INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

PCNN: A Lightweight Parallel Conformer Neural Network for Efficient Monaural Speech Enhancement

Xinmeng Xu[1], *Weiping Tu*[1,2,3,*] , *Yuhong Yang*[1,3]

[1]NERCMS, School of Computer Science, Wuhan University, China
[2]Hubei Luojia Laboratory, China
[3]Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China

{xuxinmeng, tuweiping, yangyuhong}@whu.edu.cn

Abstract

Convolutional neural networks (CNN) and Transformer have wildly succeeded in multimedia applications. However, more effort needs to be made to harmonize these two architectures effectively to satisfy speech enhancement. This paper aims to unify these two architectures and presents a Parallel Conformer for speech enhancement. In particular, the CNN and the self-attention (SA) in the Transformer are fully exploited for local format patterns and global structure representations. Based on the small receptive field size of CNN and the high computational complexity of SA, we specially designed a multi-branch dilated convolution (MBDC) and a self-channel-time-frequency attention (Self-CTFA) module. MBDC contains three convolutional layers with different dilation rates for the feature from local to non-local processing. Experimental results show that our method performs better than state-of-the-art methods in most evaluation criteria while maintaining the lowest model parameters.

Index Terms: speech enhancement, self-attention, convolutional neural network, parallel conformer

1. Introduction

Speech enhancement (SE) aims to estimate target speech from a noisy recording, which may consist of ambient noise, interfering speech, and room reverberation. It is a pre-processor for many speech processing applications, such as speech recognition [1], speaker verification [2], and hearing aids design [3]. With the recent advances in supervised learning, deep neural networks (DNNs) are applied to several SE models. Typically, DNN-based SE models operate in the short-time Fourier transform (STFT) domain and estimate the clean target speech from the noisy signal via direct spectral mapping [4, 5], or time-frequency (TF) masking [6, 7].

Convolutional neural network (CNN) represents a successful backbone network architecture, which performs the filter processing on speech frames in parallel. It thus is structurally well-suited to focus on local patterns, such as harmonic structures [8]. Meanwhile, CNN captures the contextual information by stacking multiple layers. While these properties bring efficiency and generalization to CNNs, they also cause two main issues. Firstly, the convolutional operation has a limited receptive field. Secondly, the convolution filters have static weights at inference. The former thus prevents the network from capturing the long-range feature dependencies [9, 10] while the latter sacrifices the adaptability to the input contents. As a result, it needs to meet the requirement in modeling the global noise distribution and generates results with noticeable noise residue.

Self-attention (SA) calculates response at a given feature region by a weighted sum of all other positions [11, 12, 13]. Benefiting from the advantage of global processing, SA achieves a significant performance boost over CNNs in SE tasks by mitigating their shortcomings, i.e., limited receptive field and inadaptability to input content [14, 15]. However, due to the global calculation of SA, its computation complexity grows quadratically with the spatial resolution, making it infeasible to fulfill the real-time demanding of SE systems. In addition, global relationships between these speech features are prone to bias and unreliable because feature regions are usually noisy [5]. In this way, calculating the self-similarity of features between the target speech and the global mixture may not be a practical option.

Inspired by the superior performance of CNN in extracting speech local format patterns and the effectiveness of Transformer in capturing the long-range dependency, we propose the parallel Conformer neural network (PCNN) for monaural speech enhancement. The proposed architecture incorporates CNN and Transformer in a parallel manner. It is followed by a hybrid fusion block containing depth-wise separable convolutions and channel attention for an adaptively and learnable performance trade-off. In addition, to deal with the small receptive field of CNN and the high computational complexity of the Transformer, we specially designed a multi-branch dilated convolution (MBDC) and a self-channel-time-frequency attention (Self-CTFA) module. In particular, the MBDC applies channel-wise attention to different dilation rates of convolutions to enlarge the size of the receptive field of local operation, in which channel-wise attention is independently performed on these three outputs for flexibly achieving feature processing from local to non-local. The self-CTFA module consists of three parallel attention branches, i.e., channel-dimension, time-dimension, and frequency-dimension, in which three 2D attention maps are calculated by three 1D energy distributions of these dimensions.

2. Model Description

In this section, we elaborate on our proposed PCNN, which enables adaptive and learnable adjustment of contribution between CNN and Transformer in the SE tasks.

2.1. Overview

We propose a parallel conformer neural network (PCNN) for SE in the time domain. As shown in Figure 1, the architecture consists of a segmentation operation, encoder, separator, masking module, decoder, and overlap-add operation.

The input of PCNN is a raw speech waveform mixture, $\mathbf{x} \in \mathbb{R}^{1\times N}$, which is firstly split into F overlapped frames of length

[*] Corresponding Author.

10.21437/Interspeech.2023-1376

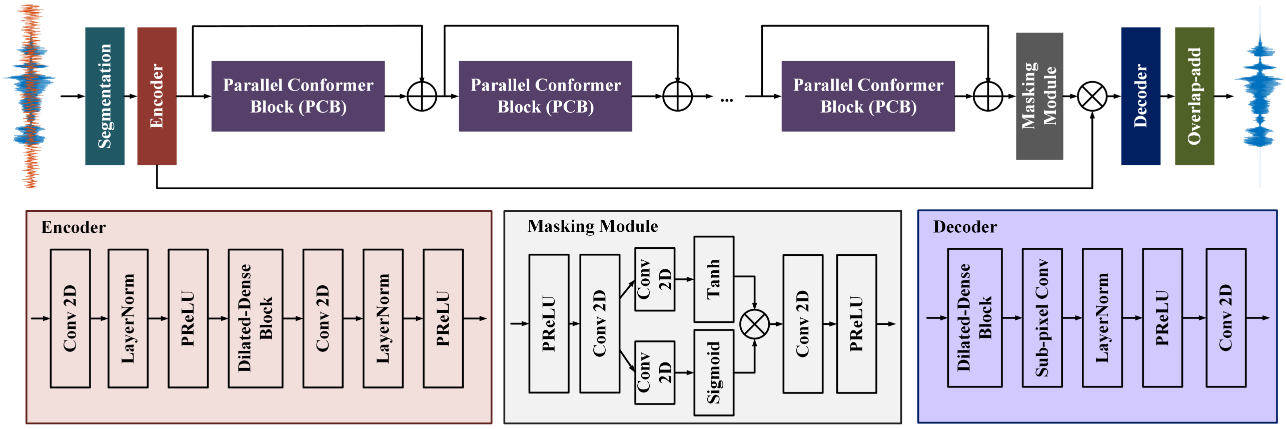

Figure 1: *Overview of the parallel conformer neural network (PCNN).*

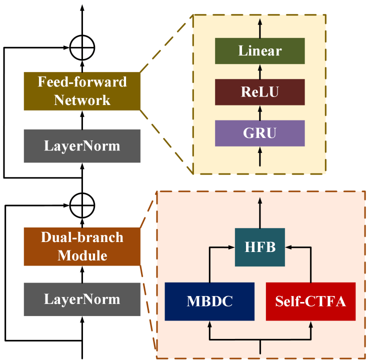

Figure 2: *The architecture of parallel conformer block (PCB). "MBDC" denotes the multi-branch dilation convolution, "Self-CTFA" denotes self channel-time-frequency attention, and "HFB" denotes hybrid fusion block.*

L with a shifting size S by **segmentation operation**. In this way, $\mathbf{x}$ is resulted into a 3-dimensional tensor $X \in \mathbb{R}^{1\times F\times L}$, which is be expressed as

$$F = \lceil (M - L)/(L - S) + 1 \rceil, \tag{1}$$

where N represents the length of the input speech mixture and $\lceil \cdot \rceil$ rounds the number involved up to the nearest integer. In addition, the **overlap-add operation** in an inverse of **segmentation operation** to merge the frames for the enhanced speech waveform reconstruction.

Encoder plays the role of feature extractor [16, 17] and contains two convolutional layers, of which the first one is increasing the number of channels to 64 using convolution with a kernel size of $(1, 1)$ and the second one halves the dimension of frame size using a kernel size of $(1, 3)$ with a stride of $(1, 2)$, in which a dilated-dense block [18] by using four dilated convolutions collaborates between them. In addition, layer normalization and PReLU [19] are adopted after these convolutional layers. On the contrary, **decoder** is responsible for the feature reconstruction and contains a dilated-dense block, and a sub-pixel convolution [20], where followed by a layer normalization, PReLU, and a 2D convolutional layer with a kernel size of $(1, 1)$ for the channel dimension recovery of enhanced speech feature into 1.

The **separator** is the crucial part of PCNN and is mainly composed of several parallel conformer blocks (PCBs) that are cascaded together. The PCB, as shown in Figure 2, include a dual-branch module containing an MBDC, Self-CTFA module, and HFB for local and global extracting and leveraging, and a feed-forward network, both preceded by layer normalization steps and with skip connections, which is specially described in Sec 2.2. Unlike conventional Transformer blocks, the feed-forward network consists of a gated recurrent unit (GRU) layer to learn the positional information [21, 22]. The **masking module** utilizes the feature output from **separator** to generate a mask to enhance speech. Concretely, the production from **separator** is doubled along the channel dimension with PReLU and convolution for matching the output of the encoder that then passes through a gated convolution operation [23] and ReLU to get the mask. The element-wise multiplication between the mask and the output of the encoder obtains the final masked encoder feature.

Two loss functions are used in our study. One is the frequency-domain loss function, which starts with calculating the STFT to create a TF representation of the mixture sample. The TF bins corresponding to the target speech are then separated and used to synthesize the source waveform using inverse STFT. In this case, the loss function is formulated by the mean square error (MSE) between the TF bins estimated target speech $\hat{S} \in \mathbb{R}^{T\times F}$ and the corresponding ground truth $S \in R^{T\times F}$,

$$\mathcal{L}_f = \frac{1}{T \times F} ||S - \hat{S}||^2 \tag{2}$$

where $|| \cdot ||^2$ denotes the l_2 norm, T, and F denotes the number of frames and frequency bins, respectively. We also use the time-domain loss based on the mean MSE between the enhanced speech and clean speech, which is defined as:

$$\mathcal{L}_t = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i), \tag{3}$$

where x and $\hat{x}$ are the clean speech and enhanced speech samples, respectively, and N represents the number of samples. In this way, we obtain the final loss by combining these two types of loss functions,

$$\mathcal{L}_{total} = \alpha \mathcal{L}_f + (1 - \alpha)\mathcal{L}_t, \tag{4}$$

where α is a tunable parameter and is set as 0.2.

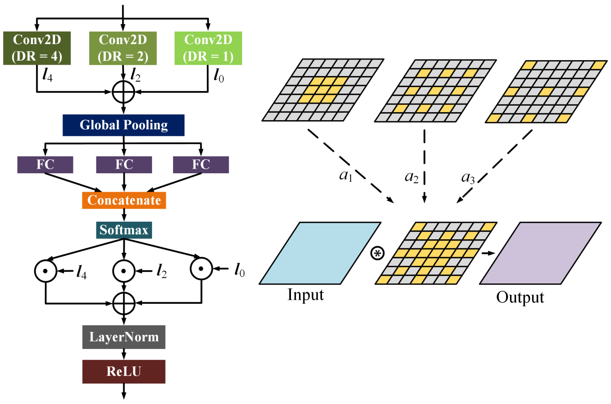

Figure 3: *The architecture of Multi-branch Dilation Convolution (MBDC).*

Table 1: *Comparison between MBDC, Dilated Convolution, and Convolution.*

Type	MBDC	Dilated Conv	Conv
Kernel Size	3×3	3×3	9×9
Dilation Rate	{1, 2, 4}	4	0
Receptive Field	9×9	9×9	9×9
Sampling Rate	33.33%	11.11%	100%
Param.(M)	N	N	9.00N

2.2. Dual-branch Module

Our proposed dual-branch module, labeled with an orange box in Figure 2, comprises three parts: MBDC for local processing, Self-CTFA module for global processing, and HFB for features fusion. The role of the dual-branch module is to leverage local and global operations adaptively.

Multi-branch Dilation Convolution (MBDC). Inspired by [24, 25], we employ the channel-wise attention mechanism to design the MBDC to perform channel selection with multiple convolutions with different dilation rates. The detailed architecture of our proposed MBDC is shown in Figure 3. In our design, we adopt three branches to carry different dilation rates of convolutional layers to generate feature maps with different receptive field sizes. The channel-wise attention is independently performed on these three outputs, and results are added together. In this way, the features can be extracted from local to non-local operations with the flexibility increasing, while the size of the receptive field is enlarged without substantial computational cost by the parallel structure of three dilated convolutional layers with the same kernel size [26]. In addition, comparison results between MBDC, dilated convolution, and convolution in Table 1 indicate that MBDC has a more significant feature sampling rate than dilated convolution while having much lower computational complexity than conventional convolution.

Self Channel-Time-Frequency Attention (Self-CTFA) Module. We show the proposed Self-CTFA module in Figure 4. The Self-CTFA module takes TF representation $\mathbf{F}_{in} \in \mathbb{R}^{C \times T \times F}$ as input, where C, T, and F denote the channels, time frames, and frequency bins, respectively. Self-CTFA module consists of three branches to generate a 1D channel-dimension energy feature $\mathbf{F}_c \in \mathbb{R}^{C \times 1}$, a 1D time-dimension energy feature $\mathbf{F}_t \in \mathbb{R}^{1 \times T}$, and a 1D frequency-dimension energy feature $\mathbf{F}_f \in \mathbb{R}^{F \times 1}$ in parallel by separately using three global pooling functions.

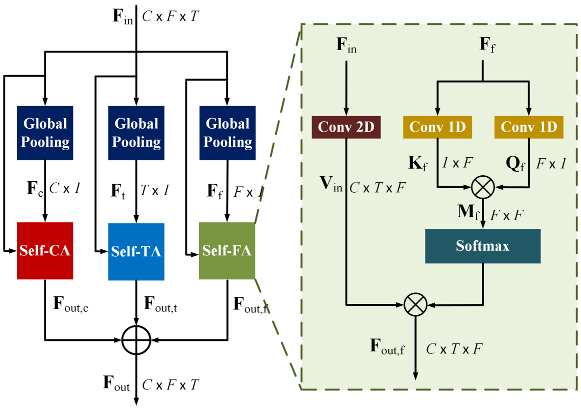
$\mathbf{F}_{\mathrm{in}}$ C x F x T
Global Pooling
Global Pooling
Global Pooling
$\mathbf{F}_{\mathrm{c}}$ C x 1
$\mathbf{F}_{\mathrm{t}}$ T x 1
$\mathbf{F}_{\mathrm{f}}$ F x 1
Self-CA
Self-TA
Self-FA
$\mathbf{F}_{\mathrm{out,c}}$
$\mathbf{F}_{\mathrm{out,t}}$
$\mathbf{F}_{\mathrm{out,f}}$
$\mathbf{F}_{\mathrm{out}}$ C x F x T
$\mathbf{F}_{\mathrm{in}}$
$\mathbf{F}_{\mathrm{f}}$
Conv 2D
Conv 1D
Conv 1D
$\mathbf{K}_{\mathrm{f}}$ 1 x F
$\mathbf{Q}_{\mathrm{f}}$ F x 1
$\mathbf{V}_{\mathrm{in}}$ C x T x F
$\mathbf{M}_{\mathrm{f}}$ F x F
Softmax
$\mathbf{F}_{\mathrm{out,f}}$ C x T x F

Figure 4: *The architecture of Self Channel-Time-Frequency Attention (Self-CTFA) Module.*

Each branch contains two sub-branches to calculate query and key using two 1D 1×1 convolutional layers. After that, the query and key of each branch are multiplied and fed into the softmax activation function to generate the attention feature map, which is defined as:

$$\begin{cases} \mathbf{M}_c = \mathrm{softmax}(\mathcal{H}_{c1}(\mathbf{F}_c) \times \mathcal{H}_{c1}(\mathbf{F}_c)^{\top}), \\ \mathbf{M}_t = \mathrm{softmax}(\mathcal{H}_{c1}(\mathbf{F}_t) \times \mathcal{H}_{c1}(\mathbf{F}_t)^{\top}), \\ \mathbf{M}_f = \mathrm{softmax}(\mathcal{H}_{c1}(\mathbf{F}_f) \times \mathcal{H}_{c1}(\mathbf{F}_f)^{\top}), \end{cases} \tag{5}$$

where $\mathbf{M}_c \in \mathbb{R}^{C \times C}$, $\mathbf{M}_t \in \mathbb{R}^{T \times T}$, and $\mathbf{M}_f \in \mathbb{R}^{F \times F}$ denote the attention feature maps of channel branch, time branch, and frequency branch, respectively, and $\mathcal{H}_{c1}$ represents the 1D 1×1 convolutional layer. Afterwards, the $\mathbf{V}_{in}$ generated from $\mathbf{F}_{in}$ separately multiples with $\mathbf{M}_c$, $\mathbf{M}_t$, $\mathbf{M}_f$, and these results are added to obtain the output of Self-CTFA module $\mathbf{F}_{out}$. In this way, the Self-CTFA module reduces the computational complexity from $\mathcal{O}(C^2TF+CT^2F+CTF^2)$ to $\mathcal{O}(C^2+T^2+F^2)$.

Hybrid Fusion Module (HFB). Considering the feature redundancy and knowledge discrepancy among MBDC and Self-CTFA module, we introduce a novel hybrid fusion block (HFB) in our approach. Specifically, we incorporate depth-wise separable convolutions and the channel attention layer into HFB to discriminatively aggregate features in spatial and channel dimensions [27].

3. Experimental Setup

3.1. Datasets

In order to evaluate the performance of the proposed model, experiments are conducted on Librispeech corpus [28]. There 6500 clean utterances are selected for the training set and 400 for the validation set, created under the random SNR levels ranging from -5dB to 10 dB. The test set contains 100 utterances under the SNR condition of -5dB, 0dB, 5dB, and 10 dB.

Noise signals from the Demand dataset [29], along with the clean speech recordings, are used to create the noisy speech for the training and validation set. The clean speech and noise recordings with a sampling frequency of 16 kHz and the frame size and frameshift for frame-level processing are set to 512 and 256.

Table 2: *Comparisons of baseline models in terms of STOI, PESQ, and SSNR.*

Test SNR	-5 dB			0 dB			5 dB			Param.	RTF
Metric	STOI (%)	PESQ	SSNR	STOI (%)	PESQ	SSNR	STOI (%)	PESQ	SSNR		
Unprocess	57.71	1.37	-5.04	71.02	1.73	-0.05	82.53	2.03	4.93	-	-
Conv-TasNet	79.81	2.23	6.65	86.76	2.42	8.23	91.46	2.86	10.12	4.58 M	0.72
GRN	80.31	2.21	6.62	86.98	2.49	8.31	91.28	2.91	10.89	3.08 M	0.68
TSTNN	83.76	2.32	6.98	89.75	2.64	8.86	93.67	3.03	11.59	4.86 M	0.86
U-Former	84.75	2.38	7.03	89.60	2.68	8.94	93.76	3.08	11.96	5.85 M	0.88
PCNN	87.24	2.51	7.84	92.17	2.83	9.71	94.82	3.13	12.34	3.15 M	0.51
PCNN (DC)	85.69	2.44	7.11	90.64	2.71	9.38	93.92	3.09	12.01	3.13 M	0.49
PCNN (CC)	85.98	2.46	7.69	91.86	2.74	9.39	94.19	3.11	12.14	3.57 M	0.54
PCNN (SA)	87.36	2.50	7.84	92.06	2.83	9.75	94.79	3.12	12.28	6.36 M	0.96
PCNN -w/o MBDC	83.32	2.30	6.86	87.96	2.59	8.78	92.89	2.98	11.87	2.98 M	0.46
PCNN -w/o Self-CTFA	82.67	2.26	6.79	87.01	2.51	8.66	91.47	2.93	10.98	2.74 M	0.43

Table 3: *Ablation study of self-CTFA by removing different components in -5 dB SNR condition.*

Metric	STOI (%)	PESQ	SSNR
PCNN	87.24	2.51	7.84
-w/o C Branch	84.81	2.47	7.51
-w/o T Branch	85.99	2.44	7.46
-w/o F Branch	85.56	2.41	7.49
-w/o C-T Branches	84.57	2.35	7.23
-w/o C-F Branches	84.59	2.32	7.15
-w/o T-F Branches	83.98	2.30	7.13
-w/o C-T-F Branches	83.05	2.26	6.98

3.2. Training and Network Parameters

The clean speech and noise recordings with a sampling frequency of 16 kHz and the frame size and frameshift for frame-level processing are set to 512 and 256. In each training epoch, we chunk a random segment of 4 seconds from an utterance if it is more significant than 4 seconds. The smaller utterances are zero-padded to match the size of the largest utterance in the batch. The Adam optimizer is used for stochastic gradient descent (SGD) based optimization, and the initial learning rate is set to 0.001. MSE is used as a loss function.

4. Results and Analysis

4.1. Model Comparison

This section compares alternative baseline models in Table 2 in terms of STOI, PESQ, and SSNR, where the numbers represent the averages over the test set in each condition. Four baseline systems are selected for the comparison, i.e., Conv-TasNet [17], GCRN [23], TSTNN [30], and U-Former [14]. In addition, we also evaluate the performance of PCNN when replacing MBDC with dilated convolution, i.e., PCNN (DC), replacing MBDC with conventional convolution, i.e., PCNN (CC), replacing self-CTFA module with self-attention module, i.e., PCNN (SA), removing MBDCs, i.e., PCNN -w/o MBDC, and removing self-CTFA module, i.e., PCNN -w/o Self-CTFA.

Table 2 shows the comparison results of the proposed PCNN and the other four baseline models. The number of parameters and real-time factory (RTF) is also presented. One can observe the following phenomena. First, the proposed model consistently outperforms all baselines in three metrics scores for different cases. Secondly, the proposed PCNN has the fewest parameters followed by GRN and lowest RTF, but PCNN has the best performance, demonstrating the higher performance effectiveness of PCNN. Thirdly, we compare the PCNN when using different components to replace MBDC and self-CTFA module in Table 2, according to the results, we observe that (1) replacing MBDC with dilated convolution achieves lower scores in these evaluation metrics with a similar number of parameters, (2) replacing MBDC with conventional convolution that has the exact size of the receptive field with dilated convolution achieves similar performance with a higher number of parameters and RTF. Finally, the comparison results between PCNN, PCNN -w/o MBDC, and PCNN -w/o Self-CTFA indicate the necessity of the proposed MBDC and self-CTFA module.

4.2. Impact of Branches in Self-CTFA Module

The proposed self-CTFA module consists of the channel (C) branch, time (T) branch, and frequency (F) branch. In this study, we evaluate variants of the self-CTFA module when removing the C, T, and F branches. We set the same parameters for the algorithm in the previous section but control the usage of different branches. Table 2 demonstrates the evaluation results of the self-CTFA module when removing other components in the -5 SNR condition. According to Table 2, we conclude that the existence of the C branch, T branch, and F branch does promote the performance of self-CTFA module. In addition, the F branch performs better than the T and C branches.

5. Conclusion

This paper proposes a parallel conformer neural network (PCNN) for SE by leveraging CNN for local detail information capture and the transformer for long-range dependencies extraction. According to the drawbacks of CNN and SA in the transformer, we develop an MBDC to address the small receptive field of CNN and a self-CTFA model to address the high computational complexity of SA. In addition, an HFB is utilized for leveraging the MBDC and self-CTFA. Through experiments, we show the superiority of the proposed method over other methods compared in this paper.

6. Acknowledgements

This work is supported by the National Nature Science Foundation of China (No. 62071342, No.62171326), the Special Fund of Hubei Luojia Laboratory (No. 220100019), the Hubei Province Technological Innovation Major Project (No. 2021BAA034) and the Fundamental Research Funds for the Central Universities (No.2042023kf1033).

7. References

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[2] K. A Al-Karawi, A. H Al-Noori, F. F. Li, T. Ritchings *et al.*, "Automatic speaker recognition system in adverse conditions—implication of noise and reverberation on system performance," *International Journal of Information and Electronics Engineering*, vol. 5, no. 6, pp. 423–427, 2015.

[3] B. Edwards, "The future of hearing aid technology," *Trends in amplification*, vol. 11, no. 1, pp. 31–45, 2007.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[5] X. Xu, W. Tu, and Y. Yang, "Selector-Enhancer: Learning dynamic selection of local and non-local attention operation for speech enhancement," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[6] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[7] S. Routray and Q. Mao, "Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network," *Computer Speech & Language*, vol. 71, p. 101270, 2022.

[8] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," *Proc. Interspeech 2017*, pp. 1993–1997, 2017.

[9] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[10] D. B. Pisoni, "Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning," *Speech communication*, vol. 13, no. 1-2, pp. 109–125, 1993.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[13] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[14] X. Xu and J. Hao, "U-Former: Improving monaural speech enhancement with multi-head self and cross attention," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 663–369.

[15] X. Xu, Y. Wang, J. Jia, B. Chen, and J. Hao, "GLD-Net: Improving Monaural Speech Enhancement by Learning Global and Local Dependency Features with GLD Block," in *Proc. Interspeech 2022*, 2022, pp. 966–970.

[16] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7847–7851.

[17] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[18] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6629–6633.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[20] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[21] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," *Proc. Interspeech 2018*, pp. 3723–3727, 2018.

[22] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part I 26*. Springer, 2020, pp. 653–665.

[23] 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, 2019.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[25] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[26] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1911–1920.

[27] K. Jiang, Z. Wang, C. Chen, Z. Wang, L. Cui, and C.-W. Lin, "Magic ELF: Image Deraining Meets Association Learning and Transformer," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 827–836.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[29] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[30] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7098–7102.