Improving Single-Channel Speech Enhancement on DEMAND Dataset Using a Modified DPCRN Approach

T.D.H. Deiyagala
Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
Email: deiayagalatdh.21@uom.lk

***Abstract*—Speech enhancement (SE) aims to improve the quality and intelligibility of speech signals distorted by environmental noise. Single-channel SE is particularly challenging due to the absence of spatial cues. In this paper, we study the dual-path convolutional recurrent network (DPCRN) baseline for time-frequency domain SE and propose a simple, practical modification to improve its performance on the DEMAND dataset. The proposed approach incorporates a lightweight spectral compression mapping and a two-stage refinement process that first estimates a spectral magnitude mask and then optimizes the real and imaginary parts of the complex spectrum. Our modification is inspired by recent advancements in attention mechanisms, adaptive convolutional layers, and multi-loss strategies. Experiments demonstrate notable improvements in perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) while keeping the model lightweight. The approach remains practical for real-time implementation and can serve as a foundation for further research in low-complexity full-band SE.**

***Index Terms*—Speech enhancement, DPCRN, dual-path RNN, spectral compression mapping, DEMAND dataset, single-channel.**

I. Introduction

Speech enhancement (SE) plays a critical role in numerous applications including telecommunication, automatic speech recognition, hearing aids, and voice-controlled systems. Real-world environments often introduce non-stationary noise such as traffic, crowd chatter, or machine sounds, which can severely degrade the intelligibility and perceived quality of speech. While traditional rule-based signal processing techniques have shown limited success in controlled scenarios, deep learning-based approaches have demonstrated significant improvements by learning robust spectral representations from large datasets [1].

Single-channel SE is especially challenging because only the noisy speech signal is available, without access to spatial cues that multi-channel approaches can exploit. Most modern SE systems operate in the time-frequency (T-F) domain, leveraging the short-time Fourier transform (STFT) to separate noise from speech. Two primary approaches exist: estimating a mask to attenuate noise components [2] or directly predicting the complex spectrum [3]. Recent models combine convolutional neural networks (CNNs) to extract local spectral patterns and recurrent neural networks (RNNs) to capture temporal dependencies, providing an effective balance between performance and computational efficiency.

The dual-path convolutional recurrent network (DPCRN) [4] has emerged as a strong baseline for T-F domain SE. DPCRN employs dual-path recurrent modeling to efficiently capture both intra-frame spectral structures and inter-frame temporal correlations. While DPCRN has achieved state-of-the-art results on several benchmarks, its performance can degrade in full-band, high-resolution scenarios or in very low-SNR environments. Recent studies have explored augmentations such as multi-head attention, adaptive convolution, and multi-loss optimization to improve SE performance without substantially increasing computational complexity [5]–[7].

In this work, we focus on improving the DPCRN baseline for single-channel SE on the DEMAND dataset. Our contribution is a practical, lightweight modification inspired by learnable spectral compression mapping (SCM) and two-stage refinement, which enables the network to focus on low and mid-frequency speech components while preserving high-frequency details. The resulting model achieves improved perceptual and objective metrics while remaining suitable for real-time deployment.

II. Related Work

DPCRN introduced by Le et al. [4] combines convolutional feature extraction with dual-path recurrent modeling. The dual-path design allows modeling both short-term intra-frame correlations and long-term inter-frame dependencies. The original DPCRN serves as a strong baseline for T-F domain SE and has inspired numerous extensions and variants.

Wan et al. [5] proposed a Multi-Loss Time-Frequency Attention model that integrates axial self-attention within

DPCRN-style blocks. By combining multi-resolution STFT losses with perceptual losses derived from WavLM, the model achieves parameter-efficient gains on SE benchmarks. This highlights the effectiveness of incorporating attention mechanisms into dual-path structures.

Wang et al. [6] explored adaptive convolution layers in CNN-based SE models, showing that drop-in adaptive convolution can improve DPCRN-light metrics with minimal computational overhead. The work also compared various kernel attention variants, demonstrating how spectral adaptation can enhance representation in challenging noise conditions.

Peracha et al. [7] studied causal SE with dynamically weighted loss functions. Their approach improves artifact control and robustness by balancing contributions of magnitude and phase reconstruction losses. This idea of adaptive loss weighting informs our choice of two-stage refinement, where coarse magnitude masking is followed by fine phase-aware optimization.

Other notable contributions in SE research include convolutional recurrent networks (CRN) [3], complex spectral mapping methods [8], and attention-based enhancements [9]. Across these studies, a common trend is the combination of local feature extraction via convolution with global sequence modeling via recurrent or attention modules. This combination provides a strong inductive bias for speech, capturing harmonic structures and temporal dynamics effectively.

Our work builds upon these foundations by integrating SCM-inspired frequency compression with a two-stage enhancement process. Unlike more computationally intensive attention-heavy models, our modification remains lightweight and straightforward to implement, making it feasible for academic experimentation and real-time applications.

III. Proposed Method

A. Overview

The proposed approach modifies the DPCRN baseline using a two-stage enhancement pipeline. First, a spectral magnitude mask (SMM) is estimated from the noisy spectrogram to suppress dominant noise components. Second, the pre-enhanced spectrogram is refined through dual-path RNN processing to reconstruct both the real and imaginary parts of the complex spectrum. To improve low- and mid-frequency representation while reducing redundant high-frequency computations, we integrate a learnable spectral compression mapping (SCM).

B. Learnable Spectral Compression Mapping

Full-band speech signals contain a disproportionate amount of energy in the low- and mid-frequency ranges. Uniform spectral processing wastes capacity on sparse high-frequency components and may hinder learning. Inspired by human auditory perception, the SCM compresses high-frequency bands logarithmically while leaving low frequencies largely unaltered. The mapping is implemented as a partially learnable dense layer: the low-band portion remains fixed, while the high-band portion is trainable. This allows the network to adjust compression dynamically during training, improving harmonics discrimination and noise suppression.

C. Two-Stage Enhancement

The first stage of the model estimates a coarse spectral magnitude mask. This mask enhances the overall signal-to-noise ratio (SNR) and reduces dominant noise, effectively providing a pre-cleaned input for the subsequent stage. The second stage is a dual-path RNN module similar to DPCRN, which processes chunks of the spectrogram along temporal and spectral dimensions. The dual-path structure enables modeling both intra-frame spectral correlations (via bidirectional LSTM) and inter-frame temporal dependencies (via LSTM). Skip connections and convolutional encoding-decoding ensure that low-level spectral details are preserved.

D. Loss Functions

To encourage accurate magnitude and phase reconstruction, we adopt a combination of power-compressed magnitude loss and real-imaginary (RI) loss. The first stage pre-trains the SMM using only magnitude loss, while joint training of both stages uses the sum of magnitude and RI losses. This approach balances coarse noise suppression with fine spectral detail reconstruction. Mathematically, the total loss can be expressed as:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{mag} + \beta \mathcal{L}_{RI} \tag{1}$$

where α and β are weighting factors for magnitude and complex-domain losses.

IV. Implementation Details

We process audio at a 48 kHz sampling rate using 25 ms Hanning windows and 12.5 ms hop size. FFT length is set to 1200, producing 601 frequency bins per frame. SCM compresses these 601 bins into 256-dimensional representations, maintaining the first 64 bins (low frequencies) intact. The DPCRN encoder uses five Conv-2D layers with increasing channels [16, 32, 48, 64, 80] and progressively downsampling kernels in time and frequency. The dual-path module uses 1 intra-chunk BiLSTM and 1 inter-chunk LSTM, both with hidden size 127. Training uses the warmup-based Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and learning rate schedule adapted from transformer-style warmup. The total parameter count is approximately 5.0M.

For dataset preparation, we use DEMAND for noise and clean speech clips from VCTK. Clips are convolved with room impulse responses from openSLR datasets to simulate

reverberation. SNR ranges from 15 dB to -5 dB for both training and evaluation. 8% of data is held out for validation. Audio augmentations include random SNR scaling and minor pitch shifts to improve robustness.

V. Experiments and Results

A. Evaluation Metrics

We evaluate using PESQ and STOI as primary objective metrics. PESQ measures perceptual speech quality, while STOI assesses short-time intelligibility. Additionally, we report DNS-MOS P.835 scores to compare with subjective-quality models on real acoustic data.

B. Baseline Comparison

The table below summarizes the performance of the proposed model compared to the original DPCRN and other baseline methods on the DEMAND test set.

TABLE I
Performance Comparison on DEMAND Dataset

Model	PESQ	STOI (%)	DNS-MOS
DPCRN	2.74	88.1	3.1
Proposed	2.81	89.0	3.3
RNNoise	2.30	82.5	2.5

Qualitative observations indicate that the proposed approach preserves low-frequency harmonics while reducing high-frequency artifacts. The SNR improvement calculation is as below:

$$SNR_{improved} = 10 \log_{10} \frac{\sum s(t)^2}{\sum (s(t) - \hat{s}(t))^2} \tag{2}$$

where $s(t)$ is clean speech and $\hat{s}(t)$ is enhanced speech.

Another simple equation used during loss optimization is the magnitude mask calculation:

$$M(f,t) = \frac{|S(f,t)|}{|S(f,t)| + |N(f,t)|} \tag{3}$$

where $S(f,t)$ and $N(f,t)$ are the STFTs of clean speech and noise.

C. Qualitative Observations

Noise types such as traffic, office chatter, and cafe ambient sounds show improved suppression in low- and mid-frequency bands. The two-stage pipeline effectively removes stationary background noise while preserving the spectral envelope of the speech. Artifacts commonly observed in single-stage magnitude-only enhancement are reduced, particularly in high-frequency consonant sounds.

VI. Discussion

The results indicate that simple, practical modifications to DPCRN can provide meaningful improvements. SCM allows more efficient allocation of model capacity to perceptually important frequency ranges. The two-stage approach reduces noise while enabling phase-aware reconstruction, highlighting the benefit of separating coarse magnitude suppression from fine RI refinement. While attention-based models may achieve higher absolute metrics, our approach maintains a balance between performance and model complexity, making it suitable for real-time systems.

Future improvements may include integrating light-weight attention mechanisms or adaptive convolution kernels, inspired by [5], [6]. Additionally, dynamically weighted losses [7] could further reduce residual artifacts in extreme low-SNR conditions.

VII. Conclusion

We presented a practical modification to the DPCRN baseline for single-channel speech enhancement on the DEMAND dataset. By incorporating learnable spectral compression mapping and a two-stage enhancement pipeline, the model improves perceptual and objective metrics while maintaining low computational overhead. The approach offers a strong foundation for low-complexity, full-band speech enhancement research and can be extended with attention mechanisms, adaptive convolutions, or dynamic loss strategies in future work.

References

[1] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 3, pp. 483–492, 2016.

[3] K. Tan and D. L. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," Interspeech 2018.

[4] X. Le, H. Chen, K.-J. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single-Channel Speech Enhancement," Interspeech 2021.

[5] Y. Wan et al., "Multi-Loss TF-Attention Model for Speech Enhancement," ICASSP 2022.

[6] Y. Wang et al., "Adaptive Convolution for CNN-based Speech Enhancement," arXiv preprint 2021.

[7] A. Peracha et al., "Causal Speech Enhancement with Dynamically Weighted Losses," ICASSP 2021.

[8] K. Tan and D. L. Wang, "Complex Spectral Mapping with a Convolutional Recurrent Network for Monaural Speech Enhancement," ICASSP 2019.

[9] C. Subakan et al., "Attention Is All You Need in Speech Separation," ICASSP 2021.