S^{3}G-Net: Lightweight Banded Network for Real-Time Speech Enhancement

Joyraj Chakraborty, Martin Reed, and Nikolaos Thomos

***Abstract*—Real-time speech enhancement on resource-constrained devices faces a persistent challenge: many causal systems rely on dual-branch magnitude/complex refinements or attention-heavy backbones, whereas ultra-light streaming models often lack explicit cross-band interaction and long-range temporal memory, resulting in unstable artifacts. We propose *Streaming Subband State-space Graph-frequency Network* (S^{3}G-Net), a causal complex-masking architecture built around a *subband state-space graph-frequency bottleneck*. Specifically, S^{3}G-Net (i) allocates frequency-dependent capacity through dynamic subband gating in the encoder with no look-ahead, (ii) restores cross-band structure via banded graph-frequency residual propagation that promotes locally coherent spectral evolution, and (iii) redesigns the temporal core into a lightweight hybrid architecture that integrates a dilated causal Temporal Convolutional Network (TCN) for long-horizon modeling with a *graph-conditioned diagonal state-space module*. The proposed graph-conditioned diagonal state-space model (SSM) uses per frame graph-frequency descriptors to modulate band-wise memory, enabling causal cross-band interaction with small attention while stabilizing streaming behavior. S^{3}G-Net is implemented as a single-stream encoder–bottleneck–decoder with no auxiliary branches or post-filtering, and uses a model with only $\sim$30k parameters requiring $\sim$60 M MAC/s. S^{3}G-Net outperforms or matches substantially larger baselines in terms of perceptual quality across in-domain and out-of-domain evaluations, while maintaining competitive intelligibility at significantly lower cost.**

***Index Terms*—Causal monaural speech enhancement, Graph-conditioned state-space model, Lightweight network, DNN.**

I. Introduction

Real-time speech enhancement (SE) on resource-constrained platforms must simultaneously satisfy three coupled requirements: high perceptual quality, streaming operation with low algorithmic latency, and tight limits on memory and MACs [1]–[3]. Modern monaural SE methods can be broadly viewed as discriminative mask/mapping approaches, generative models, and hybrid or metric-aware systems. Since classical single-channel SE methods often degrade in non-stationary noise, data-driven deep neural network (DNN) approaches that learn time–frequency (T–F) filtering have become standard [4]–[6]. In particular, complex-spectrum modeling has proven important for perceptual quality, from causal complex spectral mapping [7] to complex-valued convolutional architectures [8]. Although diffusion, metric-aware, and recent Mamba/attention-style discriminative SE models can improve enhancement quality [9]–[13], their cost, causality assumptions, or heavier designs remain less suited to strict ultra-light streaming.

Recent efficient architectures, e.g., DeepFilterNet2 [14], FSPEN [15], LiSenNet [16] demonstrate that real-time causal SE is feasible on embedded hardware. However, high quality in ultra-small footprints remains challenging: subband splitting or aggressive full-band simplification can weaken cross-band coordination [15], [16], while short-context temporal modeling may cause musical noise and streaming instability under rapidly varying interference [14]. Compression (pruning/quantization or reduced precision) reduces cost [17], [18], but often trades numerical fidelity and capacity for efficiency.

In this paper, we propose S^{3}G-Net, a strictly causal complex masking model that closes the quality efficiency gap without post-filtering. The core contribution is a subband state-space graph–frequency bottleneck that combines (i) dynamic subband gating for frequency-dependent capacity allocation, (ii) a lightweight GraphFreq residual for band-local coupling and streaming stability, and (iii) long-context temporal modeling via a graph-conditioned diagonal SSM (S4D/Mamba family) paired with a compact dilated causal Temporal Convolutional Network (TCN). On VoiceBank+DEMAND [19] and the DNS [20] evaluation sets, S^{3}G-Net matches or exceeds larger real-time SE baselines while using only $\sim$30k parameters and requiring $\sim$60 M multiply-accumulations (MAC)/s.

Our contributions are:

- we propose a causal, single-stream complex-masking architecture designed for ultra-low computational cost;
- we introduce a novel subband state-space graph-frequency bottleneck combining dynamic subband gating, GraphFreq coupling, and graph-conditioned diagonal state-space temporal modeling;
- we extensively validate strong quality–efficiency trade-offs on standard benchmarks and conduct detailed ablations and complexity analysis.

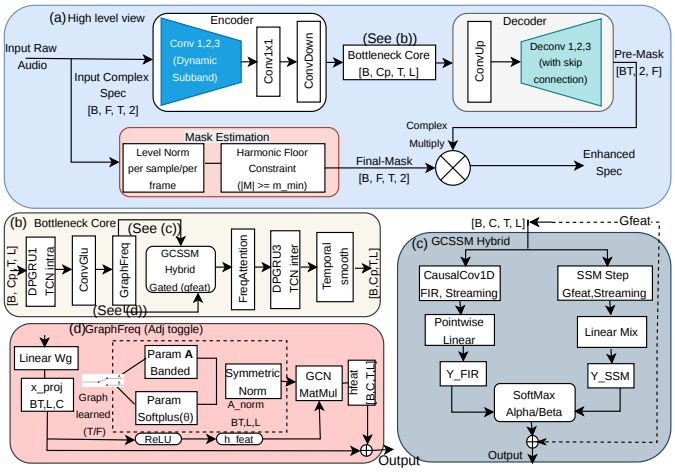
II. Proposed Methodology

In real-time monaural speech communication, the observed signal is $y[n] = s[n] + v[n]$, where $s[n]$ denotes clean speech and $v[n]$ denotes non-stationary noise. The goal is to estimate $s[n]$ causally while suppressing background/transient interference and preserving intelligibility, timbre, and onsets. We operate in the left-aligned Short-Time Fourier Transform (STFT) domain and enhance both magnitude and phase using a complex ratio mask (CRM). Given the noisy spectrum $X = \mathrm{STFT}\{y[n]\}$, with $X \in \mathbb{R}^{B\times F\times T\times 2}$, and a predicted mask $M \in \mathbb{R}^{B\times F\times T\times 2}$, the enhanced spectrum is

$$\hat{S} = M \odot X, \tag{1}$$

where $\odot$ denotes element-wise complex multiplication. The waveform is reconstructed by inverse STFT (iSTFT) of $\hat{S}$. All operators are causal; the encoder and decoder process

J. Chakraborty is with the Department of Engineering Science, University of Oxford, UK. M. Reed and N. Thomos are with the School of Computer Science and Electronic Engineering, University of Essex, UK.



Proposed S^{3}G-Net spectral enhancement architecture.

each STFT frame only along the frequency axis and therefore introduce no additional algorithmic delay.

A. Network architecture

Fig. II summarizes the proposed S^{3}G-Net and its hyperparameters. We propose a causal complex-masking SE model for low-power streaming that remains stable in non-stationary noise, addressing common failure modes of ultra-light designs, i.e., frequency misallocation, temporal jitter (musical noise), and weak cross-band consistency [14], [16], [21], [22]. The model is a single-stream encoder–bottleneck–decoder with frequency-axis skip connections. The band-aware encoder with dynamic subband gating compresses $F \to L$, the lightweight bottleneck couples causal dual-path temporal modeling (dilated depth-wise TCN + diagonal SSM) with a band-local GraphFreq residual, and the decoder restores frequency resolution to predict the CRM—all without auxiliary branches, post-filtering, or look-ahead. We use $(B{\cdot}T, C, F)$ for frequency-only blocks and (B, C, T, L) for temporal blocks, with $F{=}257$ and $L{=}32$.

B. Frequency encoder and dual-path causal modeling

Ultra-light causal encoders may misallocate frequency capacity: fixed receptive fields may be too narrow to capture low-band structure while remaining too coarse for rapidly varying high-frequency content. [15], [16]. We, therefore, use a *frequency-only* two-kernel mixture with a *frame-wise, channel-wise* gate that selects the effective receptive field under strict causality (Fig. II(a)). For each frame t and $x_t \in \mathbb{R}^{B\times C_{\text{in}}\times F}$,

$$\mathbf{g}_t = \sigma\big(W_g\,\mathrm{GAP}_F(x_t)\big), \qquad \mathbf{g}_t \in \mathbb{R}^{B\times C_{\text{out}}\times 1}, \tag{2a}$$

$$y_t = \mathbf{g}_t \odot \mathrm{conv_f}(x_t) + (1-\mathbf{g}_t)\odot \mathrm{conv_c}(x_t), \tag{2b}$$

where $\mathrm{conv_f}$ and $\mathrm{conv_c}$ are 1-D convolutions along F with kernels $k_f{=}3$ and $k_c \in \{6, 8\}$. The gate is broadcast over F, enabling causal frame-/channel-wise selection. A $1{\times}1$ projection maps $C_{\text{out}}{=}\,C\,{=}32$ to $C_p{=}16$, yielding band-compressed features $Y \in \mathbb{R}^{B\times C_p\times T\times L}$.

Given Y (with $C_p{=}C/2$), we apply a dual-path causal core (Fig. II(b)) that separates cross-band binding over L from temporal modeling over T. The intra-path runs a unidirectional Gated Recurrent Unit (GRU) across the L bands per frame, while the inter-path uses an 8-layer causal depth-wise dilated TCN along time (kernel 3, doubling dilations) with $1{\times}1$ mixing and residual skips. Depth-wise separability keeps complexity linear in C_p, and dilation expands receptive field without look-ahead. The resulting features are refined by GraphFreq (Sec. II-C) and the GC–SSM (Sec. II-D) for band-consistent, streaming-stable estimates.

C. Graph-frequency residual

The dynamic subband encoder operates in a frequency-local manner, which in ultra-light streaming models can lead to overly independent band estimates and narrowband artifacts under non-stationary noise. To inject a minimal cross-band prior at low cost, we apply a single-step per-frame graph residual propagation over the compressed band axis L. Fig. II(c) summarizes the fixed versus learned adjacency variants used in our ablations. For each frame t, let $X_t \in \mathbb{R}^{B\times L\times C_p}$ denote bottleneck features (with $C_p{=}C/2$) on which we apply a small channel projection $W_{\text{proj}} \in \mathbb{R}^{C_p\times C_p}$. We then initialize Θ as a $(2w{+}1)$-diagonal band (we use $w{=}3$), which bounds parameters and MACs to $O(LC_pw)$ per frame and biases the coupling to be band-local. Then a nonnegative adjacency $A = \mathrm{softplus}(\Theta)$ is defined through symmetric normalization, $A_n = D^{-1/2}AD^{-1/2}$ with $D = \mathrm{diag}(A\mathbf{1})$. The propagation is implemented as a residual update:

$$\tilde{X}_t \;=\; \phi\big(A_n\,X_t W_{\text{proj}}\big), \qquad X_t^{+} \;=\; X_t + \tilde{X}_t, \tag{3}$$

where $\phi(\cdot)$ is a pointwise nonlinear function (e.g., ReLU). The residual form preserves the original representation while encouraging cross-band consistency. To condition the temporal memory block (Sec. II-D), we derive a per-frame band descriptor from the propagation update by pooling over channels:

$$g_t^{\text{graph}} \;=\; \mathrm{GAP}_{C_p}(\tilde{X}_t) \;\in\; \mathbb{R}^{B\times L}. \tag{4}$$

Stacking g_t^{graph} yields $g \in \mathbb{R}^{B\times T\times L}$, enabling frame-wise band conditioning in the GC–SSM while preserving causality.

D. Graph-Conditioned Diagonal SSM

The dual-path core provides causal context, while the GraphFreq residual yields a per-frame band descriptor g_t that summarizes cross-band diffusion. The remaining challenge in streaming SE is adaptive temporal smoothing, i.e., suppressing non-stationary interference without smearing onsets or inducing musical noise. Classical causal backbones [8], [14], [16], [20], [21] improve temporal modeling but incur increased computational and memory costs. Diagonal SSMs offer an efficient alternative [23], but such lightweight SE typically rely on fixed per-channel dynamics limiting band-wise adaptivity.

We introduce a *graph-conditioned diagonal SSM* that keeps the linear-time diagonal scan while making its dynamics *frame- and band-adaptive* using g_t. The hybrid time core finite impulse response (FIR) branch + conditioned diagonal scan + lightweight mixing) is shown in Fig. II(d). Let $x_t \in \mathbb{R}^{B\times C_p\times L}$ denote the input at frame t and $s_t \in \mathbb{R}^{B\times C_p\times L}$ the diagonal state. The scan is

$$\begin{aligned} s_{t+1} &= \tilde{a}_t \odot s_t + \tilde{b}_t \odot x_t, \\ y_t &= \tilde{c}_t \odot s_t + \tilde{d}_t \odot x_t, \end{aligned} \tag{5}$$

where $y_t \in \mathbb{R}^{B\times C_p\times L}$ and $\odot$ is element-wise. The base coefficients $(a,b,c,d)\in\mathbb{R}^{C_p}$ introduce only $O(C_p)$ parameters; stability is enforced by $a=\tanh(a_{un})$, $a_{un}\in\mathbb{R}^{C_p}$. To adapt dynamics across time and bands, we gate these coefficients with the descriptor $g_t\in\mathbb{R}^{B\times L}$ (broadcast over C_p):

$$\begin{aligned}\tilde{a}_t &= a\odot\sigma(\alpha g_t^{\text{graph}}+b_a), \quad \tilde{b}_t = b\odot\sigma(\beta g_t^{\text{graph}}+b_b),\\ \tilde{c}_t &= c\odot\sigma(\gamma g_t^{\text{graph}}+b_c), \quad \tilde{d}_t = d\odot\sigma(\delta g_t^{\text{graph}}+b_d),\end{aligned} \tag{6}$$

with trainable scales $\alpha,\beta,\gamma,\delta$ and biases b_a,b_b,b_c,b_d broadcast over (B,L). This yields a stable diagonal scan whose effective time constants vary with the graph-conditioned evidence. Hence, bands/frames that require longer retention can increase $|\tilde{a}_t|$, while transient-dominated regions can shorten memory to avoid smearing. The scan costs $O(TLC_p)$ with $O(LC_p)$ state. A small $C_p\times C_p$ mixer after the scan (plus the parallel FIR branch) restores cross-channel expressivity at negligible cost for small C_p, and the resulting features are refined by lightweight frequency attention and causal smoothing (Sec. II-E) before CRM prediction.

E. Attention, temporal smoothing, and decoder

After the dual-path core, we apply a residual ConvGLU before the GraphFreq/GC–SSM blocks to add low-cost nonlinearity. Then, a lightweight per-frame frequency attention (4 heads) refines cross-band mixing, followed by a causal depth-wise temporal convolution (kernel 5) to smooth frame-to-frame variations and reduce mask jitter. The decoder upsamples along frequency using three transposed-convolution stages with skip connections where each stage uses a 1×1 fusion before upsampling. A high-frequency-aware LevelNorm stabilizes the predicted complex ratio mask. Finally, an optional harmonic-floor constraint is applied to the mask magnitude and the resulting mask $\mathbf{M}$ is used to produce the enhanced spectrum.

F. Loss Function

Following [15], [16], [20], we employ a six-term training objective addressing complementary aspects of causal enhancement. Multi-resolution STFT loss enforces signal reconstruction fidelity; differentiable PESQ and STOI surrogates promote perceptual quality and intelligibility; a frozen WavLM feature loss ensures representation consistency; a CBAK penalty suppresses background artifacts; and total variation (TV) regularization on the mask magnitude encourages smoothness [20], [24], [25]. The overall loss is

$$\mathcal{L}=\sum_{i=1}^{6}\lambda_i\,\mathcal{L}_i, \qquad \mathcal{L}_i\in\{\mathcal{L}_m\}_{m\in\ \{\text{STFT,PESQ,STOI,}\atop\text{WavLM,CBAK,TV}\}} \tag{7}$$

with $(\lambda_1,\ldots,\lambda_6) = (0.88,\,0.03,\,0.10,\,0.02,\,0.08,\,0.04)$. $\mathcal{L}_{\text{STFT}}$ averages power-compressed magnitude and complex reconstruction losses over $N\in\{2^7,2^8,2^9,2^{10}\}$ with hop $N/4$ (p=0.30). $\mathcal{L}_{\text{CBAK}}$ suppresses residual energy in non-speech regions via a clean-STFT soft Voice Activity Detection (VAD), and TV regularization penalizes abrupt time–frequency changes in the estimated mask magnitude, $|M|\approx|\hat{S}|/(|X|+\epsilon)$, avoiding instability near spectral zeros:

$$\mathcal{L}_{\text{TV}}=\tfrac{1}{BFT}\big(\,\|\Delta_t|M|\|_1+\tfrac{1}{2}\,\|\Delta_f|M|\|_1\big)\,. \tag{8}$$

TABLE I
OBJECTIVE RESULTS ON VOICEBANK+DEMAND (VB+DEMAND) TEST SET. "–" DENOTES NOT REPORTED.

Model	Params [M]	MACs [G]	Causal	PESQ	CSIG	CBAK	COVL	STOI
Noisy	–	–	–	1.97	3.34	2.44	2.63	0.921
xLSTM-SENet [28]	2.20	40.36†	No	3.48	4.74	3.93	4.22	0.960
MBTU-SE [29]	0.40	–	No	3.39	4.65	3.83	4.12	0.95
Mamba-SEUNet [23]	6.21	9.09	No	3.59	4.80	4.02	4.32	0.960
RNNoise [30]	0.06	0.04	Yes	2.33	3.40	2.51	2.84	0.922
DCCRN [8]	3.70	14.36	Yes	2.79	3.74	3.13	2.75	0.938
TSRM [31]	2.63	–	Yes	2.98	4.36	3.58	3.70	–
FRCRN [21]	10.27	12.30	Yes	3.21	4.23	3.64	3.73	0.942
DeepFilterNet2 [14]	1.78	0.35	Yes	3.08	4.30	3.40	3.70	0.943
CCFNet+ [6]	0.62	1.47	Yes	3.03	4.27	3.55	3.61	0.950
FSPEN [15]	0.096	0.09	Yes	2.97	–	–	–	0.942
LiSenNet [16]	0.037	0.06	Yes	3.07	–	–	–	0.939
Proposed S^{3}G-Net	**0.030**	**0.063**	**Yes**	**3.607**	**4.36**	**3.29**	**3.92**	**0.938**

† MACs are converted from the reported 80.71G FLOPs in MambAttention [27] using 1 MAC ≈ 2 FLOPs; the value is indicative due to differing FLOP/MAC conventions.

TABLE II
EVALUATION OF MODELS TRAINED ON VOICEBANK+DEMAND. WE REPORT PESQ/ESTOI ON THE UNSEEN VOICEBANK+DEMAND-EX TEST SET AND DNSMOS OVRL ON REAL RECORDINGS FROM DNS2020.

Model	VoiceBank+DEMAND-Ex		DNS2020 (Real Rec.)
	PESQ ↑	eSTOI ↑	DNSMOS OVRL ↑
Noisy	1.63	0.63	2.26
DeepFilterNet2 [14]	2.41	0.71	2.54
FSPEN [15]	2.35	0.68	2.42
Proposed S^{3}G-Net	**2.47**	**0.69**	**2.53**

III. EXPERIMENTAL RESULTS

We evaluate S^{3}G-Net on VoiceBank+DEMAND [19], using the standard 16 kHz split without speaker overlap. Training/validation uses 4 s segments with batch size 8. Enhancement is performed in the left-aligned causal complex-STFT domain with Hann windows, N_{fft}=512, and hop = 256 samples, giving 257 frequency bins. During training, a shared random complex rotation $R(\phi)$, $\phi\sim\mathcal{U}[-\pi,\pi]$, is applied to both X and M to reduce global phase bias, and perceptual losses use iSTFT waveforms reconstructed with the same window/hop. We further evaluate generalization on VB-DemandEx [26], [27], introduced in MambAttention [27], as an extended VoiceBank+DEMAND-style benchmark, and on DNS Challenge 2020 [20]. Full-utterance results use WB-PESQ/STOI; models are trained with Adam for 220 epochs, 5×10^{-4} learning rate, and StepLR decay 0.98 every 4 epochs. Code: https://github.com/drchk/S3G-Net-SE.

A. System evaluation

We compare S^{3}G-Net to representative non-causal [23], [28], [29] and causal systems [6], [8], [14]–[16], [30], [31] on VB+DEMAND (Table I). S^{3}G-Net attains the highest overall PESQ (3.607) with only 30 k parameters and 0.063 G MACs, and it also achieves the highest COVL (3.92) among causal systems with competitive STOI (0.938), while using ≈200× and ∼6× less compute over FRCRN and DeepFilterNet2. Although several baseline causal models still depend on a few future frames [16], S^{3}G-Net avoids any form of look-ahead. While several baselines also employ MR-STFT objectives, our gains are consistent across complementary metrics (COVL/STOI) and are achieved with a single-stream causal architecture without

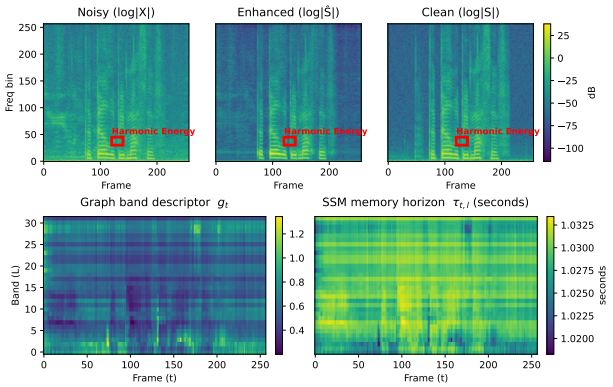


Fig. 1. Enhanced Spectrogram with Graph-Conditioned τ Map.

TABLE III
EXPERIMENTAL RESULTS ON THE DNS CHALLENGE TEST SET.

Method	WB-PESQ	STOI
Noisy	1.58	0.915
FRCRN [21]	3.23	0.976
CCFNet+ [6]	3.06	0.972
Proposed S^{3}G-Net	3.613	0.951

auxiliary branches or post-filtering. Even under reconstruction-driven training, $\mathcal{L}_{\mathrm{STFT}}$ alone yields PESQ 2.98, before incorporating perceptual or self-supervised supervision, which increases to 3.11 after adding $\mathcal{L}_{\mathrm{CBAK,TV}}$, indicating that the architecture provides a strong causal prior. For robustness (Table II), VB+DEMAND-trained models are evaluated on unseen VB+DEMAND-Ex [27] and DNS2020 real recordings [20]. S^{3}G-Net achieves the highest PESQ on VB+DEMAND-Ex (2.47) and remains competitive on DNS2020 (DNSMOS OVRL 2.53), suggesting good transfer under domain shift. Similar conclusions are derived on the DNS Challenge test set (Table III), where S^{3}G-Net achieves WB-PESQ 3.613 and STOI 0.951. We further analyze the source of these improvements in Fig. 1, which illustrates the intended GraphFreq–SSM coupling. For example, a small example region in Fig. 1 shows the enhanced harmonic structure coincides with increased SSM gating $g_a(t,l) = \sigma(w_a g_{t,l} + b_a)$ and a longer effective memory horizon $\tau_{t,l} = -1/\log(|\bar{a}_{\mathrm{eff}}(t,l)| + \varepsilon)$ (i.e., $\Delta_{\mathcal{R}}\tau > 0$), which is consistent with band-adaptive temporal retention when voiced structure is supported by cross-band diffusion.

It is worth noting that on a notebook-class Intel Core i5 CPU, our proposed S^{3}G-Net reaches a real-time factor (RTF) of ≈ 0.10 on VoiceBank+DEMAND, comparable to or lower than the RTFs reported for DeepFilterNet2 [14] and LiSenNet [16], without relying on post filtering as [14], [16]. Further, all metrics and costs correspond to a single-pass predictor.

Ablation Study: Table IV shows that removing the inter-path dilated TCN gives the largest single-module drop (**-TCN_inter**; PESQ: -0.352, STOI: -0.020), confirming it as the main causal long-context stabilizer against temporal modulation. Removing the second dual-path stage (**–2nd DPGRU**; PESQ: -0.297) also degrades quality, while ablating GraphFreq (**–GraphFreq**; PESQ: -0.137, STOI: -0.016) hurts both quality and intelligibility, supporting the role of cross-band diffusion in preserving harmonic/formant structure. ConvGLU has a smaller but consistent effect (**–ConvGLU**; PESQ: -0.057). The fixed

TABLE IV
VOICEBANK+DEMAND ABLATION. Δ = DROP FROM FULL.

Variant	Params [k]	MACs [G]	PESQ	ΔP↓	STOI	ΔS↓
Full S^{3}G-Net (GC–SSM)	30	0.063	3.607	0.000	0.938	0.000
GraphFreq choice						
GraphFreq fixed (banded mix)	30	0.063	3.607	0.00	0.938	0.00
GraphFreq learned (Theta)	30	0.063	3.531	0.076	0.937	0.001
Module removal						
–ConvGLU	29	0.062	3.550	0.057	0.931	0.007
–GraphFreq	28	0.061	3.470	0.137	0.922	0.016
–TCN_inter	25	0.055	3.255	0.352	0.918	0.020
–2nd DPGRU	24	0.056	3.310	0.297	0.919	0.019
Loss ablation						
$\mathcal{L}_{\mathrm{STFT}}$ only	30	0.063	2.98	–	0.934	–
$+\mathcal{L}_{\mathrm{CBAK,TV}}$	30	0.063	3.11	–	0.936	–
$+\mathcal{L}_{\mathrm{PESQ}}$	30	0.063	3.471	–	0.927	–
$+\mathcal{L}_{\mathrm{WavLM}}$	30	0.063	3.559	–	0.933	–
$+\mathcal{L}_{\mathrm{STOI}}$ (= Full)	30	0.063	3.607	–	0.938	—

banded GraphFreq adjacency slightly outperforms learned Θ (PESQ 3.607 vs. 3.531), suggesting that a structured local prior is more stable at this scale. Module and loss ablations address complementary questions. The PESQ-related term shifts the model to a stronger perceptual operating point, but this gain is not due to the loss alone. With the same objective, removing the TCN, GraphFreq residual, ConvGLU, or second dual-path stage consistently reduces PESQ and/or STOI. Thus, the final performance reflects the interaction between metric-aligned supervision and the compact causal architecture, whose cross-band diffusion and long-context modeling allow perceptual and intelligibility losses to be used without destabilizing speech structure under a 30k-parameter budget. With $\mathcal{L}_{\mathrm{STFT}}$ alone, S^{3}G-Net remains competitive with causal STFT-trained baselines [14]–[16], [30], [32], although perceptual quality is limited (PESQ 2.98). Adding $\mathcal{L}_{\mathrm{CBAK,TV}}$ improves background cleanliness and smoothness (3.11). $\mathcal{L}_{\mathrm{PESQ}}$ gives the largest PESQ gain (3.471) but lowers STOI (0.927 vs. 0.934), showing the quality–intelligibility trade-off. $\mathcal{L}_{\mathrm{WavLM}}$ improves perceived quality and linguistic consistency (3.559/0.933), and the final $\mathcal{L}_{\mathrm{STOI}}$ restores the best intelligibility while preserving the top PESQ (3.607/0.938). This shows that the loss terms are complementary, while S^{3}G-Net provides the compact 30k-parameter structure needed to exploit them effectively.

IV. CONCLUSION

In this paper, we presented S^{3}G-Net, a causal, single-pass CRM model for real-time SE under edge constraints. A single branch couples a band-aware encoder with dynamic sub-band gating, a per-frame graph–frequency residual, and a lightweight time core that combines a diagonal state-space module with a dilated causal TCN. This architecture concentrates capacity on rapidly varying bands, enforces harmonic/near-diagonal spectral coupling, and provides both adaptive short-range memory and long-range temporal context without look-ahead or post-filtering, thus, latency and compute reflect the deployed predictor. S^{3}G-Net improves perceptual quality in the causal regime, narrowing the gap to heavier non-causal systems, especially under non-stationary interference. Ablations show the inter-path TCN drives long-range modeling, stacked dual-path layers add depth, the graph residual stabilizes spectra, and ConvGLU reduces mask jitter at negligible cost.

REFERENCES

[1] C. K. A. Reddy, N. Shankar, G. S. Bhat, R. Charan, and I. Panahi, "An individualized super-gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1601–1605, 2017.

[2] J. Chakraborty, M. Reed, N. Thomos, G. Pratt, and N. Wilson, "Neuro-fuzzy voice quality improvement for low-power, half-duplex, communication," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3609–3622, 2025.

[3] X. Rong, T. Sun, X. Zhang, Y. Hu, C. Zhu, and J. Lu, "GTCRN: A speech enhancement model requiring ultralow computational resources," in *Proc. of IEEE ICASSP*, 2024, pp. 971–975.

[4] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. of Interspeech*, 2018, pp. 3229–3233.

[5] Z. Guo, J. Du, S. M. Siniscalchi, J. Pan, and Q. Liu, "Controllable conformer for speech enhancement and recognition," *IEEE Signal Processing Letters*, 2024.

[6] F. Dang, H. Chen, Q. Hu, P. Zhang, and Y. Yan, "First coarse, fine afterward: A lightweight two-stage complex approach for monaural speech enhancement," *Speech Communication*, vol. 146, pp. 32–44, 2023.

[7] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[8] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. of Interspeech*, 2020, pp. 2472–2476.

[9] C. Wang, J. Gu, D. Yao, J. Li, and Y. Yan, "GALD-SE: Guided anisotropic lightweight diffusion for efficient speech enhancement," *IEEE Signal Processing Letters*, 2024.

[10] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proc. of Interspeech*, 2021, pp. 201–205.

[11] S. Abdulatif, R. Cao, and B. Yang, "CMGAN: Conformer-based metric-gan for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2477–2493, 2024.

[12] R. Chao, W.-H. Cheng, M. L. Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, "An investigation of incorporating mamba for speech enhancement," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 302–308.

[13] N. L. Kühne, J. Jensen, J. Østergaard, and Z.-H. Tan, "Exploring resolution-wise shared attention in hybrid mamba-u-nets for improved cross-corpus speech enhancement," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 16 327–16 331.

[14] H. Schröter, A. Maier, A. N. Escalante-B, and T. Rosenkranz, "DeepFilterNet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *Proc. of IEEE Int. Workshop on Acoustic Signal Enhancement*, 2022, pp. 1–5.

[15] L. Yang, W. Liu, R. Meng, G. Lee, S. Baek, and H.-G. Moon, "FSPEN: an ultra-lightweight network for real time speech enahncment," in *Proc. of IEEE ICASSP*, 2024, pp. 10 671–10 675.

[16] H. Yan, J. Zhang, C. Fan, Y. Zhou, and P. Liu, "LiSenNet: Lightweight sub-band and dual-path modeling for real-time speech enhancement," in *Proc. of IEEE ICASSP*, 2025, pp. 1–5.

[17] Y.-C. Lin, C. Yu, Y.-T. Hsu, S.-W. Fu, Y. Tsao, and T.-W. Kuo, "SEOFP-NET: Compression and acceleration of deep neural networks for speech enhancement using sign-exponent-only floating-points," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1016–1031, 2022.

[18] K. Tan and D. Wang, "Compressing deep neural networks for efficient speech enhancement," in *Proc. of IEEE ICASSP*, 2021, pp. 8358–8362.

[19] C. Valentini Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *Proceedings of 9th ISCA Speech Synthesis Workshop*, Sep. 2016, pp. 159–165.

[20] H. Wang and B. Tian, "ZipEnhancer: Dual-path down-up sampling-based zipformer for monaural speech enhancement," in *Proc. of IEEE ICASSP*, 2025, pp. 1–5.

[21] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *Proc. of IEEE ICASSP*, 2022, pp. 9281–9285.

[22] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "FullSubNet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *Proc. of IEEE ICASSP*, 2022, pp. 7857–7861.

[23] J. Wang, Z. Lin, T. Wang, M. Ge, L. Wang, and J. Dang, "Mamba-SEUNet: Mamba UNet for monaural speech enhancement," in *Proc. of IEEE ICASSP*, 2025, pp. 1–5.

[24] M. Kolbæk, Z.-H. Tan, and J. Jensen, "On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 283–295, 2018.

[25] N. Babaev, K. Tamogashev, A. Saginbaev, I. Shchekotov, H. Bae, H. Sung, W. Lee, H.-Y. Cho, and P. Andreev, "FINALLY: fast and universal speech enhancement with studio-like quality," *Advances in Neural Information Processing Systems*, vol. 37, pp. 934–965, 2024.

[26] N. L. Kühne, J. Jensen, J. Østergaard, and Z.-H. Tan, "VB-DemandEx," Jul. 2025, dataset. [Online]. Available: https://doi.org/10.57967/hf/5907

[27] N. L. Kühne, J. Jensen, J. Østergaard, and Z.-H. Tan, "MambAttention: Mamba with multi-head attention for generalizable single-channel speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 820–833, 2026.

[28] N. L. Kühne, J. Østergaard, J. Jensen, and Z.-H. Tan, "xLSTM-SENet: xLSTM for Single-Channel Speech Enhancement," in *Proc. of Interspeech*, 2025, pp. 5148–5152.

[29] H. Wang, C. Wang, X. Wang, L. Yu, and Y. Jiang, "MBTU-SE: A speech enhancement network integrates enhanced taylor multi-branch linear transformer with u-net architecture," *IEEE Signal Processing Letters*, vol. 32, pp. 4309–4313, 2025.

[30] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *Proc. of 20th IEEE Int. workshop on multimedia signal processing*, 2018, pp. 1–5.

[31] J. Lee and H.-G. Kang, "Two-stage refinement of magnitude and complex spectra for real-time speech enhancement," *IEEE Signal Processing Letters*, vol. 29, pp. 2188–2192, 2022.

[32] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. of IEEE ICASSP*, 2022, pp. 7407–7411.