Speech Enhancement with Lightweight Neural Networks

ELEC5305 Acoustics, Speech and Signal Processing
Author: Wenye Zhang
SID: 540473182
Data: 16 November

Abstract

This project investigates lightweight neural speech enhancement in the short-time Fourier transform (STFT) domain, with the aim of improving perceptual quality under realistic noise conditions while keeping the model small enough for CPU deployment. The proposed architecture estimates a complex time–frequency mask for the noisy STFT and is built around a DPCRN-style dual-path bottleneck, combining a shallow encoder–decoder with recurrent modelling along both time and frequency. On top of this backbone, two design choices are explored: (i) complex convolutions in the encoder–decoder to process real and imaginary parts jointly, and (ii) a frame-wise self-attention block at the bottleneck to provide global temporal context at moderate cost. The model is trained and evaluated on the VoiceBank+DEMAND corpus using a joint loss on compressed STFT magnitudes and time-domain waveforms. With fewer than one million trainable parameters, the full configuration achieves reasonable performance on PESQ, STOI and ΔSI-SDR. Ablation experiments show that removing the frame-wise attention block leads to a clearer degradation in these metrics than removing complex convolutions, even though the latter saves more parameters. This suggests that global temporal context at the bottleneck is particularly important in the lightweight setting, while the benefit of full complex convolutions in all encoder–decoder layers is comparatively modest. The work provides a concrete baseline and several design insights for future lightweight speech enhancement models.

Contents

1. Introduction

Speech enhancement aims to recover clean and intelligible speech from signals corrupted by environmental noise and reverberation, and is widely used in online meetings, voice assistants and automatic speech recognition (ASR). A common and practical choice is to work in the short-time Fourier transform (STFT) domain and estimate a time–frequency mask that modifies the noisy spectrum before inverse STFT. Recent neural approaches show that modelling the complex spectrum, rather than magnitude only, can better exploit the relationship between real and imaginary parts and improve perceptual quality (Hu et al., 2020).

At the same time, many state-of-the-art models are still relatively heavy and are designed with GPU deployment in mind. This limits their usefulness on CPU-only laptops or embedded devices, where model size, latency and power consumption are constrained. In this project I focus on single-channel, lightweight neural speech enhancement in the STFT domain. The goal is to design and study a compact model that estimates a complex-valued time–frequency mask while remaining suitable for near real-time CPU execution. Concretely, I build on a dual-path convolutional recurrent network (DPCRN) backbone that combines a convolutional encoder–decoder with dual-path recurrent blocks in the time–frequency domain (Le et al., 2021), and extend it by introducing complex-valued convolutions and a frame-wise self-attention module inspired by recent time–frequency attention architectures (Wang et al., 2022).

The remainder of this report reviews related work on lightweight and complex-domain speech enhancement, describes the proposed method in detail, presents experimental results, and discusses the effectiveness and limitations of the approach.

2. Literature Review

This section reviews lightweight neural speech enhancement work that is directly relevant to my project. Rather than describing each paper in detail, I group the methods into three themes: (i) lightweight encoder–decoder designs, (ii) complex-domain and dual-path modelling, and (iii) attention and full-band/sub-band strategies. I then summarise how these ideas motivate the DPCRN-based approach adopted in this project.

2.1 Lightweight encoder–decoder architectures

A first line of work focuses on building compact encoder–decoder backbones with carefully designed blocks. UL-UNAS treats lightweight speech enhancement as an architecture search problem, using neural architecture search to discover an ultra-lightweight U-Net with affine PReLU and causal time–frequency attention (Rong et al., 2025). MA-Net and related studies investigate resource-efficient multi-attentional networks and compare several shallow encoder–decoder and recurrent structures under strict complexity constraints (Wahab et al., 2024). These works are valuable because they explicitly report the trade-off between perceptual performance and model size rather than only presenting large "competition" systems.

However, most of these architectures operate on magnitude spectra or real-valued features. Phase is either ignored or handled implicitly, which limits their ability to recover fine phase details compared with complex-domain models. In addition, while parameter counts and multiply–accumulate (MAC) operations are reported, detailed analysis of CPU latency and real-time behaviour is often limited. For my project, these papers mainly provide design principles—keep the encoder–decoder shallow, use narrow channels, and avoid overly large recurrent modules—rather than a concrete complex-domain backbone to copy.

2.2 Complex-domain and dual-path modelling

Complex-domain models explicitly treat real and imaginary parts as coupled variables. DCCRN is a representative example, simulating complex-valued operations in both convolutional and recurrent layers and showing clear gains over magnitude-only baselines (Hu et al., 2020). This confirms that modelling the complex spectrum can improve perceptual quality, but the original DCCRN is still relatively heavy for CPU-only deployment.

Dual-path architectures such as DPCRN offer a more lightweight alternative. DPCRN combines a convolutional encoder–decoder with dual-path RNN modules that separately model dependencies along frequency and time in the time–frequency (T–F) domain (Le et al., 2021). With fewer than one million parameters, DPCRN achieves competitive single-channel enhancement and explicitly targets real-time scenarios. Related sub-band dual-path designs, such as LiSenNet, further show that chunk-based dual-path processing can capture both local and long-range context with modest hidden sizes (Yan et al., 2023). The downside is that some of these models work directly in the time domain or rely on sophisticated sub-band layouts that are difficult to implement within a small project.

From my perspective, these works suggest that a DPCRN-style dual-path backbone is a good starting point: it is compact, has a clear separation between local convolutional feature extraction and longer-range recurrent modelling, and can be naturally extended to complex-domain processing by replacing real convolutions with complex ones.

2.3 Attention and full-band / sub-band strategies

Attention mechanisms and band-structured networks offer another way to improve efficiency. Full-band/sub-band fusion models use separate branches to process local sub-band patterns and global full-band context, achieving a good balance between performance and parameter count (Chen & Zhang, 2022; Hao & Li, 2023). Coarse–fine strategies in the complex domain similarly allocate more capacity to a refinement stage after an initial coarse enhancement (Dang et al., 2023). These ideas show that not all frequencies or time–frequency locations need the same computational budget.

At the same time, recent time–frequency models for separation show how attention can be used to share information across frames. TF-GridNet stacks spectral modules, sub-band temporal modules and full-band self-attention over frames in the complex T–F domain, and achieves strong separation performance (Wang et al., 2022). UL-UNAS and MA-Net also include lightweight attention blocks in their encoders (Rong et al., 2025; Wahab et al., 2024). The common message is that small, well-placed attention modules can significantly strengthen spectral–temporal modelling. The challenge for my project is to use this idea without adopting the full complexity of TF-GridNet or multi-branch fusion networks.

2.4 Summary and motivation for this project

Across these studies, several themes emerge. Lightweight encoder–decoder designs highlight the importance of shallow networks and narrow channels but often underuse complex-domain information. Complex-domain and dual-path frameworks demonstrate that jointly modelling real and imaginary parts and using dual-path RNNs can improve quality with manageable parameter counts, yet models such as DCCRN remain relatively heavy and some dual-path variants are architecturally complex. Attention and band-structured methods show that selective use of attention and sub-band processing is effective, but full implementations like TF-GridNet are designed for GPU-oriented separation tasks rather than lightweight enhancement .

The gap that motivates this project is therefore a simple, implementable backbone that simultaneously:

1. operates in the complex STFT domain,

2. uses dual-path modelling to capture time–frequency context, and

3. includes a lightweight attention module for frame-wise information sharing,
while keeping the overall complexity reasonable for CPU execution.

To address this, I start from a DPCRN-style dual-path encoder–decoder (Le et al., 2021), explicitly introduce DCCRN-inspired complex convolutions in the encoder and decoder (Hu et al., 2020), and add a simplified frame-wise self-attention block at the bottleneck, following the spirit of full-band attention in TF-GridNet (Wang et al., 2022).

3. Methodology

This project adopts a lightweight, STFT-based speech enhancement model that estimates a complex-valued time–frequency (T–F) mask for the noisy spectrum. The overall architecture (Fig. 1) follows a DPCRN-style encoder–decoder backbone with dual-path recurrent blocks, and it is extended in two ways: (i) the encoder and decoder are implemented with explicit complex convolutions to better handle real and imaginary parts; and (ii) a frame-wise self-attention module is inserted in the bottleneck to enhance cross-frame modeling at moderate cost. This section first introduces the STFT-based complex mask formulation, and then details the backbone, complex convolution module, and frame-wise attention module.

3.1 STFT-based complex mask

Given a noisy time-domain signal y(t) and a clean reference signal s(t), this project applies the short-time Fourier transform (STFT) with a Hann window and fixed hop size to obtain complex spectra

$$Y(f,k) = \mathscr{F}\{y(t)\}, \quad S(f,k) = \mathscr{F}\{s(t)\}$$

where f denotes the frequency bin and k the frame index.
Instead of directly predicting $\widehat{S}(f, k)$, the network estimates a complex-valued T–F mask $M(f, k)$ and computes the enhanced spectrum as

$$\hat{S}(f,k)=M(f,k)\,Y(f,k)$$

This mask-based formulation has two advantages. First, by conditioning on the noisy spectrum Y, the network can focus on predicting corrections rather than absolute values, which is empirically easier to learn. Second, allowing $M(f, k)$ to be complex enables explicit phase modification, which has been shown to improve perceptual quality compared with magnitude-only masks (Hu et al., 2020). After obtaining S(f, k), this project applies inverse STFT (iSTFT) to reconstruct the enhanced waveform $\hat{s}(t)$

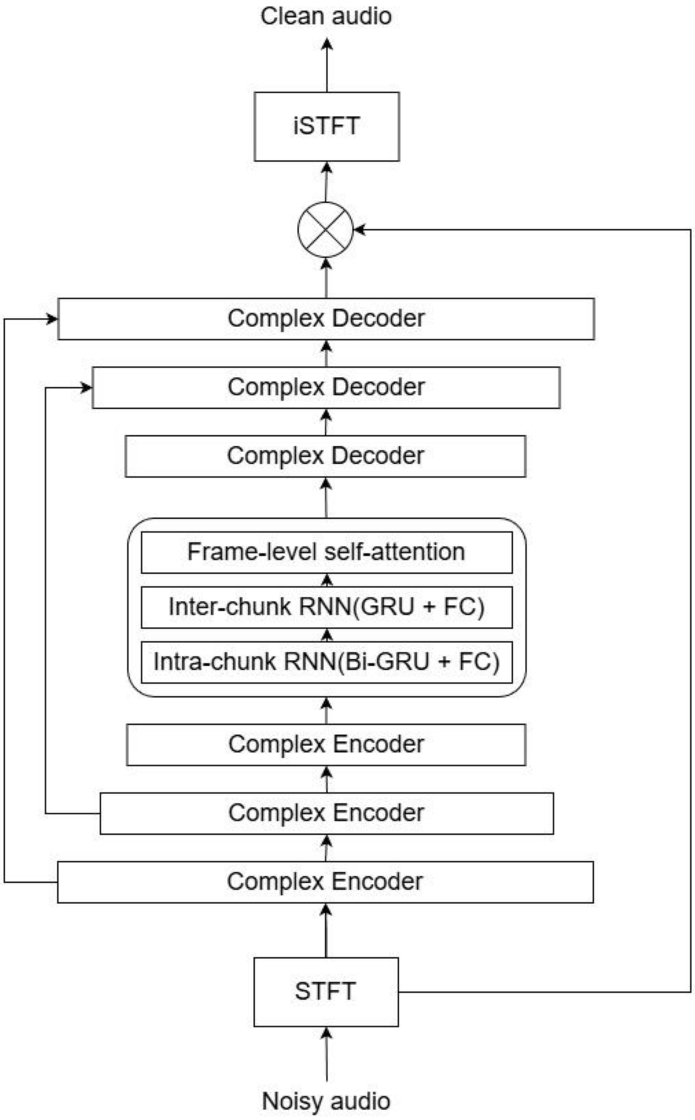

Figure 1: Overall network framework

3.2 Dual-path recurrent bottleneck

Between the complex encoder and decoder, the model uses a dual-path recurrent bottleneck to capture long-range dependencies in both time and frequency. Let

$$X_g \in \mathbb{R}^{B\times C_g\times T\times F}$$

denote the real-valued bottleneck feature map obtained after stacking the real and imaginary channels from the encoder, where *B* is batch size, *Cg* is the channel dimension, *T* is the number of frames and *F* is the down-sampled frequency dimension.

The dual- path idea follows DPCRN (Le et al., 2021): instead of running one large RNN over the whole *(T, F)* grid, this project alternates intra-frequency and inter-time RNNs, each operating on a 1-D sequence while keeping the other axis folded into the batch.

Intra-frequency RNN.
For the intra stage, we treat each time frame as a sequence over frequency. The feature is reshaped to

$$H_{in} \in \mathbb{R}^{(BT)\times F\times C_g}$$

and a bidirectional GRU models dependencies along the F axis:

$$\widetilde{H}_{in} = \mathrm{BiGRU}(H_{in})$$

A linear layer projects the GRU output back to Cg channels, and layer normalization plus a residual connection map it back to the original shape *[B,Cg,T,F]*. This stage allows the model to exploit correlations between neighbouring and distant frequency bins within each frame, which is important for modelling formants and harmonic structure.

Inter-time RNN.
For the inter stage, we instead treat each frequency bin as a sequence over time. The feature is reshaped to

$$H_{out} \in \mathbb{R}^{(BT)\times F\times C_g}$$

and a (causal) GRU runs along the T axis:

$$\widetilde{H}_{\mathrm{out}} = \mathrm{GRU}(H_{out})$$

Compared with using a single large 2-D RNN or transformer on the full T–F plane, the dual-path design offers a better performance–complexity trade-off:

1. Each GRU only sees a 1-D sequence, so the hidden size can be relatively small while still covering long context.
2. Alternating intra-frequency and inter-time stages means that information can propagate across both axes, approximating 2-D context with significantly fewer parameters and multiply–accumulate operations.

This makes the dual-path bottleneck a natural choice for a lightweight model that still needs strong temporal and spectral modelling, which is exactly the role it plays in DPCRN (Le et al., 2021) and in this project.

3.3 Complex convolution module

The encoder and decoder blocks use complex convolutions to process real and imaginary parts jointly, as illustrated in Fig. 2. Let the complex feature map be

$$X = Xr + jXi,$$

where *Xr* and *Xi* are the real and imaginary parts. Following the formulation of complex convolution in DCCRN (Hu et al., 2020), a complex kernel

$$W = Wr\ +\ jWi$$

acts on *X* as

$$Y = X *\ W,$$

which can be decomposed into real and imaginary parts:

$$Y_r = X_r *\ W_r - X_i *\ W_i,$$
$$Y_i = X_r *\ W_i + X_i *\ W_r$$

where * denotes 2-D convolution.

In the implementation, the input channels are arranged as *[Xr,Xi]*. Two real-valued Conv2d layers compute *Xr* * *Wr, Xi* * *Wr, Xr* * *Wi* and *Xi* * *Wi*; these are then combined according to the above equations to produce the new real and imaginary parts. Each complex encoder block consists of:

1. a complex convolution (ComplexConv2d),

2.complex batch normalization (separate BN for real and imaginary channels),

3. complex PReLU activation (independent PReLU for each part).

Decoder blocks use the same structure but with transposed convolutions.

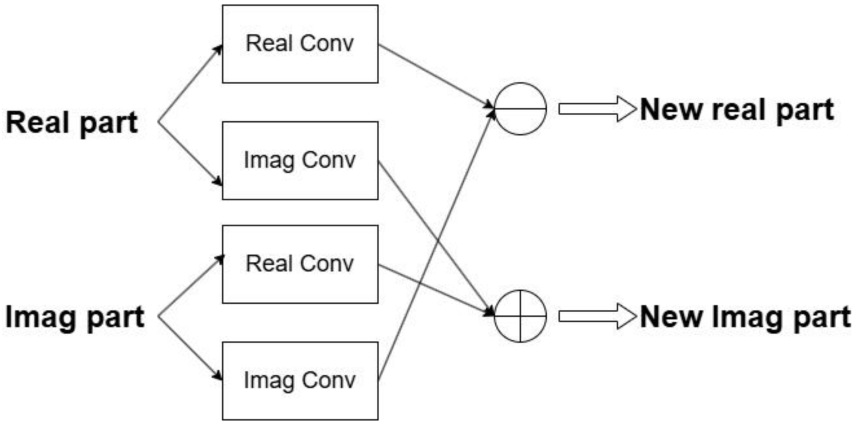

Figure 2: The operational logic of complex convolution

3.4 Frame-wise self-attention module

To further strengthen temporal modelling without making the network much larger, the bottleneck is augmented with a frame-wise multi-head self-attention module, shown in Fig. 3. The input is the bottleneck feature

$$X_g \in \mathbb{R}^{B\times C_g\times T\times F}$$

The idea is to treat each time frame as a token whose feature dimension aggregates information over frequency, and to allow all frames to attend to each other.

First, three times 1×1 convolutions generate query, key and value tensors:

$$\mathrm{Q},\, \mathrm{K} \in \mathbb{R}^{B\times(\mathrm{EL})\times T\times F},\, \mathrm{V} \in \mathbb{R}^{B\times(\mathrm{D_h L})\times T\times F}$$

where L is the number of heads, E is the query/key dimension per head, and $D_\mathrm{h}=C_g/L$ is the value dimension per head. For each head L, we reshape to

$$\mathrm{Q_L},\, \mathrm{K_L} \in \mathbb{R}^{T\times FE},\, \mathrm{V_L} \in \mathbb{R}^{T\times(\mathrm{D_h}F)}$$

Self-attention for head L is then computed as

$$A_l=\mathrm{softmax}\left(\frac{Q_l K_l^\top}{\sqrt{LFE}}\right), \quad O_l=A_l V_l$$

where $A_l \in \mathbb{R}^{T\times T}$ contains frame-to-frame attention weights and $O_l \in \mathbb{R}^{T\times(FD_\mathrm{h})}$ is the attended output. Outputs from all heads are concatenated, reshaped back to *[B,Cg,T,F]*, passed through a final 1×1 convolution and normalization layer, and added to the original Xg via a residual connection.

Compared with full 2-D attention over both time and frequency, the proposed frame-wise attention only operates along the time axis and folds the frequency axis into the token features, reducing the complexity from *O((TF)^2)* to *O(T^2)*. Placing this module only once at the bottleneck—where F is already down-sampled—still allows each frame to access information from all other frames, similar in spirit to full-band attention in TF-GridNet (Wang et al., 2022), but at much lower cost. Together with the dual-path bottleneck and complex convolutions, this gives a lightweight way to add global temporal context that matches the CPU-oriented design goal of the project.

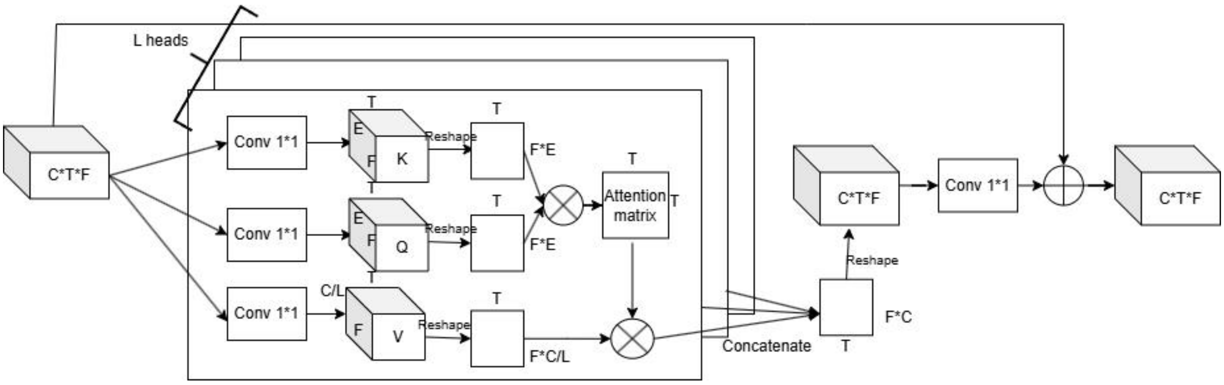

Figure 3: Inter-frame self-attention module framework diagram

4.Experiment

4.1 Dataset

This project uses the VoiceBank+DEMAND single-channel speech enhancement corpus. In this dataset, clean speech from multiple English speakers is taken from the VoiceBank (VCTK) recordings and artificially mixed with diverse real-world noises from the DEMAND database, covering domestic, office and outdoor environments at several signal-to-noise ratio (SNR) levels (Valentini-Botinhao, 2017). The result is a paired noisy–clean corpus that has become a standard benchmark for supervised speech enhancement, and is widely used to evaluate both complex-domain and lightweight neural models. It provides a controlled but realistic setting to assess how well the proposed architecture can improve perceptual speech quality under different noise conditions.

4.2 Loss function

The model is trained to predict a complex mask that improves both the spectral representation and the time-domain waveform. To balance these two aspects, I use a combined loss consisting of a compressed magnitude loss in the STFT domain and an L1 loss in the time domain.

Let *S(f, k)*and $\hat{S}(f, k)$ be the clean and enhanced complex spectra, and *s(t)* and $\hat{s}(t)$ be the corresponding time-domain signals. The spectral loss is defined on compressed magnitudes as

$$\mathscr{L}_{mag}=E\left[\left(|\hat{S}(f,k)|^{\gamma}-|S(f,k)|^{\gamma}\right)^{2}\right]$$

The waveform loss is given by the mean absolute error between the enhanced and clean signals:

$$\mathcal{L}_{wav}=E[\,|\hat{s}(t)-s(t)|\,]$$

The total training objective is a weighted sum of the two terms:

$$\mathcal{L}=\alpha\,\mathcal{L}_{mag}+(1-\alpha)\,\mathcal{L}_{wav}$$

with $\alpha=0.5$ in all experiments.

The spectral term encourages the network to match the overall time–frequency structure of the clean signal, while the waveform term stabilises the optimisation and directly penalises audible distortions in the time domain. Using both is a simple way to constrain the model from two complementary views without introducing additional complexity into the architecture.

4.3 Result

The ablation study on the proposed lightweight model is summarised in Table 1. All experiments are conducted on the VoiceBank+DEMAND test set and evaluated using PESQ, STOI and ΔSI-SDR. The full model, which uses complex convolutions in the encoder–decoder and a frame-wise self-attention block at the bottleneck, has 0.676 M parameters and achieves 3.035 PESQ, 94.8% STOI and 10.68 dB ΔSI-SDR. This serves as the main reference configuration for the rest of the analysis.

Removing complex convolutions while keeping the rest of the architecture unchanged reduces the parameter count to 0.411 M. In this “no complex convolution” variant, the real and imaginary parts of the STFT are simply treated as separate real channels processed by standard convolutions. The resulting performance shows only a small but consistent drop compared to the full model (PESQ 3.028, STOI 94.4%, ΔSI-SDR 10.57 dB). This indicates that explicitly modelling the coupling between real and imaginary parts does improve quality slightly, but the gain is modest relative to the additional parameters introduced by full complex convolutions.

In the second ablation, the frame-wise self-attention block at the bottleneck is removed while complex convolutions are kept. This “no frame-wise self-attention” variant has 0.638 M parameters, which is close to the full model, but the performance is degraded more noticeably: PESQ drops to 2.973, STOI to 94.6%, and ΔSI-SDR to 10.52 dB. Compared with this variant, the full model improves PESQ by about 0.06 and ΔSI-SDR by 0.16 dB. These differences are not large in absolute terms, but they are consistent across metrics and show that inserting a single frame-wise attention block provides a clear benefit with only a slight increase in model size.

Overall, Table 1 shows that all three configurations remain in the sub-million-parameter range, and that the proposed full model offers the best balance between complexity and perceptual quality. The detailed implications of these ablations for complex convolution design and temporal modelling are discussed in the following Analysis section.

	Parameter(M)	**PESQ**	**STOI(%)**	**Delta SI-SDR(dB)**
Final model	0.676	**3.035**	**94.8**	**10.68**
No complex convolution	**0.411**	3.028	94.4	10.57
No frame-wise Self-attention	0.638	2.973	94.6	10.52

Table 1: Metrics of lightweight models on the VoiceBank-DEMAND dataset and ablation experiments

5. Analysis

The ablation results mainly reflect two design choices: whether to use complex convolutions in the encoder–decoder, and whether to include the frame-wise self-attention block at the bottleneck. Overall, the full model gives the best PESQ and ΔSI-SDR, but the way these gains are achieved is not equally efficient across components.

From a parameter–performance perspective, the frame-wise attention appears to be the more cost-effective component. Removing this block leads to a noticeable drop in PESQ and ΔSI-SDR, even though the parameter count is only slightly reduced. In other words, a relatively small attention module placed at the bottleneck contributes a clear and consistent improvement in perceptual quality. This matches the intuition from TF-GridNet-style designs: allowing each frame to aggregate information from all other frames can strengthen temporal modelling even when the attention layer itself is lightweight.

By contrast, the effect of complex convolutions is more subtle. When they are removed and the real and imaginary parts are simply treated as separate real channels, the model becomes significantly smaller, but the degradation in PESQ and ΔSI-SDR is quite modest. This suggests that, in the current architecture and training setup, the complex convolution design does help, but the gain is not proportional to its parameter cost. A plausible explanation is that most of the "heavy lifting" is done by the dual-path recurrent bottleneck and the attention block, while the encoder–decoder mainly needs to provide a reasonably good time–frequency representation. Under a loss function that still focuses strongly on magnitude structure, the network may not fully exploit the more precise real–imaginary coupling that complex convolutions provide.

This raises the question of whether there is a fundamental problem with the way complex convolutions are used here. Conceptually, the idea of modelling real and imaginary parts jointly is sound and consistent with existing complex-domain work. The potential issue is more practical: implementing every encoder and decoder layer as a full complex convolution is relatively expensive for a "lightweight" setting, and the rest of the architecture may already be strong enough to approximate the needed phase behaviour without strictly enforcing complex arithmetic everywhere. In that sense, the current design is likely "over-allocating" parameters to complex convolutions and "under-allocating" them to other parts that might have a larger impact on quality.

Taken together, the analysis suggests that the overall direction—DPCRN-style bottleneck plus a small attention block—is reasonable for lightweight enhancement, but the complex convolution part could be further simplified or redesigned. For example, it may be more effective to use lighter variants (e.g., depthwise or partially shared complex kernels), or to restrict complex operations to a few key layers while reallocating some of the saved capacity to the bottleneck or attention module. This would keep the model within a similar parameter budget, while potentially improving the performance–complexity balance observed in the current experiments.

6. Conclusion

This project investigated lightweight STFT-based speech enhancement using a DPCRN-style dual-path bottleneck, complex convolutions in the encoder–decoder, and a frame-wise self-attention block. The model estimates a complex time–frequency mask for the noisy STFT and was trained and evaluated on the VoiceBank+DEMAND corpus. With fewer than one million trainable parameters, the full configuration achieves reasonable PESQ, STOI and ΔSI-SDR, indicating that a compact dual-path backbone combined with a single attention layer can still provide useful enhancement under realistic noise conditions.

The ablation results suggest that frame-wise attention is more cost-effective than enforcing complex convolutions at every encoder–decoder layer: removing the attention block causes a clearer performance drop than removing complex convolutions, even though the latter saves more parameters. This indicates that global temporal context at the bottleneck is particularly important in this setting, while the current complex convolution design offers only modest gains for its cost. A natural next step would be to simplify or partially replace complex convolutions (e.g., with lighter or selectively applied variants) and reallocate the saved capacity to the bottleneck and attention module, as well as to extend evaluation to additional datasets and explicit CPU runtime measurements. Overall, the work provides a concrete baseline and some practical design hints for future lightweight neural speech enhancement models.

References

Cao, Y., Xu, S., Zhang, W., Wang, M., & Lu, Y. (2025). Hybrid lightweight temporal-frequency analysis network for multi-channel speech enhancement. EURASIP Journal on Audio, Speech, and Music Processing, 2025, 21. https://doi.org/10.1186/s13636-025-00408-3

Chen, Z., & Zhang, P. (2022). Lightweight full-band and sub-band fusion network for real time speech enhancement. In Interspeech 2022: Conference of the International Speech Communication Association (pp. 921–925).

Dang, F., Chen, H., Hu, Q., Zhang, P., & Yan, Y. (2023). First coarse, fine afterward: A lightweight two-stage complex approach for monaural speech enhancement. Speech Communication, 146, 32–44. https://doi.org/10.1016/j.specom.2022.11.004

Fan, Z., Guo, Z., Lai, Y., & Kim, J. (2025). TSDCA-BA: An ultra-lightweight speech enhancement model for real-time hearing aids with multi-scale STFT fusion. Applied Sciences, 15(15), 8183. https://doi.org/10.3390/app15158183

Hao, X., & Li, X. (2023). Fast FullSubNet: Accelerate full-band and sub-band fusion model for single-channel speech enhancement. arXiv preprint arXiv:2212.09019.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In Interspeech 2020: Conference of the International Speech Communication Association (pp. 2472–2476). https://doi.org/10.21437/Interspeech.2020-2537

Le, X., Chen, H., Chen, K., & Lu, J. (2021). DPCRN: Dual-path convolution recurrent network for single channel speech enhancement. In Interspeech 2021: Conference of the International Speech Communication Association (pp. 2826–2830). https://doi.org/10.21437/Interspeech.2021-296

Rong, X., Wang, D., Hu, Y., Zhu, C., Chen, K., & Lu, J. (2025). UL-UNAS: Ultra-lightweight U-Nets for real-time speech enhancement via network architecture search. arXiv preprint arXiv:2503.00340.

Rosenbaum, T., Winebrand, E., Cohen, O., & Cohen, I. (2025). Deep-learning framework for efficient real-time speech enhancement and dereverberation. Sensors, 25(3), 630. https://doi.org/10.3390/s25030630

Wahab, F. E., Ye, Z., Saleem, N., & Ullah, R. (2024). Compact deep neural networks for real-time speech enhancement on resource-limited devices. Speech Communication, 156, 103008. https://doi.org/10.1016/j.specom.2023.103008

Wahab, F. E., Ye, Z., Saleem, N., Ullah, R., & Hussain, A. (2025). MA-Net: Resource-efficient multi-attentional network for end-to-end speech enhancement. Neurocomputing, 619, 129150. https://doi.org/10.1016/j.neucom.2024.129150

Wang, Z.-Q., Cornell, S., Choi, S., Lee, Y., Kim, B.-Y., & Watanabe, S. (2023). TF-GridNet: Integrating full- and sub-band modeling for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 3221–3236. https://doi.org/10.1109/TASLP.2023.3304482

Yan, H., Zhang, J., Fan, C., Zhou, Y., & Liu, P. (2025). LiSenNet: Lightweight sub-band and dual-path modeling for real-time speech enhancement. In ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1–5). IEEE. https://doi.org/10.1109/ICASSP49660.2025.10888272

Valentini-Botinhao, C. (2017). Noisy speech database for training speech enhancement algorithms and TTS models [Data set]. University of Edinburgh.