SFG-MAMBA: RELIABILITY-AWARE SPEECH FRAME GRAPH PROPAGATION WITH STATE-SPACE MODELING FOR SPEECH ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-channel speech enhancement (SE) remains challenging when degradation exhibits temporal heterogeneity. Most existing methods process speech as temporally ordered sequences, without frame-level reliability estimation or direct reuse of information from reliable frames across time. To address this problem, we propose SFG-Mamba, a framework that incorporates a reliability-aware Speech Frame Graph (SFG) to model reliability-guided cross-temporal information transfer. A Speech Time-Frequency Graph Convolution (STGC) module further integrates non-local graph information with local time-frequency representations, followed by bidirectional Mamba for long-range temporal modeling. Experiments on VoiceBank+DEMAND and DNS demonstrate the effectiveness of SFG-Mamba, achieving PESQ scores of 3.57 and 3.52, respectively. Furthermore, additional experiments are designed to investigate the performance drop under increasing temporal heterogeneity. SFG-Mamba achieves the best performance among the compared methods.

1 INTRODUCTION

Single-channel speech enhancement (SE) aims to recover clean speech from noisy observations and supports a broad range of applications, including automatic speech recognition, teleconferencing, and hearing assistance Loizou (2013). Although recent neural approaches have substantially improved enhancement quality, their performance remains limited when the corruption pattern within a speech signal is highly heterogeneous. We refer to this condition as non-uniform degradation. Specifically, some speech frames retain relatively intact speech structures, while others are severely masked by transient or intermittent interference This condition is often encountered in real-world recordings and may partly explain the performance drop of existing enhancement methods when moving from simulated noisy speech to real-world recordings. This performance gap highlights a challenge: how to recover heavily corrupted segments without distorting relatively intact ones. Motivated by this challenge, this work focuses on improving speech enhancement performance under non-uniform degradation.

To quantify the temporal non-uniformity degradation, we design Temporal Degradation Dispersion (TDD) as a proxy metric. Since clean speech is unavailable during inference, we adopt MCRA Cohen & Berdugo (2002) to estimate the noise power spectrum directly from the noisy observation. As illustrated in Figure 1(a), we define noise events as timestamps at which the estimated noise energy exceeds a threshold, and compute the intervals (Δt) between consecutive events, including the utterance boundaries. TDD is defined as $\mathrm{std}(\mathbf{d})/(\mathrm{mean}(\mathbf{d}) + \epsilon)$, where a larger TDD indicates a more uneven temporal distribution of degradation. Furthermore, we investigate the performance drop of existing enhancement methods as TDD increases under a fixed overall noise level, as shown in Figure 1(b). The result indicate that temporal degradation heterogeneity introduces an additional challenge beyond noise intensity itself.

Most SE models represent speech as a temporally ordered sequence of frames Xu et al. (2014). Convolutional networks capture local time-frequency patterns Ouyang et al. (2019), recurrent networks model sequential dependencies Weninger et al. (2015), and transformer-based methods capture long-range dependencies through attention-based interactions Wang et al. (2021). However, under non-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

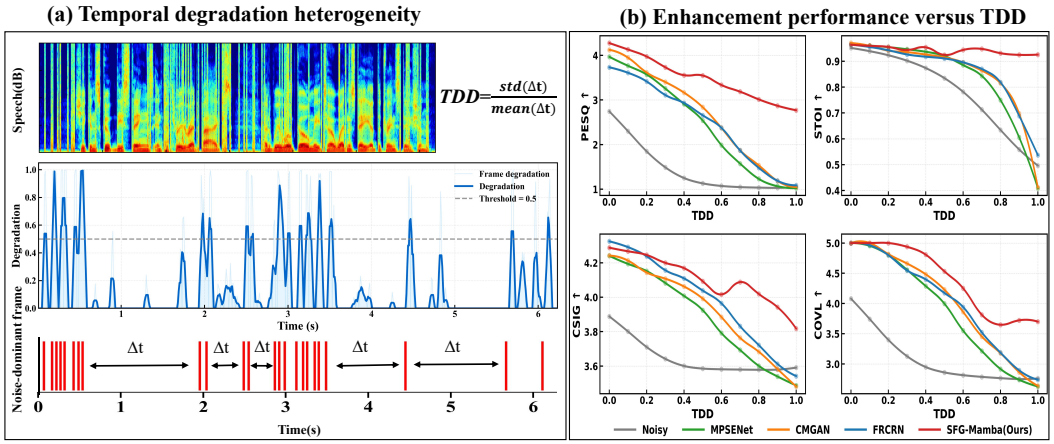


Figure 1: Temporal degradation heterogeneity in speech enhancement. (a) Irregular frame-level degradation. (b) Enhancement performance under different TDD levels.

uniform degradation, reconstruction-relevant acoustic evidence may be unevenly distributed across time. Specifically, relatively clean frames may be concentrated in limited temporal regions. Processing speech as a temporal sequence may weaken the contribution of highly informative frames because their information is aggregated as part of the sequence context and dispersed during contextual integration. Ideally, highly informative frames should serve as information sources that directly support degraded-frame reconstruction. Therefore, a mechanism is needed to directly transfer information from these frames across temporal positions while preserving local time-frequency structure.

Prior work has introduced graph structures into speech processing to model non-local dependencies between frames Saleem et al. (2025); however, these graphs largely rely on fixed topological priors Passos et al. (2023); Wang et al. (2025), which can neither identify trustworthy speech frames nor determine which degraded frames those trustworthy frames should connect to. This suggests that graph connections or plain similarity alone are insufficient, and a reliability-aware directed graph is needed, where edge weights are determined by frame similarity.Accordingly, we use the feature similarity between speech frames as the edge weight reflecting the acoustic compatibility between source and target frames, yielding a weighted directed graph.

To this end, we propose SFG-Mamba, a single-channel speech enhancement framework that incorporates a directed reliability-aware Speech Frame Graph (SFG) to establish cross-frame connections for cross-temporal information transfer. Specifically, speech frames are treated as graph nodes and directed edges are established from source nodes to target nodes. Frames with higher reliability are selected as source nodes, while lower-reliability frames serve as target nodes. Source-to-target connections are further determined by acoustic compatibility, allowing information from reliable sources to be selectively transferred to acoustically matched targets. Note that clean references are unavailable in practical scenarios, so frame-level reliability is estimated directly from the noisy speech using reference-free noise estimation. To preserve local acoustic structure while incorporating cross-temporal information, we further introduce a Speech Time-Frequency Graph Convolution (STGC) module that integrates graph-based features with local time-frequency representations. These features are then unified within a bidirectional Mamba-based architecture, which efficiently models long-range temporal dependencies through selective State-Space modeling for final speech reconstruction. Our main contributions are summarized as follows:

- We propose a reliability-aware speech frame graph (SFG) tailored to non-uniform degradation. It treats speech frames as nodes and builds directed edges from reliable source frames to degraded target frames based on acoustic compatibility, enabling selective cross-time information transfer beyond temporal ordering.
- We propose SFG-Mamba, a unified framework that integrates SFG and speech time-frequency graph convolution (STGC) for cross-frame and local time-frequency modeling, followed by bidirectional Mamba for long-range temporal modeling.

- Extensive experiments on the VoiceBank+DEMAND and DNS datasets demonstrate its effectiveness and robustness, with particular advantages under non-uniform degradation.

2 Related Work

Traditional speech enhancement methods rely on sequential modeling for context, using CNNs and RNNs for local context (e.g., DCCRN Hu et al. (2020) and PercepNet Valin & Skoglund (2019)) and Transformers Subakan et al. (2021) with self-attention Vaswani et al. (2017) for global context. However, these methods organize speech primarily along temporal order, making it difficult to explicitly capture cross-temporal associations between distant but similar frames; under severe noise, global attention may further aggregate noise-dominated frames Zhao & Ma (2023).

Graph Neural Networks (GNNs) model non-local relationships by treating speech structures as nodes and edges, and have been applied to speech emotion recognition Li et al. (2023), speaker verification, and speech enhancement Wang et al. (2025). For instance, Tzirakis et al. Tzirakis et al. (2021) used predefined spatial adjacency, while Chau et al. Chau et al. (2024) adopted feature-distance-based graph construction. However, existing graph SE methods learn or construct acoustic topology Velickovic et al. (2018), but do not explicitly constrain message sources using reference-free reliability under temporally heterogeneous degradation, neither selecting reliable source frames nor determining which degraded frames to associate them with, and thus fail to suppress propagation from noise-dominated frames under non-uniform degradation.

State Space Models Gu & Dao, especially Mamba Jiang et al. (2025), offer linear-complexity long-range modeling via selective scan and have shown promise in audio processing Erol et al. (2024). Yet existing Mamba-based methods still treat speech as sequential representations, lacking explicit modeling of graph-structured acoustic dependencies among distant frames.

3 Method

We consider single-channel speech enhancement under additive noise:

$$y(n) = x(n) + d(n), \tag{1}$$

where $y(n)$, $x(n)$, and $d(n)$ denote noisy speech, clean speech, and noise, respectively. Applying the short-time Fourier transform (STFT) yields

$$Y(t, f) = X(t, f) + D(t, f), \tag{2}$$

where t and f index time frames and frequency bins. SFG-Mamba estimates the clean magnitude and phase spectra from Y to reconstruct the enhanced waveform.

As shown in Figure 2, the noisy waveform $y \in \mathbb{R}^L$ is transformed into magnitude and phase spectra, $Y_m, Y_p \in \mathbb{R}^{T \times F}$, where L, T, and F denote the waveform length, number of frames, and number of frequency bins, respectively. The magnitude spectrum is power-law compressed and concatenated with the phase spectrum to form

$$Y_{in} = \mathrm{Concat}(Y_m^c, Y_p) \in \mathbb{R}^{T \times F \times 2}, \tag{3}$$

where $0 < c < 1$ is the compression exponent and the batch dimension is omitted for clarity.

The input is processed by the Speech Time-Frequency Graph Convolution (STGC) module, which progressively fuses non-local features from Speech Frame Graph (SFG) with local features from Local Time-Frequency (LTF). Within SFG, noise-estimation-based frame reliability guides source selection and weights graph propagation. The fused representation is then processed by M two-stage Mamba modules incorporating bidirectional temporal modeling. Finally, magnitude and phase decoders produce the estimated clean spectra $\hat{Y}_m$ and $\hat{Y}_p$, and the enhanced waveform is:

$$\hat{x} = \mathrm{iSTFT}\left(\hat{Y}_m e^{\mathrm{i}\hat{Y}_p}\right). \tag{4}$$

where $\hat{x} \in \mathbb{R}^L$ denotes the reconstructed enhanced speech waveform.

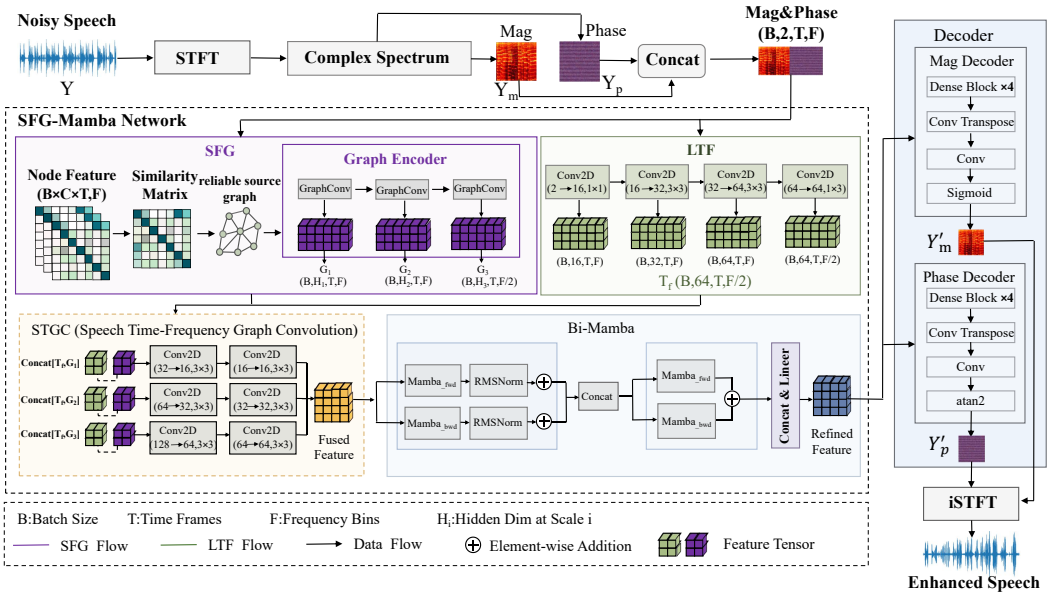


Figure 2: Overview of SFG-Mamba. STGC progressively integrates reliability-aware graph propagation in SFG with local time-frequency modeling in LTF. Bi-Mamba captures bidirectional temporal dependencies, followed by magnitude and phase reconstruction

3.1 STGC Module (Speech Time-Frequency Graph Convolution)

The STGC module comprises two branches and a cross-fusion step. The graph branch constructs a reliability-aware Speech Frame Graph (SFG) to propagate non-local cross-frame information, while the local time-frequency branch (LTF) extracts local time-frequency patterns; the two are progressively fused as described below.

3.1.1 SFG (Speech Frame Graph)

To selectively reuse information across temporal frames, we construct a reliability-aware Speech Frame Graph (SFG). Each frame is represented as a graph node, and relatively reliable frames are selected as propagation sources. For each sample, the magnitude and phase features in Y_{in} are flattened within each frame to obtain $H^{(0)} \in \mathbb{R}^{T\times 2F}$, where $H_t^{(0)}$ denotes the feature vector of frame t. The batch dimension is omitted for clarity. Frame reliability is estimated from local noise statistics. Specifically, minima controlled recursive averaging (MCRA) estimates the noise power spectrum $\hat{\lambda}_d(t,f)$ from the uncompressed noisy power spectrum $P_y(t,f) = |Y(t,f)|^2$. We compute

$$\gamma(t,f) = \frac{P_y(t,f)}{\hat{\lambda}_d(t,f)+\varepsilon}, \qquad \hat{\xi}(t,f) \quad = \max\big(\gamma(t,f)-1,0\big), \tag{5}$$

where $\varepsilon > 0$ ensures numerical stability, γ denotes the a posteriori SNR, and $\hat{\xi}$ is a coarse a priori SNR estimate. The time-frequency reliability proxy and frame-level score are defined as

$$r(t,f) = \frac{\hat{\xi}(t,f)}{1+\hat{\xi}(t,f)}, \qquad q_t \quad = \frac{\sum_{f=1}^{F} P_y(t,f)r(t,f)}{\sum_{f=1}^{F} P_y(t,f)+\varepsilon}. \tag{6}$$

Let $V = \{1,\dots,T\}$ denote the set of all speech-frame indices. The K frames with the highest reliability scores are selected as reliable source nodes, $\mathcal{V}_s$ is reliable source, and $\mathcal{V}_t$ is degraded target:

$$\mathcal{V}_s = \mathrm{TopK}(q,K), \qquad \mathcal{V}_t = \{j \in V \setminus \mathcal{V}_s \mid q_j < \tau_q\}. \tag{7}$$

where τ_q denotes the reliability threshold used to identify degraded frames. A directed bipartite graph is then constructed between the reliable source set $\mathcal{V}_s$ and the degraded target set $\mathcal{V}_t$:

$$G = (\mathcal{V}_s, \mathcal{V}_t, E), \qquad E = \{(i,j) \mid i \in \mathcal{V}_s,\ j \in \mathcal{V}_t\}. \tag{8}$$

Thus, only the selected reliable source frames contribute outgoing messages, while graph-propagated information is delivered only to the identified degraded target frames. For each directed source-target pair $(i, j) \in E$, we measure their acoustic compatibility using the cosine similarity between frame features:

$$s_{ij} = \frac{\left(h_i^{(0)}\right)^\top h_j^{(0)}}{\left\|h_i^{(0)}\right\|_2 \left\|h_j^{(0)}\right\|_2 + \epsilon}, \qquad i \in \mathcal{V}_s,\ j \in \mathcal{V}_t, \tag{9}$$

where s_{ij} measures the acoustic similarity between source frame i and target frame j. Since cosine similarity may take negative values, we map it to a non-negative compatibility weight is $\tilde{s}_{ij} = \frac{1+s_{ij}}{2}$. The resulting $\tilde{s}_{ij}$ is used as the edge weight of the directed graph, so that acoustically compatible source-target pairs contribute more strongly to information propagation. To further adapt the propagation strength to the current source-target representations, we introduce a learnable pairwise gate:

$$g_{ij}^{(l)} = \sigma\left((w_g^{(l)})^\top [h_i^{(l)} \| h_j^{(l)}] + b_g^{(l)}\right), \tag{10}$$

where $\|$ denotes feature concatenation and $\sigma(\cdot)$ is the sigmoid function. The final propagation coefficient jointly considers source reliability, source-target acoustic compatibility, and the learned pairwise gate is $\alpha_{ij}^{(l)}$, and graph propagation at layer l is formulated as $h_j^{(l+1)}$:

$$\alpha_{ij}^{(l)} = q_i\, \tilde{s}_{ij}\, g_{ij}^{(l)}, \qquad h_j^{(l+1)} = \sum_{i \in \mathcal{N}_{src}(j)} \alpha_{ij}^{(l)} W_h^{(l)} h_i^{(l)}, \tag{11}$$

where $\mathcal{N}_{\mathrm{src}}(j) = \mathcal{V}_s$ denotes the source-neighbor set of target frame $j \in \mathcal{V}_t$, and $W_h^{(l)}$ is a learnable projection. The SFG output provides non-local information for subsequent fusion with local time-frequency features.

3.1.2 LTF (LOCAL TIME-FREQUENCY)

The LTF branch extracts local time-frequency patterns from the joint magnitude and phase representation, complementing the non-local information provided by SFG. It uses the input Y_{in} defined above, which contains the power-law-compressed magnitude and phase spectra. The branch comprises N convolutional blocks, each consisting of a two-dimensional convolution, batch normalization, and a ReLU activation:

$$\mathcal{C}_i(Z) = \mathrm{ReLU}\left(\mathrm{BN}(W_i * Z + b_i)\right), \quad i = 1, \ldots, N, \tag{12}$$

where Z denotes the input feature map, W_i and b_i are the convolutional weights and biases, respectively, and $*$ denotes two-dimensional convolution. Stacking these blocks expands the receptive field, allowing LTF to capture local spectral structures and their temporal variations. Within STGC, the convolutional features are progressively fused with the corresponding SFG features, as described below. STGC progressively integrates non-local cross-frame information from SFG with local time-frequency features from LTF. Let $H^{(i)}$ denote the output of the i-th SFG layer. The fusion process is formulated as

$$Z_0 = Y_{in}, \qquad Z_i = \mathcal{C}_i(Z_{i-1}) + \mathcal{A}_i(H^{(i)}), \quad i = 1, \ldots, N, \tag{13}$$

where $\mathcal{C}_i$ is the convolutional block defined in the LTF branch, and $\mathcal{A}_i$ aligns the SFG output with the dimensions of the corresponding convolutional feature map. Addition is performed element-wise. At each stage, graph features are incorporated into the convolutional stream, allowing subsequent blocks to process the fused representation. The final output $Z_N \in \mathbb{R}^{B \times C_{out} \times T \times F}$ is passed to the subsequent Bi-Mamba modules, where B, C_{out}, T, and F denote the batch size, output channel count, number of frames, and number of frequency bins, respectively.

3.2 BI-MAMBA (BI-DIRECTIONAL MAMBA)

To capture both past and future temporal dependencies, we adopt a Bi-Mamba, which processes the joint representation through forward and backward scans and fuses their outputs. Specifically, the outputs of the SFG and LTF branches are first stacked along the channel dimension to form the

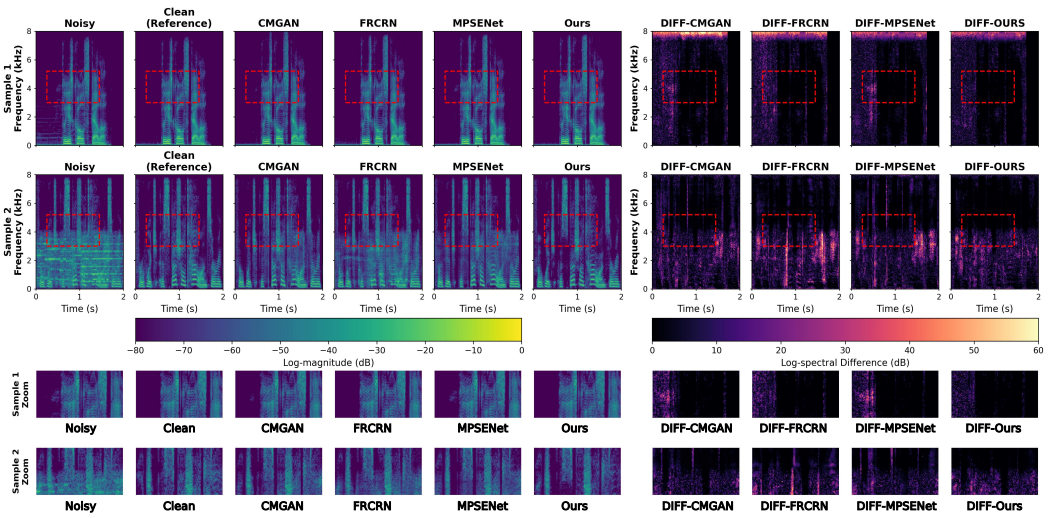


Figure 3: Visual comparison of speech enhancement results. The left part shows the log- magniude spectrograms (dB) of noisy input, clean reference and enhanced outputs from CMGAN, FRCRN, MPSENet, and our proposed method. The right part presents the log-spectral difcrence (dB) between cach enhanced result and the clean reference. The red dashed boxes highlight the regions that are zoomed in and displayed at the bottom. Sample 1 is from the Voicebank+DEMAND dataset, and Sample 2 is from the DNS dataset.

joint representation is $A_{in} \in \mathbb{R}^{C\times N}$, where C and N denote the channel count and the sequence length, respectively. The forward Mamba block models the sequence from time step 1 to N, and the time-reversed sequence is processed by the backward Mamba block:

$$H_{forward} = \mathrm{Mamba}_f(A_{in}), \qquad H_{backward} = \mathrm{Mamba}_b\big(\mathrm{Reverse}(A_{in})\big), \tag{14}$$

where $H_{forward}, H_{backward} \in \mathbb{R}^{C\times N}$. The two hidden representations are then concatenated along the channel dimension and compressed by a linear layer to obtain the final output:

$$H_0 = \mathrm{Linear}\Big(\mathrm{Concat}\big(H_{forward}, H_{backward}\big)\Big), \qquad H_0 \in \mathbb{R}^{C\times N}. \tag{15}$$

To jointly optimize spectral fidelity, phase consistency, perceptual quality, and feature representation, the generator is trained with a multi-objective loss consisting of magnitude loss, phase loss, complex-spectrum loss, and a discriminator-based perceptual loss. Specifically, the magnitude and complex-spectrum losses are computed using mean squared error (MSE), the phase loss combines instantaneous phase, group delay, instantaneous angular frequency losses, the perceptual loss further improves the quality of the generated speech by encouraging higher perceptual-quality scores.

4 Experiments

4.1 Dataset

To evaluate generalization on large-scale data, we used the DNS dataset Reddy et al. (2020) for training and evaluation, which contains over 500 hours of clean English speech from 2,150 speakers, 150 noise types, and 60,000 noisy segments. For comparison with existing studies, we also used the VoiceBank+DEMAND dataset Valentini-Botinhao et al. (2016), including 11,572 training utterances from 28 speakers mixed with 10 noise types and 872 test utterances from 2 unseen speakers mixed with 5 unseen noise types, with SNRs ranging from 0–15 dB for training and 2.5–17.5 dB for testing.

All speech signals were resampled to 16 kHz. STFT features were extracted using a 400-point Hanning window with a 100-point frame shift, corresponding to 25 ms and 6.25 ms, respectively. All models were trained on two RTX 3090 GPUs for 100 epochs with a batch size of 4, using AdamW with an initial learning rate of 0.0005, which was halved every 30 epochs.

Table 1: Results on VoiceBank+DEMAND and DNS. Higher is better; boldface denotes our method.

(a) VoiceBank+DEMAND

Method	PESQ	CSIG	CBAK	COVL	STOI
Noisy	1.97	3.35	2.44	2.63	0.91
SEGAN	2.16	3.48	2.94	2.80	0.92
SE-Conformer	3.13	4.45	3.55	3.82	0.95
FRCRN	3.21	4.23	3.64	3.73	0.96
MetricGAN	2.86	3.99	3.18	3.42	0.95
MetricGAN+	3.15	4.14	3.16	3.64	0.96
DPT-FSNet	3.33	4.58	3.72	4.00	0.96
DB-AIAT	3.31	4.61	3.75	3.96	0.96
CMGAN	3.41	4.63	3.79	4.12	0.96
MPSENet	3.50	4.73	3.88	4.22	0.96
HRLF-Net	3.51	4.23	3.95	4.25	0.96
SFG-Mamba	**3.57**	**4.78**	**4.03**	**4.29**	**0.96**

(b) DNS

Method	PESQ	STOI	SI-SDR
Noisy	1.58	0.91	9.07
FullSubNet	2.78	0.96	17.29
SE-Conformer	3.03	0.95	18.71
FRCRN	3.11	0.96	19.78
Inter-SubNet	2.99	0.97	18.05
VPIDM	3.22	0.97	19.78
MetricGAN+KAN	2.89	0.96	17.80
aTENNuate	2.98	0.95	15.40
CMGAN	3.15	0.96	16.52
MPSENet	3.32	0.98	17.12
HRLF-Net	3.41	0.98	18.35
SFG-Mamba	**3.52**	**0.98**	**19.84**

4.2 Comparison results

We compared the proposed SFG-Mamba with several state-of-the-art speech enhancement models on the VoiceBank+DEMAND and DNS Challenge datasets.

VoiceBank+DEMAND dataset: SEGAN Pascual et al. (2017), SE-Conformer Kim & Seo (2021), FRCRN Zhao et al. (2022), MetricGAN Fu et al. (2019), MetricGAN+ Fu et al. (2021), DPT-FSNet Dang et al. (2022), DB-AIAT Yu et al. (2022), CMGAN Cao et al. (2022), MPSENet Lu et al. (2023), HRLF-Net Jia et al. (2024).

DNS dataset: FullSubNet Hao et al. (2021), FRCRN Zhao et al. (2022), Inter-SubNet Chen et al. (2023), VPIDM Guo et al. (2024), MetricGAN+KAN Mai & Goetze (2025), and aTENNuate Pei et al. (2025).

For VoiceBank+DEMAND, we report PESQ, CSIG, CBAK, COVL, and STOI, while PESQ, STOI, and SI-SDR are used for evaluation on DNS. PESQ measures perceptual speech quality; CSIG, CBAK, and COVL estimate signal quality, background noise quality, and overall quality, respectively; STOI evaluates speech intelligibility; and SI-SDR measures signal reconstruction fidelity.

As shown in Table 1(a), on the VoiceBank+DEMAND dataset, SFG-Mamba outperforms all baseline models across most metrics. Specifically, the PESQ of SFG-Mamba reaches 3.57, surpassing the MPSENet (3.50) by 0.07. The improvements in CSIG, CBAK, and COVL indicate that our model not only reduces noise but also preserves speech quality and background naturalness. Demonstrating that SFG-Mamba framework effectively balances speech enhancement and intelligibility.

On the DNS Challenge dataset, as shown in Table 1(b), SFG-Mamba achieves a PESQ of 3.52, which is higher than all baselines while maintaining competitive STOI and SI-SDR scores. The advantage of SFG-Mamba lies in its explicit modeling of cross-temporal acoustic dependencies via the Speech Frame Graph, combined with phase-aware amplitude reconstruction and bidirectional Mamba for long-range sequence modeling. This allows the model to better recover harmonic structures masked by noise, resulting in improved perceptual quality and waveform fidelity.

In addition to the objective evaluation results, Figure 3 provides a visual comparison of the enhanced spectrograms and the corresponding log-spectral differences. The left part shows the log-magnitude spectrograms of the noisy speech, clean reference, and enhanced outputs produced by different methods, while the right part presents the spectral difference maps with respect to the clean reference. It can be observed that CMGAN and FRCRN still retain noticeable residual noise components, particularly in high-frequency regions, whereas MPSENet tends to over-smooth harmonic structures and suppress fine spectral details. In contrast, the proposed SFG-Mamba preserves clearer harmonic patterns and speech formant structures while effectively removing background noise. Moreover, the spectral difference maps of SFG-Mamba exhibit lower residual energy and fewer distortion artifacts compared with the baseline methods. The zoomed-in regions further demonstrate that SFG-Mamba better reconstructs harmonic continuity and temporal details, leading to enhanced perceptual quality and waveform fidelity across both the VoiceBank+DEMAND and DNS datasets.

Table 2: Results of the ablation studies on VoiceBank+DEMAND dataset

Method	PESQ	CSIG	CBAK	COVL	STOI
SFG-Mamba(Full)	3.57	4.78	4.03	4.29	0.96
SFG-Mamba (-STGC)	2.97	4.01	3.11	3.76	0.91
SFG-Mamba (-SFG+GCN)	3.10	4.18	3.21	3.62	0.92
SFG-Mamba (-BiMamba+Conformer)	3.24	4.35	3.23	3.95	0.94

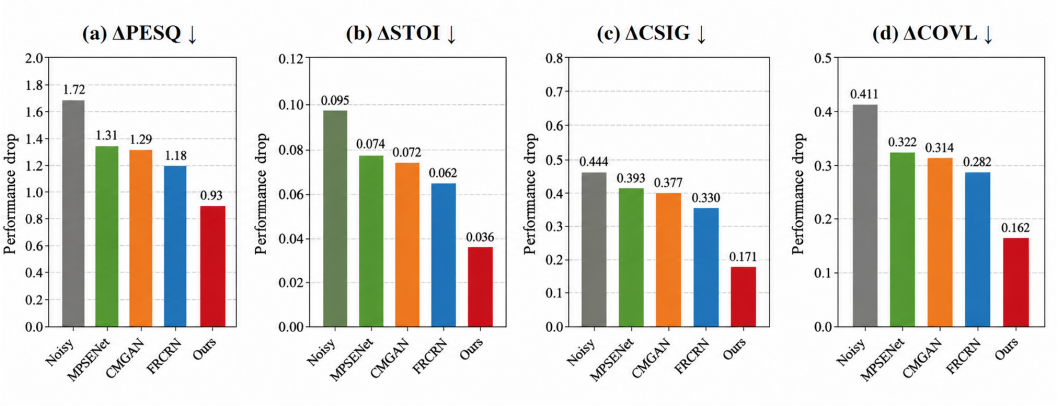


Figure 4: Performance drop induced by non-uniform degradation. Each panel reports ΔPESQ, ΔSTOI, ΔCSIG, and ΔCOVL for SFG-Mamba and the baselines. SFG-Mamba achieves the smallest drop across all metrics, indicating superior robustness to temporal irregularity.

4.3 Impact Analysis of TDD-Controlled Non-uniform Degradation

To isolate the effect of temporal distribution from the overall noise level, we construct a controlled evaluation where the global SNR is held fixed and only the temporal arrangement of noise energy is varied. Conventional evaluations vary the global SNR and thus confound noise strength with temporal irregularity; in contrast, our setup directly exposes how the non-uniformity of degradation, as measured by TDD, affects enhancement performance.

Given a clean utterance, we construct the additive noise as the sum of two components: a burst component formed by $K{=}10$ transient interference events, and a continuous background component sampled from environmental noise. The TDD of each arrangement is controlled by the temporal placement of the burst events: placing them in tighter clusters yields progressively larger TDD, whereas distributing them more uniformly yields smaller TDD. To mimic realistic burst noise, the energy fraction of the background component is tied to the target TDD as $f(\mathrm{TDD}) = 0.10\ \mathrm{TDD}^4$, so that at $\mathrm{TDD} = 0$ the intervals between bursts are essentially clean, while larger TDD injects increasing background energy between clustered bursts. The coupling between burst clustering and background energy is intentional: in real recordings, transient interference rarely occurs in isolation without residual ambient noise, so larger TDD is accompanied by more background energy. The total noise energy is held constant across all arrangements, keeping the global SNR fixed at 0 dB. Each arrangement is characterized by its TDD, computed as described in the introduction ($\mathrm{TDD} = \mathrm{std}(\Delta t)/\,\mathrm{mean}(\Delta t)$ over noise-dominant frames). The observed TDD values span $[0, 1]$, which we divide into 11 equally spaced levels with a balanced number of utterances per level. We evaluate on the VoiceBank+DEMAND test set and report PESQ, STOI, CSIG, and COVL, comparing SFG-Mamba against representative baselines including FRCRN, MPSENet, and CMGAN.

Figure 4 reports the performance drop of each method induced by non-uniform degradation, measured as ΔPESQ, ΔSTOI, ΔCSIG, and ΔCOVL, the decrease of each metric between the lowest and the highest TDD level. A smaller drop indicates greater robustness to temporal irregularity. As expected, all methods suffer from increasing TDD, yet the drop is markedly uneven across methods: the recurrent and attention-based baselines degrade substantially, whereas SFG-Mamba consistently yields the smallest drop across all four metrics. This robustness stems from the reliability-aware speech frame graph, which explicitly selects reliable source frames and propagates their informa-

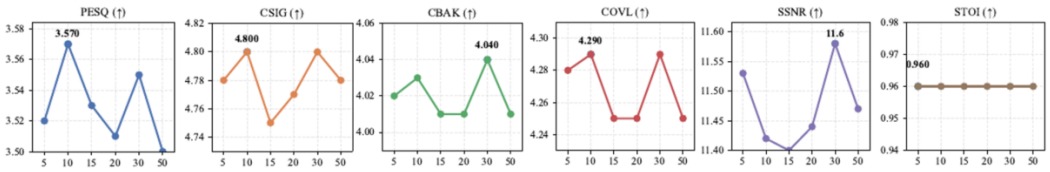


Figure 5: Sensitivity analysis of the Top-K sources.

tion to degraded targets, thereby mitigating the uneven temporal distribution of acoustic evidence that sequential models fail to exploit. These results substantiate our central claim that non-uniform degradation, quantified by TDD, significantly affects speech enhancement performance, and that explicit reliability-aware cross-frame propagation is effective in mitigating its impact.

4.4 Ablation Study

To verify the contribution of each core module, we conduct ablation experiments on the Voice-Bank+DEMAND dataset, as shown in Table 2. The complete model is compared with three variants: SFG-Mamba(-STGC) removes the STGC module to evaluate the contribution of Speech Time-Frequency Graph Convolution to speech enhancement. SFG-Mamba(-SFG+GCN) replaces the proposed SFG module with a vanilla GCN using a fixed similarity graph. Following previous graph-based speech processing methods Tzirakis et al. (2021), each speech frame is regarded as a graph node, and graph edges are constructed according to cosine similarity between node features. SFG-Mamba(-BiMamba+Conformer) replaces the bidirectional Mamba module with a Conformer module while keeping the number of parameters comparable, aiming to evaluate the effectiveness of Bi-Mamba in modeling long-range temporal dependencies.

As shown in Table 2, all ablated variants degrade performance, confirming the contributions of STGC, SFG, and Bi-Mamba. Removing STGC causes the largest drop, replacing SFG with a vanilla GCN also reduces PESQ and STOI, and substituting Bi-Mamba with Conformer further degrades performance, demonstrating the complementary roles of local time-frequency modeling, reliability-aware graph propagation, and long-range sequence modeling.

We further analyze the graph connectivity parameter K, varying $K \in \{5, 10, 15, 20, 30, 50\}$ while keeping all other settings unchanged in Figure 5, the model improves from $K = 5$ to the optimum at $K = 10$, then gradually declines as excessive source connections allow acoustically irrelevant frames to propagate. STOI remains relatively stable. This suggests that moderate graph connectivity provides the best trade-off between information aggregation and robustness.

5 Conclusion

This paper proposes the SFG-Mamba speech enhancement network, which explicitly captures cross-temporal acoustic similarity by constructing a SFG and addresses the feature masking problem under non-uniform degradation by leveraging acoustic prior information from similar frames at a distance. In terms of method design, we developed a STGC module. Through deep interaction between Speech Frame Graph and time-frequency feature modeling, we achieved a unification of global dependencies and local sequence consistency. Additionally, we introduced a Bi-Mamba module to efficiently capture long-range semantic information while maintaining sequence coherence. Experimental results demonstrate that SFG-Mamba performs exceptionally well on mainstream datasets such as VoiceBank+DEMAND, achieving a PESQ score of 3.57, which outperforms existing methods. Further ablation experiments confirm the effectiveness of the synergy between graph structure modeling and State-Space models. More importantly, this paper demonstrates that speech enhancement is not merely a problem of local time-frequency modeling but also requires the full exploitation of the intrinsic cross-temporal similarity structure within the speech signal. This perspective provides a new modeling paradigm for speech enhancement and lays the foundation for future research.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the English prose of the manuscript, and for assistance with LaTeX typesetting. We have not used generative AI tools for formulating the research ideas, designing the method, running the experiments, or generating code; all experimental results, figures, and tables are produced from our own implementations. Additionally, we used generative AI tools for language polishing (grammar, clarity, and consistency of terminology and notation). We have reviewed all AI-assisted work: every AI-edited passage was checked by the authors for technical accuracy and consistency with our experimental results; all equations and notation were verified against our implementation; and all citations were confirmed through a manual literature survey to avoid plagiarism. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ruizhe Cao, Sherif Abdulatif, and Bin Yang. Cmgan: Conformer-based metric gan for speech enhancement. *arXiv preprint arXiv:2203.15149*, 2022.

Hoang Ngoc Chau, Tien Dat Bui, Huu Binh Nguyen, Thanh Thi Hien Duong, and Quoc Cuong Nguyen. A novel approach to multi-channel speech enhancement based on graph neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1133–1144, 2024.

Jun Chen, Wei Rao, Zilin Wang, Jiuxin Lin, Zhiyong Wu, Yannan Wang, Shidong Shang, and Helen Meng. Inter-subnet: Speech enhancement with subband interaction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Israel Cohen and Baruch Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE signal processing letters*, 9(1):12–15, 2002.

Feng Dang, Hangting Chen, and Pengyuan Zhang. Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6857–6861. IEEE, 2022.

Mehmet Hamza Erol, Arda Senocak, Jiu Feng, and Joon Son Chung. Audio mamba: Bidirectional state space model for audio representation learning. *IEEE Signal Processing Letters*, 31:2975–2979, 2024.

Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, pp. 2031–2041. PMLR, 2019.

Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. Metricgan+: An improved version of metricgan for speech enhancement. *arXiv preprint arXiv:2104.03538*, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces.

Zilu Guo, Qing Wang, Jun Du, Jia Pan, Qing-Feng Liu, and Chin-Hui Lee. A variance-preserving interpolation approach for diffusion models with applications to single channel speech enhancement and recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 3025–3038, 2024.

Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6633–6637. IEEE, 2021.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. 2020.

Jing Huang, Liu Yang, and Heng Liu. Bimad: bidirectional mamba-based anomaly detection for multivariate time series. *Signal, Image and Video Processing*, 20(12):655, 2026.

Lizhen Jia, Yanyan Xu, and Dengfeng Ke. A dual-branch speech enhancement model with harmonic repair. *Applied Sciences*, 14(4):1645, 2024.

Xilin Jiang, Cong Han, and Nima Mesgarani. Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Eesung Kim and Hyeji Seo. Se-conformer: Time-domain speech enhancement using conformer. In *Interspeech*, pp. 2736–2740, 2021.

Yan Li, Yapeng Wang, Xu Yang, and Sio-Kei Im. Speech emotion recognition based on graph-lstm neural network. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):40, 2023.

Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2 edition, 2013.

Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra. *arXiv preprint arXiv:2305.13686*, 2023.

Yemin Mai and Stefan Goetze. Metricgan+ kan: Kolmogorov-arnold networks in metric-driven speech enhancement systems. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Zhiheng Ouyang, Hongjiang Yu, Wei-Ping Zhu, and Benoit Champagne. A fully convolutional neural network for complex spectrogram processing in speech enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5756–5760. IEEE, 2019.

Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

Leandro A Passos, João Paulo Papa, Javier Del Ser, Amir Hussain, and Ahsan Adeel. Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement. *Information Fusion*, 90:1–11, 2023.

Yan Ru Pei, Ritik Shrivastava, and FNU Sidharth. Optimized real-time speech enhancement with deep ssms on raw audio. In *Proc. Interspeech*, pp. 51–55, 2025.

Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *arXiv preprint arXiv:2005.13981*, 2020.

Nasir Saleem, Kia Dashtipour, Arif Reza Anwary, Adeel Hussain, Khubaib Ahmed, and Amir Hussain. Temporal-aware graph neural network with conformer model for audiovisual speech enhancement. In *Proc. AVSEC 2025*, pp. 14–18, 2025.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25. IEEE, 2021.

Panagiotis Tzirakis, Anurag Kumar, and Jacob Donley. Multi-channel speech enhancement using graph neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3415–3419. IEEE, 2021.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, pp. 146–152, 2016.

Jean-Marc Valin and Jan Skoglund. A real-time wideband neural vocoder at 1.6 kb/s using lpcnet. *Interspeech 2019*, pp. 3406–3410, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *stat*, 1050:4, 2018.

Kai Wang, Bengbeng He, and Wei-Ping Zhu. Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain. In *ICASSP 2021-2021 IEEE international Conference on acoustics, speech and signal processing (ICASSP)*, pp. 7098–7102. IEEE, 2021.

Tingting Wang, Tianrui Wang, Meng Ge, Qiquan Zhang, and Xi Shao. A learning time-graph frequency representation for monaural speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International conference on latent variable analysis and signal separation*, pp. 91–99. Springer, 2015.

Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM transactions on audio, speech, and language processing*, 23(1):7–19, 2014.

Guochen Yu, Andong Li, Chengshi Zheng, Yinuo Guo, Yutian Wang, and Hui Wang. Dual-branch attention-in-attention transformer for single-channel speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7847–7851. IEEE, 2022.

Shengkui Zhao and Bin Ma. Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Shengkui Zhao, Bin Ma, Karn N Watcharasupat, and Woon-Seng Gan. Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9281–9285. IEEE, 2022.