1016/j.specom.2024.103161