IEICE
TRANSACTIONS

on Fundamentals of Electronics,
Communications
and Computer Sciences

DOI:10.1587/transfun.2024EAP1173

Publicized:2025/01/31

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Monaural Speech Enhancement with Full-Convolution Attention Module and Post-Processing Strategy

Lin ZHOU[†a)], ***member***, **Yanxiang CAO**[†], **Qirui WANG**[†], **Yunling CHENG**[†], **Chenghao ZHUANG**[†], **and Yuxi DENG**[†], ***Nonmembers***

SUMMARY The performance of phase-aware speech enhancement has improved dramatically in recent years. Combined with complex convolutions, deep complex U-Net and deep complex convolution recurrent network (DCCRN) have achieved superior performance in monaural phase-aware speech enhancement. However, these methods optimize the models with loss only in the time domain and ignore the global correlations along the frequency axis that capture the harmonic information between frequency bands. Also, the algorithms based on self-attention exhibit high computational complexity. To strike the balance between performance and computational cost, we propose a new monaural phase-aware method in the time-frequency domain on the deep complex U-Net structure. Specifically, this proposed method incorporates a dual-path recurrent neural network (DPRNN) block in the bottleneck to model both frequency-domain correlation and time-domain correlation. Additionally, attention modules are implemented between the complex encoder and decoder layers. This introduces a more effective way of enhancing the representation of the model, rather than directly concatenating their outputs. Finally, a post-processing module is introduced to mitigate the over-suppression of speech and residual noise. We conduct ablation studies to validate the effectiveness of the dual-path method and the post-processing module. Also, compared to several recent speech enhancement models, the proposed algorithm demonstrates remarkable improvements in terms of objective metrics.

key words: speech enhancement, phase-aware, deep learning, time-frequency domain

1. Introduction

Speech often has poor intelligibility and perceptual quality due to additive background noise and other interference. Monaural speech enhancement aims to separate clean speech from background interference, with applications ranging from front-end modules in automatic speech recognition (ASR) systems to hearing aids [1].

With the development of deep learning, deep neural network (DNN) based speech enhancement has achieved significant performance improvement over conventional signal processing-based methods. According to the signal domain, existing DNN-based speech enhancement methods can be classified into time domain [2,3,4] and time-frequency (TF) domain methods [5,6,7]. To some extent, time-domain methods overcome the drawbacks of conventional TF domain methods, which use the noisy phase to reconstruct the time-domain waveform of estimated speech. However, the time-domain methods ignore the advantages that in TF representations, speech has certain structures and noise is easier to separate from noisy input. And by learning the complex spectrogram or the complex masks, the TF domain methods can also take the phase information into account.

Due to the above advantages, much research has been done in the TF domain. Early TF mask-based methods recover the target speech using the estimated magnitude and noisy phase [8,9,10]. It was found that using even clean magnitude with noisy phase, the reconstructed time-domain speech would still have a large distortion [11]. This means that phase has a significant impact on the quality of the reconstructed speech, thus introducing the study of both phase and magnitude estimation. Complex spectrogram mapping [12,13] is an attempt to incorporate phase estimation. However, using a mask to incorporate phase estimation achieves better performance [14], and mask-based methods are gradually becoming mainstream due to their fast convergence and finite dynamic range [15]. Unlike previous masks, a phase-sensitive mask (PSM) is one of the first attempts to incorporate phase information. PSM compensates to some extent for the distortion caused by using the noisy phase. Subsequently, a complex ideal ratio mask (cIRM) [16] is proposed to enhance both the complex noisy spectrogram's real and imaginary components, thus implicitly estimating the phase information. Theoretically, cIRM can accurately recover the complex TF spectrogram.

To deal with complex spectrograms and thus to incorporate phase estimation, deep complex U-Net (DCU-Net) [11] combines deep complex network and U-Net for phase-aware speech enhancement. By introducing complex convolution into the convolutional recurrent network (CRN), deep complex convolutional recurrent network (DCCRN) [17] also attempts to replace the traditional long short-term memory (LSTM) with complex LSTM. By optimizing the scale-invariant source-to-noise ratio (SI-SNR) loss and estimating the cIRM, DCCRN won first place in the real-time track of the INTERSPEECH2020 Deep Noise Suppression (DNS) challenge. Funnel deep complex U-Net (FDCU) [18] applies complex convolution in one encoder and two decoders, where one decoder estimates magnitude by ideal ratio mask (IRM) and another decoder estimates phase by mapping. Although the above methods improve the

†The authors are all with the School of Information Science and Engineering, Southeast University, Nanjing, China.

This work is supported by the Key Research and Development Program of Jiangsu Province under Grant BE2022059-3, and the National Key Research and Development Program of China under Grant 2020YFC2004003 and Grant 2020YFC2004002.

a) E-mail: linzhou@seu.edu.cn (Corresponding author)

performance by a large margin, the long-term correlations along the frequency axis are ignored and the models only optimize the time-domain loss. As Yin [19] et.al. point out that the harmonic correlations along the frequency axis are significant for speech enhancement, and the conventional convolutional neural network (CNN) kernels can't capture the global correlations in spectrograms. Therefore, it is an intuitive idea to include a specific module that models the global correlations along the frequency axis.

In recent years, dual-path methods have shown excellent performance in speech signal processing. To model long sequences more effectively, the dual-path recurrent neural network (DPRNN) structure has been proposed for speech separation in the time domain [20]. Under the framework of DPRNN, a dual-path convolutional recurrent network (DPCRN) [21] has been proposed for speech enhancement in the TF domain. The dual-path block is capable of modeling the global dependencies along the frequency axis within a single frame and the long-term dependencies along the time dimension between frames. [22] also extends the dual-path method to transformers in the TF domain, processing features along the time and frequency paths alternately. Given n the sequence length and d the representation dimension, the complexity of self-attention is $O(n^2 \cdot d)$ while the complexity of RNN is $O(n \cdot d^2)$. Using the dual-path method, the representation dimension d is small compared to the sequence length n. Thus, the RNN-based dual-path method has lower complexity and is used in our proposed model. Also, the dual-path methods enable the model to implement sub-band processing and full-band processing on TF representations, simultaneously modeling the long-range dependence of the time and frequency dimensions. The outstanding performance demonstrated the validity of the dual-path methods with strong modeling capability.

Attention mechanisms are widely used due to the ability to allocate limited computational resources to important features and their powerful modeling capabilities. Inspired by the squeeze-and-excitation module [23], a frequency dimension adaptive attention module [24] is proposed that uses global averaging pooling to access global information in the frequency dimension and then uses full connection layers to generate attention masks for frequency bins. In [25], a time-frequency attention module takes into account the energy distribution of speech in TF representations to accurately predict masks or spectrograms. [26] uses a module that focuses on cross-channel and spatial information of TF representations in complex convolution-based methods, and this module can be integrated into any complex-valued network.

Motivated by previous work, we propose a new mask-based phase-aware speech enhancement method in the TF domain. The main contributions are summarized as follows:

- We embed a dual-path RNN block in the bottleneck between the complex encoder and decoder to capture long-term correlations in both time and frequency dimensions.
- To improve the representational power of the model, we use attention-based modules instead of direct skip connections before feeding to the complex decoder layers.
- We propose a post-processing module to further suppress the residual noise or to repair the over-suppressed speech information within the TF bins, which applies complex convolution to enable the information interaction between the estimated spectrograms of speech and noise.

The rest of this paper is organized as follows. Section 2 provides an overview of the proposed speech enhancement framework and describes each module of the framework in detail. Section 3 describes the experimental setting. Section 4 presents the experimental results and analysis. Finally, Section 5 is the conclusion of the paper followed by references.

2. Proposed Framework

2.1 Problem Formulation

In the time domain, the noisy speech can be represented as a combination of clean speech and additive noise: $y(t) = s(t) + n(t)$, where $y(t)$, $s(t)$, and $n(t)$ refer to the noisy speech, the clean speech, and the noise, respectively. Using the Short-Time Fourier Transform (STFT), the formula can be expressed in the TF domain as:

$$Y(t,f) = S(t,f) + N(t,f) \tag{1}$$

where $Y(t,f)$, $S(t,f)$, and $N(t,f)$ denote the spectrogram of the noisy speech, the clean speech, and the noise at a particular TF bin with frame index t and frequency index f.

To recover the speech spectrogram after the decoder, cIRM in Cartesian coordinates is used, and the ground truth can be computed as follows:

$$\begin{cases} M(t,f) = M_r(t,f) + j \cdot M_i(t,f) \\ M_r(t,f) = \frac{Y_r(t,f)S_r(t,f)+Y_i(t,f)S_i(t,f)}{Y_r^2(t,f)+Y_i^2(t,f)} \\ M_i(t,f) = \frac{Y_r(t,f)S_i(t,f)-Y_i(t,f)S_r(t,f)}{Y_r^2(t,f)+Y_i^2(t,f)} \end{cases} \tag{2}$$

With the mask $M(t,f)$, we can get the clean spectrogram in the form of:

$$S = Y \cdot M = S_r + j \cdot S_i \tag{3}$$

where M denotes the cIRM at a TF bin and subscript $(\cdot)_r$ and $(\cdot)_i$ denote the real and imaginary parts of a complex variable, respectively. Also, j represents the imaginary unit. To simplify the notation, the index (t,f) is omitted if there is no conflict.

2.2 Overall Structure

As shown in Fig. 1, the proposed model is based on the U-Net structure and consists of a dual-path RNN block, attention modules, a post-processing module, a complex encoder and a complex decoder. The complex encoder uses convolutional kernels to model local dependencies and downsamples the input spectrograms in the frequency dimension to create multi-resolution features. Since the convolution-based encoder only models the local dependencies, the dual-path RNN block is used at the bottleneck layer to capture long-term correlations. The attention module between the complex encoder and decoder layers reuses the output of the encoder layer along with the output of the previous decoder layer or DPRNN block to generate the attention mask. The complex decoder upsamples feature maps in the frequency dimension to recover the original resolution, and the final decoder layer outputs the estimated cIRM that is multiplied by the noisy spectrogram to obtain a preliminary estimated speech spectrogram. The post-processing module outputs masks for the preliminary estimated noise spectrogram to further estimate the residual spectrogram, which is then subtracted from the preliminary estimated speech spectrogram to obtain the final estimated speech spectrogram.

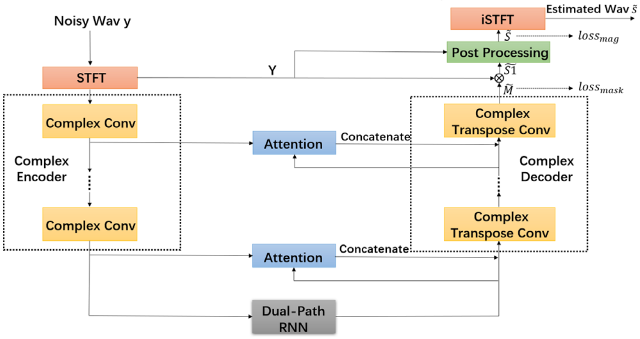


Fig. 1 The specific structure of the proposed model.

2.3 Complex Encoders and Decoders

In this paper, the complex encoder contains the complex Conv2D module, which conducts four traditional convolution operations. Let the input and output feature maps be $V = V_r + j \cdot V_i$ and $U = U_r + j \cdot U_i$, respectively. With the complex convolution filter $W = W_r + j \cdot W_i$, the complex convolution can be expressed as:

$$\begin{cases} U = V \circledast W = U_r + j \cdot U_i \\ U_r = V_r * W_r - V_i * W_i \\ U_i = V_r * W_i + V_i * W_r \end{cases} \qquad (4)$$

where $\circledast$ and $*$ stand for the complex and traditional Conv2d operation, respectively.

Six complex convolution layers are used in the complex encoder, and each complex convolution layer is followed by a batch normalization and parametric rectified linear unit (PReLU) activation. The complex decoder adopts a symmetric structure as the complex encoder, but the difference is that the complex decoder adopts complex transpose convolution layers, and there is no batch normalization and activation function after the last decoder layer because the cIRM takes an infinite range of values.

2.4 Dual-Path RNN Block

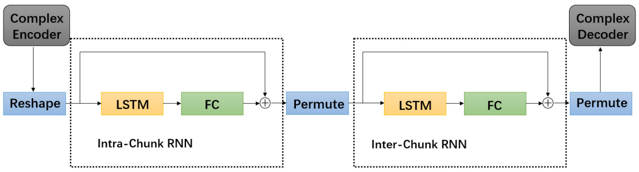


Fig. 2 Structure of the dual-path RNN block.

This block is built with reference to [20] and [21]. In contrast to [20], a variable-length training strategy is used in the experiments, where the audios are zero-padded to the largest length within a batch, and no layer normalization is used in this block. This block contains two parts, intra-chunk RNN, and inter-chunk RNN, as Fig. 2 shows. Similar to DPCRN, a single frame is considered as a chunk, with intra-chunk RNN modeling global correlations along the frequency dimension within a single frame and inter-chunk RNN modeling temporal global correlations at a certain frequency between frames. Thus, with this block, intra-frame spectral patterns and temporal correlations at certain frequency band are efficiently captured simultaneously. Following consecutive downsampling along the frequency dimension in the encoder, the number of frequency bands decreases. Using a bidirectional LSTM to model the correlations of downsampled frequency bands is not appropriate. Consequently, a unidirectional LSTM is used in the intra-chunk RNN to model correlations along the frequency dimension.

Each LSTM layer is followed by a full connection layer to restore feature size and the residual connection is applied to mitigate gradient vanishing. Since the dimensions of the dependencies to be modeled are different, a dimensional rearrangement between the two RNNs is required.

2.5 Attention Module

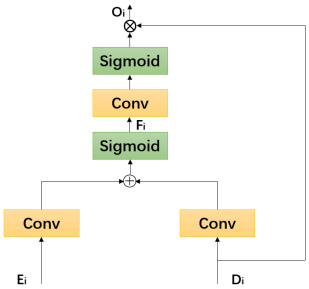


Fig. 3 Structure of the attention modules.

The output feature maps of the encoder layer contain more noise, and concatenating them with feature maps in the decoder layer to some extent interferes with the denoising process and makes the denoising process unstable. Thus, the attention modules are introduced and the structure of the attention modules is shown in Fig. 3 where E_i and D_i denote the output feature maps of the ith complex encoder and decoder layer, respectively.

The gating-based attention modules rely solely on convolutional operations, which has the advantage of processing feature maps causally with stride of 1 and causal padding. And the attention module introduces an attention mechanism in all three dimensions: frequency bin, frame and channel which enables the model to leverage the correlation between these three dimensions and select the most important information based on the attention mechanism.

To reduce the parameters, the number of output channels in the first two convolution layers is reduced to half of the number of input channels and then is restored to the number of input channels by the last convolution. The sigmoid activation is used to create the attention mask, which is multiplied by D_i element by element. In detail, the output of this module can be calculated as:

$$\begin{cases} F_i = \sigma\left(Conv(E_i) + Conv(D_i)\right) \\ O_i = \sigma(Conv(F_i)) \cdot D_i \end{cases} \tag{5}$$

where $Conv(\cdot)$ and $\sigma(\cdot)$ represent the normal 2D convolution operation and the sigmoid activation, F_i and O_i represent the middle feature maps and the output of attention module.

2.6 Post-Processing Module

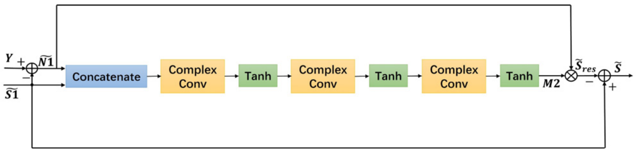


Fig. 4 Structure of the proposed post-processing module.

Since the complex convolution can introduce the information interaction between the real and imaginary features, the post-processing module uses the complex convolution to further mitigate both the residual noise and the speech information over-suppression. As shown in Fig. 4, Y and $\widetilde{S1}$ represent the noisy spectrogram and the preliminary estimated speech spectrogram obtained by cIRM. $\widetilde{N1}$ represents the estimated noise spectrogram obtained by subtracting $\widetilde{S1}$ from Y:

$$\widetilde{N1} = Y - \widetilde{S1} \tag{6}$$

Thus, the PP module has two input signals $\widetilde{S1}$ and $\widetilde{N1}$. Both $\widetilde{S1}$ and $\widetilde{N1}$ are complex spectra representing speech and noise, respectively, which include both real and imaginary parts:

$$\begin{aligned} \widetilde{S1} &= \widetilde{S1}_r + j \cdot \widetilde{S1}_i \\ \widetilde{N1} &= \widetilde{N1}_r + j \cdot \widetilde{N1}_i \end{aligned} \tag{7}$$

After the last hyperbolic tangent activation, PP module output the complex mask $M2$ and estimates residual spectrogram $\tilde{S}_{res}$ based on $M2$ and $\widetilde{N1}$ according to the following formula:

$$\tilde{S}_{res} = M2_r \cdot \widetilde{N1}_r + j \cdot M2_i \cdot \widetilde{N1}_i \tag{8}$$

where $M2_r$ and $M2_i$ represent the real and imaginary parts of a complex mask $M2$ respectively.

In Eq.(8), $M2_r$ applies to the real part $\widetilde{N1}_r$ and $M2_i$ applies to the imaginary part $\widetilde{N1}_i$.

Finally, the estimated residual spectrogram $\tilde{S}_{res}$ is subtracted from $\widetilde{S1}$ to obtain the final enhanced speech spectrogram $\tilde{S}$. This process is described as follows:

$$\begin{aligned} \tilde{S} &= \widetilde{S1} - \tilde{S}_{res} \\ &= \left(\widetilde{S1}_r - M2_r \cdot \widetilde{N1}_r\right) + j \cdot \left(\widetilde{S1}_i - M2_i \cdot \widetilde{N1}_i\right) \end{aligned} \tag{9}$$

Within a specific TF bin in the real or the imaginary spectrogram, the hyperbolic tangent activation enables $\tilde{S}_{res}$ and $\widetilde{N1}$ to have the same or opposite plus/minus sign, the two cases corresponding to $\tilde{S}_{res}$ estimating residual noise or over-suppressed speech information, respectively. Thus, the post-processing module is able to suppress the residual noise and repair the over-suppressed speech information at the same time.

3. Experiments

3.1 Data Generation

The performance of the proposed model was evaluated on the dataset by Valentini et.al [27]. This dataset has been frequently used in speech enhancement studies and was therefore convenient for comparison with other models. This dataset provides pre-simulated pairs of noisy and clean speech, where the clean speech was derived from the Voice Bank corpus and the noise was obtained from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND). The clean speech was read in English by 30 speakers with different accents, including both male and female speakers. Each speaker read approximately 400 utterances, with clean samples from 28 speakers used in training and clean samples from 2 speakers used in testing. The noise from DEMAND included 13 types of noise. 8 types from DEMAND and 2 synthetic types (ssn and bable) were used for training with the signal-to-noise ratio (SNR) set to 0dB, 5dB, 10dB, and 15dB. 5 types were used for testing with SNR set to 2.5dB, 7.5dB, 12.5dB, and 17.5dB. Thus, the final training set consisted of 11572 utterances in 40 different noisy environments, and the test set consisted of 824 utterances in 20 different noisy environments. All waveforms are downsampled to 16 kHz.

3.2 Network Setup

Six complex convolution layers are used in the complex encoder. The kernel size and stride are all set to (5,2) and (2,1) in frequency and time dimension, respectively. The complex decoder uses complex transpose convolution and has the same kernel size and stride as the complex encoder. The number of output channels in the complex convolution layers is [32,64,96,128,192,256], while the number of output channels in the complex transpose convolution layer is [192,128,96,64,32,2]. In the attention modules, the kernel size and stride of the three convolutional layers are all set to (5,2) and (1,1). The hidden LSTM units in the DPRNN block are 128. In the post-processing module, the kernel size of the three complex convolution layers is (5,2) and the stride is (1,1). And the number of output channels is [32,64,2]. Note that all the convolutions are set to be causal. Using the above parameter settings, the total parameters are 3.88M.

3.3 Training Strategy

The short-time Fourier transform with the Hanning window was performed on the noisy speech. The window length and hop size were chosen to be 25ms and 6.25ms, and the FFT length was 512.

We train the model using loss with two terms as shown in Fig.1. $loss_{mask}$ only updates the parameters of the U-Net based structure and $loss_{mag}$ updates the parameters of the whole model. We optimized the model using the mean square error (MSE) loss in the TF domain. Specifically, with the complex decoder outputting $\widetilde{M} = \widetilde{M}_r + j\widetilde{M}_i$ and the post-processing module outputting the final estimated complex speech spectrogram $\tilde{S}$, the loss function was defined as:

$$\begin{cases} loss_{mask} = MSE\left(\widetilde{M}_r,\ M_r\right) + MSE\left(\widetilde{M}_i,\ M_i\right) \\ \qquad loss_{mag} = MSE(|\tilde{S}|^{0.7}, |S|^{0.7}) \\ \qquad loss = loss_{mask} + loss_{mag} \end{cases} \tag{10}$$

Power-law compression is applied to the magnitude spectrogram, which instructs the model to pay more attention to the low-magnitude TF bins.

For model training, the batch size was set to 4. We used the Adam optimizer and the initial learning rate was set to 0.001. Exponential decay was used and the decay rate was set to 0.95. We use PyTorch to train the model with a NVIDIA GeForce RTX 3090

3.4 Objective Evaluation Metric

To evaluate the performance of our proposed algorithm, 4 objective metrics were used: the perceptual evaluation of speech quality (PESQ) [28] and the composite measures for signal distortion (CSIG), noise distortion (CBAK), overall speech quality (COVL) [29]. CSIG, CBAK, and COVL denote the mean opinion score prediction of the signal distortion, the intrusiveness of background noise, and the overall effect, respectively, with CSIG focusing on the speech signal only.

4. Results and Discussion

4.1 Ablation Study

To validate the effectiveness of the proposed method, we performed an ablation study. The baseline model is defined only with a complex encoder, a complex decoder, and a dual-path RNN block between them to model long-term correlation in both the time and frequency dimensions. The proposed model adds the attention modules and the post-processing module to the baseline. Table 1 shows the results of the ablation study. Note that all models in Table 1 used the same loss function and the same train strategy as presented in 3.3.

Table 1 Results of ablation study with respect to different modules.

Model	Param.(M)	PESQ	CSIG	CBAK	COVL
Baseline	1.88	2.93	4.02	3.42	3.47
Baseline with attention modules	3.87	3.00	4.01	3.47	3.50
Baseline with PP module	1.88	2.94	4.18	3.44	3.57
The proposed model	3.88	3.01	4.22	3.49	3.62

Compared to the baseline, the addition of attention modules slightly decreased CSIG by 0.01 and improved PESQ, CBAK, and COVL by 0.07, 0.05, and 0.03, respectively. The baseline model directly reused the output of the complex encoder layer through the skip connection. With attention modules, the output of the complex encoder layer was used only to generate the attention map, significantly improving the model's representation power. This can also be demonstrated by comparing baseline with post-processing (PP) module and the proposed model, where adding the attention modules improved PESQ, CSIG, CBAK, and COVL by 0.07, 0.04, 0.05, and 0.05, respectively.

Also, the addition of post-processing (PP) module to the baseline improved CISG and COVL by 0.16 and 0.1 with negligible parameter increase, leaving PESQ and CBAK slightly increased. Compared to baseline with attention modules, the proposed model can further improve CSIG and COVL by 0.21 and 0.12, which achieved the best performance. The effectiveness of the post-processing module will be further demonstrated in 4.2.

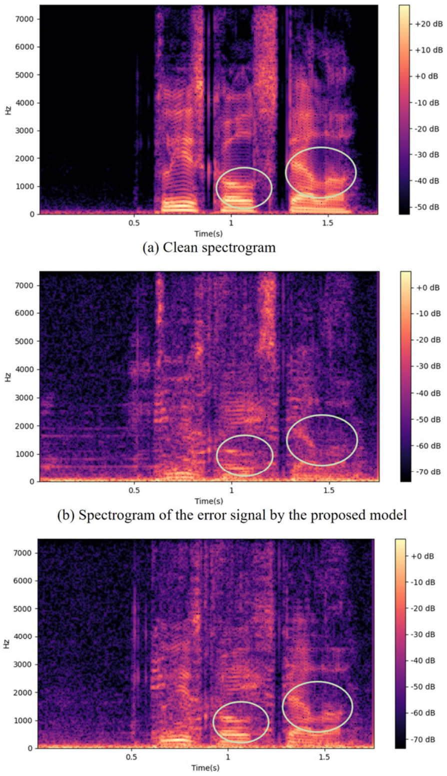
4.2 Effectiveness of Using the Post-Processing Module

Table 2 Objective metrics of the proposed model and baseline with attention modules according to SNR settings.

	SNR(dB)	PESQ	CSIG	CBAK	COVL
the proposed model	2.5	2.45	3.69	3.02	3.05
	7.5	2.89	4.15	3.39	3.52
	12.5	3.24	4.47	3.68	3.87
	17.5	3.55	4.69	3.96	4.16
baseline with attention modules	2.5	2.46	3.50	3.00	2.96
	7.5	2.91	3.94	3.39	3.42
	12.5	3.22	4.27	3.65	3.75
	17.5	3.52	4.56	3.91	4.06

To further explore the effectiveness of the post-processing module, we summarized details for different SNR levels in Table 2. Adding the post-processing module had little effect on the CBAK. For the 2.5dB and 7.5dB test samples, after the addition of the post-processing module, the PESQ decreased by 0.01 and 0.02, but the CSIG increased significantly by 0.19 and 0.21 and the COVL increased significantly by 0.09 and 0.1. For the 12.5dB and 17.5dB test samples, the PESQ increased by 0.02 and 0.03 after the addition of the post-processing module. The CSIG increased significantly by 0.2 and 0.13, and the COVL increased significantly by 0.12 and 0.1.

The error signal can be obtained by subtracting the denoised speech from the corresponding clean speech. To demonstrate the effectiveness of the PP module in a more intuitive way, Fig.5 shows the spectrograms of the clean speech and the error signal by the two models. The harmonics in the spectrogram are important structural information of the speech signal in TF domain. According to the local spectrogram labelled in Fig.5, the spectrogram of the error signal obtained by baseline with attention modules (the proposed model without PP module) has a greater number of harmonic structures compared to the addition of the PP module.



(a) Clean spectrogram

(b) Spectrogram of the error signal by the proposed model

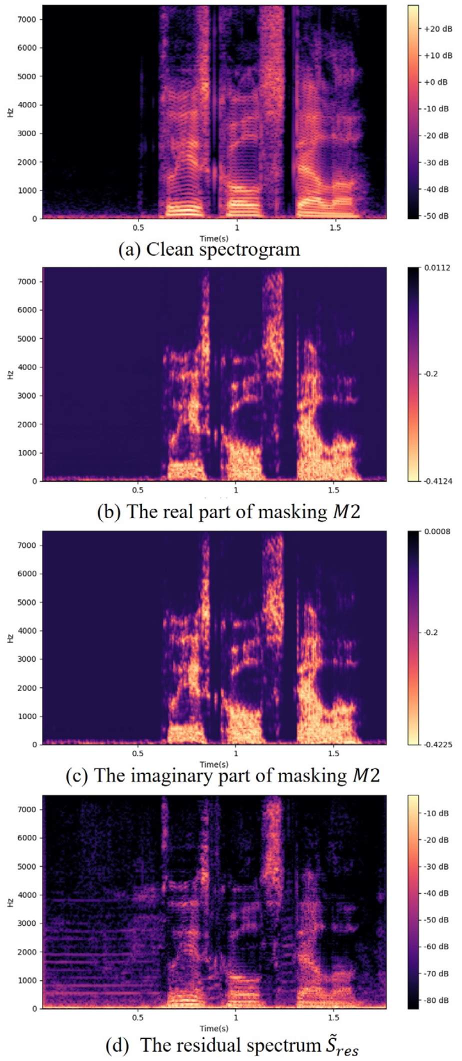
(c) Spectrogram of the error signal by baseline with attention modules

Fig. 5 Spectrograms of p232_001.wav.

Fig.5 compares the error signal spectrogram of the proposed algorithm with PP module and the baseline with attention modules, confirming the effectiveness of the PP module. Additionally, Fig.6 illustrates the real and imaginary parts of the mask $M2$ output by the PP module, along with the residual spectrum $\tilde{S}_{res}$. Fig.6 is utilized to further analyze the characterization and correlation of the PP module's output mask $M2$ and residual spectrum $\tilde{S}_{res}$.

According to Eq.(8) and Eq.(9), if the mask $M2$(subscript r or i is omitted) is negative, it indicates that $\tilde{S}_{res}$ represents over-suppressed speech at a certain TF bin of the real or imaginary spectrogram. Conversely, if $M2$ is positive, $\tilde{S}_{res}$ is identified as the residual noise. From Fig.6, it can be seen that regions with negative $M2$ correspond to $\tilde{S}_{res}$ containing harmonic-like speech components, whereas regions with positive $M2$ values indicate $\tilde{S}_{res}$ as residual noise. This suggests that the results in Fig.6 are consistent with the analysis in Section 2.6.

Moreover, from Fig.6, it is observed that the range of negative $M2$ is larger than the range of positive $M2$ in the clean speech duration. At the same time, most of the positive $M2$ have a smaller magnitude compared to the absolute value of negative $M2$. These indicate that during the cIRM masking process, the signal distortion mainly due to the speech over-suppression rather than the residual noise. Thus, the PP module repairs the over-suppressed speech information and mitigates the signal distortion so that CSIG and COVL improve significantly while CBAK improves slightly. This observation is consistent with the results in Table 2.



(a) Clean spectrogram

(b) The real part of masking $M2$

(c) The imaginary part of masking $M2$

(d) The residual spectrum $\tilde{S}_{res}$

Fig. 6 The output of PP module

In summary, the PP module has the ability to mitigate both the residual noise and the speech information over-suppression. The result indicates that without PP module, the model tends to over-suppress the speech information in the spectrogram and the PP module repairs the over-suppressed speech information so that CISG and COVL improve significantly while CBAK improves slightly. With negligible parameter increase, the post-processing module significantly improves the CSIG and COVL at all SNR levels, indicating that the post-processing module repairs over-suppressed speech information in the proposed model. The results are in line with the original purpose of adding the post-processing module, i.e., to further reduce speech distortion and improve speech quality.

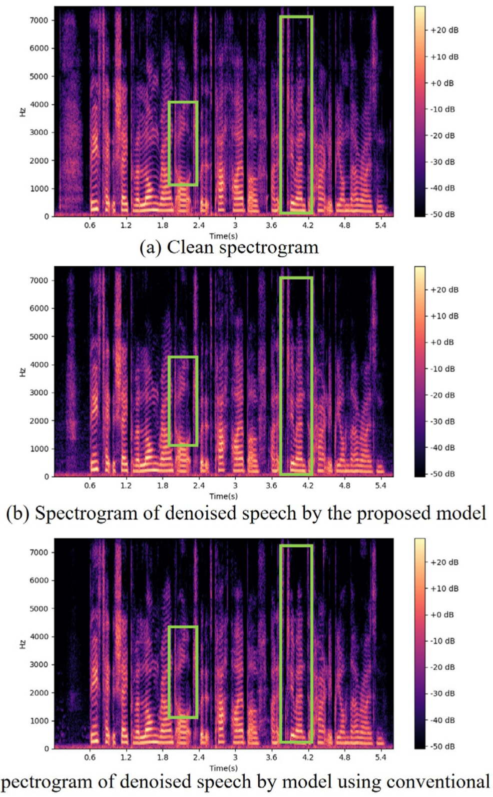
4.3 Effectiveness of Dual-Path RNN Method

We further demonstrate the effectiveness of the dual-path method by replacing the dual-path RNN block with conventional LSTM in the proposed model. Specifically, in model using conventional LSTM, an LSTM layer with 128 hidden units and a full connection layer are used instead of the dual-path RNN block. The proposed model is shown in Fig.1 and contains the attention modules, the post-processing module, the dual-path RNN block (use dual-path method), the complex encoders and the complex decoders. Different from the proposed model, model using conventional LSTM uses a two-layer LSTM and a dense layer in the bottleneck rather than the dual-path RNN block. The results are shown in Table 3. Compared to model using conventional LSTM, applying the dual-path method improves PESQ, CSIG, CBAK, and COVL by 0.1, 0.03, 0.08, and 0.06, respectively.

Table 3 Objective metrics comparison between the proposed model and model using conventional LSTM.

Model	Param.(M)	PESQ	CSIG	CBAK	COVL
Model using conventional LSTM	4.27	2.91	4.19	3.41	3.56
The proposed model	3.88	3.01	4.22	3.49	3.62

Fig.7 shows the spectrograms of the clean speech and the denoised speech by the two models. Compared to spectrogram by model using conventional LSTM, the proposed model is able to recover clearer harmonic structures due to the ability to capture long-range correlations along the frequency dimension.



(a) Clean spectrogram

(b) Spectrogram of denoised speech by the proposed model

(c) Spectrogram of denoised speech by model using conventional LSTM

Fig. 7 Spectrograms of p257_008.wav.

In the conventional LSTM, the input at each moment is the feature of a single frame for all channels at all frequencies. In the dual-path method, the intra-chunk RNN processes the feature of each frame in parallel and browses the frequency bands to model the spectral pattern within the frame. And the inter-chunk RNN processes the feature of each frequency in parallel to model the temporal correlation at a given frequency. With lower complexity, the dual-path method is able to model the harmonic correlation of spectrograms and has better speech enhancement performance.

4.4 Performance Comparison with Other Advanced Methods

The proposed model is compared with other methods, either in the time domain or complex TF domain. Among them, DEMUCS[30] and SADNUnet[31] operate in the time domain and others in the TF domain with PHASEN[19], DCCRN-E[17], MPCRN[32] in the polar coordinate and DCUnet[11], DCCRN+[15] in the Cartesian coordinate. Similar to our proposed method, DCUnet, DCCRN and DCCRN+ are models based on complex convolution and DCCRN+ and MPCRN are models based on dual-path RNN. Self-attention based methods[22] often have higher computational complexity compared to convolution recurrent network based methods. Due to higher complexity, self-attention based dual-path methods are excluded from the comparison. The results are shown in Table 5, which also summarizes the parameters, the FLOPs and the causality of the model.

Table 4 Objective metrics comparison for different models. Higher scores indicate better performance, with bold text indicating the best performance for each metric.

Model	Causal	FLOPS	Param.(M)	PESQ	CSIG	CBAK	COVL
Noisy	-	-	-	1.97	3.35	2.44	2.63
DCU-10	×		1.4	2.67	3.72	3.56	3.19
DCU-16	×		2.3	2.93	4.07	**3.75**	3.48
PHASEN	×	206G	8.76	2.99	4.21	3.55	**3.62**
DEMUCS	√	77.8G	18.87	2.93	**4.22**	3.25	3.52
DCCRN-E	√	25.2G	3.7	2.73	3.73	3.22	3.22
DCCRN+	√	-	3.3	2.84	-	-	-
SADNUnet	√	-	2.63	2.82	4.18	3.47	3.51
MPCRN	√	-	2.09	2.96	4.16	3.50	3.56
Proposed	√	35G	3.88	**3.01**	**4.22**	3.49	**3.62**

From Table 4, our proposed model outperforms the other models in almost all metrics except CBAK. Specifically, the proposed model outperforms other causal models in terms of all objective metrics. Compared to DCCRN-E, with similar parameters, our model achieved performance gains of 0.28, 0.49, 0.27, and 0.4 in terms of PESQ, CSIG, CBAK, and COVL, respectively. Both based on complex convolutions, the dual-path RNN block and attention modules along with our proposed loss function significantly improve PESQ and the post-processing module improves CSIG and COVL by a large margin. Among the non-causal models, using less than half of the parameters of PHASEN, our model reached a comparable performance. PHASEN interacts between the amplitude data and the phase data to improve the accuracy of amplitude and phase estimation. Information interaction between real and imaginary features is enabled at each complex convolution, resulting in fewer parameters and superior performance.

5. Conclusion

In this paper, complex dual-path convolution recurrent network for phase-aware speech enhancement is proposed. First, based on the complex encoder and decoder, we use a dual-path RNN block at the bottleneck layer to model long-term correlations along both time and frequency dimensions, which shows more efficient performance than using conventional RNN. Second, the addition of the attention modules to the skip connection improves the ability of the model to represent features and suppresses the flow of high-noise features into the decoder layer, thus significantly improving PESQ. Finally, we propose a post-processing module to resolve the underestimation and overestimation of noise, which can significantly increase the CSIG and COVL of the test speech in all SNR levels.

Acknowledgments

This work is supported by the Key Research and Development Program of Jiangsu Province under Grant BE2022059-3, and the National Key Research and Development Program of China under Grant 2020YFC2004003 and Grant 2020YFC2004002

References

[1] J. Abdulbaqi, Y. Gu, SH. Chen, I. Marsic, "Residual recurrent neural network for speech enhancement," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6659-6663, 2020.

[2] A. Pandey, DL. Wang, "TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6875-6879, 2019.

[3] S. Pascual, A. Bonafonte, J. Serra, "SEGAN: speech enhancement generative adversarial network," 18th Annual Conference of The International Speech Communication Association (INTERSPEECH), pp. 3642-3646, 2017.

[4] D. Rethage, J. Pons, X. Serra. "A wavenet for speech denoising," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5069-5073, 2018.

[5] Y. Zhao, DL. Wang, I. Merks I, T. Zhang, "DNN-based enhancement of noisy and reverberant speech," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6525-6529, 2016.

[6] K. Tan, DL. Wang, "A convolutional recurrent neural network for real-time speech enhancement," 19th Annual Conference of The International Speech Communication Association (INTERSPEECH), pp. 3229-3233, 2018.

[7] C. Zheng, XL. Peng, Y. Zhang, S. Srinivasan, Y. Lu, "Interactive speech and noise modeling for speech enhancement," 35th AAAI Conference on Artificial Intelligence, Electronic Network, vol.35, pp. 14549-14557, 2021.

[8] Y. Xu, J. Du, L.R. Dai, CH. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1) (2015)7-19.

[9] L. Zhang, GZ. Bao, J. Zhang, ZF. Ye, "Supervised single-channel speech enhancement using ratio mask with joint dictionary learning," Speech Communication, 82(2016)38-52.

[10] YH. Tu, J. Du, CH. Lee, "DNN training based on classic gain function for single-channel speech enhancement and recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.910-914, 2019.

[11] HS. Choi, JH. Kim, J. Huh, A. Kim, JW. Ha, K. Lee, "Phase-aware speech enhancement with deep complex u-net," International Conference on Learning Representations(ICLR),pp.1-20, 2019.

[12] K. Tan, DL. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6865-6869, 2019.

[13] K. Tan, DL. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28(2020)380-390.

[14] YX. Wang, A. Narayanan, DL. Wang, "On training targets for supervised speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12)(2014)1849-1858.

[15] SB. Lv, YX. Hu, SM. Zhang, L. Xie, "DCCRN+: channel-wise subband DCCRN with SNR estimation for speech enhancement," 22th Annual Conference of The International Speech Communication Association (INTERSPEECH), pp. 2816–2820, 2021.

[16] DS. Williamson, YX. Wang, DL. Wang, "Complex ratio masking for monaural speech separation," IEEE/ACM Transactions on Audio Speech and Language Processing, 24(3)(2016) 483-492.

[17] YX. Hu, Y. Liu, SB. Lv, MT. Xing, SM. Zhang, YH. Fu, J. Wu, BH.

Zhang, L. Xie, "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," 21th Annual Conference of The International Speech Communication Association (INTERSPEECH), pp. 2472–2476, 2020.

[18] YH. Sun, LJ. Yang, HF. Zhu, J. Han, "Funnel deep complex u-net for phase-aware speech enhancement," 22th Annual Conference of The International Speech Communication Association (INTERSPEECH), pp.161-165, 2021.

[19] DC. Yin, C. Luo, ZW. Xiong, WJ. Zeng, "PHASEN: a phase-and-harmonics-aware speech enhancement network," 34th AAAI Conference on Artificial Intelligence (AAAI), pp. 9459- 9465, 2020.

[20] Y. Luo, Z. Ghen, T. Yoshioka, "Dual-Path RNN: efficient long sequence modeling for time-domain single-channel speech separation," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 46-50, 2020.

[21] XH. Le, HS. Chen, K. Chen, J. Lu, "DPCRN: dual-path convolution recurrent network for single channel speech enhancement," 22th Annual Conference of The International Speech Communication Association (INTERSPEECH), pp. 2811-2815, 2021.

[22] DANG, Feng; CHEN, Hangting; ZHANG, Pengyuan. "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6857-6861, 2022.

[23] J. Hu, L. Shen, S. Albanie, G. Sun, EH. Wu, "Squeeze-and-excitation networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8)(2020)2011-2023.

[24] QQ. Zhang, Q. Song, A. Nicolson, T. Lan, HZ. Li, "Temporal convolutional network with frequency dimension adaptive attention for speech enhancement," 22th Annual Conference of The International Speech Communication Association (INTERSPEECH), pp. 166-170, 2021.

[25] QQ. Zhang, Q. Song, ZH. Ni, A. Nicolson, HZ. Li, "Time-frequency attention for monaural speech enhancement," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7852-7856, 2022.

[26] SK. Zhao, TH. Nguyen, B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6648-6652, 2021.

[27] C. Valentini-Botinhao, W. Xin, S. Takaki S, J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," 9th ISCA Speech Synthesis Workshop, pp.160-165, 2016.

[28] AW. Rix, JG. Beerends, MP. Hollier, AP. Hekstra, "Perceptual evaluation of speech quality (PESQ): a new method for speech quality assessment of telephone networks and codecs," IEEE International Conference On Acoustics, Speech, And Signal Processing(ICASSP), pp. 749-752, 2001.

[29] Y. Hu, PC. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," IEEE Trans Audio Speech Language Process, 16(1) (2008):229-238.

[30] A. Defossez, G. Synnaeve, Y. Adi, "Real time speech enhancement in the waveform domain," 21th Annual Conference of The International Speech Communication Association (INTERSPEECH), pp. 3291–3295, 2020.

[31] XX. Xiang, XJ. Zhang, HZ. Chen, "A nested u-net with self-attention and dense connectivity for monaural speech enhancement," IEEE Signal Processing Letters, 29(2022)105-109.

[32] ZHANG, Yuewei; ZOU, Huanbin; ZHU, Jie. "Magnitude-and-phase-aware Speech Enhancement with Parallel Sequence Modeling," IEEE Workshop on Automatic Speech Recognition and Understanding(ASRU), pp.1-8, 2023.

Lin Zhou received the B.E. degree in information engineering and the Ph.D. degree in signal and information processing from Southeast University, Nanjing, China, in 2000 and 2005, respectively. She is currently an Associate Professor with the School of Information Science and Engineering, Southeast University. Her research interests include speech signal processing and deep learning.

Yanxiang Cao received the Bachelor's degree from Qingdao University, China, in 2021. Currently, he is a graduate student in the School of Information and Engineering, Southeast University, China.

Qirui Wang received the Bachelor's degree from Nanjing University of Science and Technology, China, in 2021. Currently, he is a graduate student in the School of Information and Engineering, Southeast University, China.

Yunling Cheng received the Bachelor's degree from Nanjing University of Science and Technology, China, in 2021. Currently, she is a graduate student in the School of Information and Engineering, Southeast University, China.

Chenghao Zhuang received the Bachelor's degree from Nanjing University of Posts and Telecommunications, China, in 2021. Currently, he is a graduate student in the School of Information and Engineering, Southeast University, China.

Yuxi Deng received the Bachelor's degree from Harbin Institute of Technology, China, in 2021. Currently, she is a graduate student in the School of Information and Engineering, Southeast University, China.