INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Monaural Speech Separation Method Based on Recurrent Attention with Parallel Branches

Xue Yang, Changchun Bao✉, Xu Zhang, Xianhong Chen

Speech and Audio Signal Processing Laboratory, Faculty of Information Technology, Beijing University of Technology, Beijing, China
yangx11@emails.bjut.edu.cn, baochch@bjut.edu.cn,
xuzhang223@emails.bjut.edu.cn, chenxianhong@bjut.edu.cn

Abstract

In many speech separation methods, the contextual information contained in the feature sequence is mainly modeled by recurrent layer and/or self-attention mechanism. However, how to combine these two powerful approaches more effectively needs to be explored. In this paper, a recurrent attention with parallel branches is proposed to first fully exploit the contextual information contained in the time-frequency (T-F) features. Then, this information is further modeled by the recurrent modules in a conventional manner. Specifically, the proposed recurrent attention with parallel branches uses two attention modules stacked sequentially. Each attention module has two parallel branches of self-attention to model dependencies along two axes and one convolutional layer for feature fusion. Thus, the contextual information contained in the T-F features can be fully exploited and further modeled by the recurrent modules. Experimental results showed the effectiveness of our proposed method.

Index Terms: speech enhancement, speech separation, contextual information

1. Introduction

Human beings have the impressive ability to focus on a particular speaker and to switch their attention freely among different talkers in complex acoustic scenarios. However, it is extremely difficult for the machine to imitate human behavior and tackle this kind of task, which is widely known as the cocktail party problem [1]. Attempting to crack this problem, speech separation plays an important role in separating the speech of each talker from the mixed signal [2]. Recently, with the help of deep learning, monaural speech separation has made a big progress. The separation performance is being improved continuously.

Under the supervised framework, one of the major challenges faced by speech separation is the label permutation problem, which involves how to correctly determine the training label assignment. Fortunately, this problem was solved delicately by deep clustering [3-4] and permutation invariant training [5-6]. Since then, various algorithms have been proposed to improve separation performance. Without taking full advantage of the phase spectrum, early methods [7-8] mainly dealt with magnitude spectrum in the time-frequency (T-F) domain and their performance did not exceed that of ideal magnitude masks. This situation has not changed until the introduction of end-to-end (E2E) method without using short-time Fourier transform (STFT). In [9], the waveform of mixed signal was directly transformed into a real-valued latent space through a convolutional layer so that the difficulty of phase estimation was circumvented. Although several stacks of dilated convolutional layers were used in [9], the receptive field of convolutional layers was limited and the contextual information could not be fully exploited.

To better model the contextual information and to solve the long sequence problem due to small frame length in E2E methods, the recurrent modules combined with dual-path strategy were adopted in [10]. Thus, the contextual information mainly along time axis was exploited by sequentially modelling the dependencies lied in the intra-segment and inter-segment. The basic block used in [10] is named as dual-path recurrent neural network (DPRNN) block. Subsequently, the self-attention mechanism was combined with recurrent modules in [11] to achieve better performance. Note that the attention module was concatenated with the recurrent module to jointly exploit the dependencies contained either in the intra-segment or in the inter-segment. Thus, the self-attention used in this kind of attention module is called axial attention. Besides, the basic block used in [11] is named as dual-path transformer network (DPTNet) block. Many recent studies followed in the footsteps of [10-11] and attempted to further improve separation performance through different approaches, such as: the multi-scale feature fusion [12-13], the additional identity information [14-15], the super-resolution technique [16], the quasi-dual-path method [17], etc.

For a while, the performance of separation method in T-F domain was not as attractive as that of E2E methods. However, recent breakthrough has been achieved with separation method performed in T-F domain. In [18], the DPTNet block was applied to model the contextual information along frequency axis and time axis, instead of that contained in intra-segment and inter-segment for E2E methods. Besides, the authors designed a novel T-F path scanning that replaced several layers of DPTNet block to model the transitions of adjacent frequency bins among adjacent frames. However, the axial attention and its subsequent recurrent module still exploited the contextual information along the same specific axis, i.e., frequency axis, time axis or the newly designed T-F path.

Generally speaking, the contextual information contained in the feature sequence is mainly modeled by recurrent layer and/or self-attention mechanism. For better performance, the attention module is usually concatenated with the recurrent module to jointly exploit the contextual information along the same specific axis. Whether such combination is the most effective is unknown and other ways to combine these two powerful approaches need to be explored. Several studies in speech enhancement have shown the effectiveness of providing more contextual information to subsequent modules. In [19], the spectro-temporal receptive field extractor was applied before the original sub-band model to further improve the network discrimination between speech signal and other interference. In [20], the time-frequency attention was achieved

10.21437/Interspeech.2023-518

through average pooling combined with convolutional layer and was incorporated before subsequent modules to guide the network to focus on more important T-F features. Inspired by these studies and the recurrent criss-cross attention in [21], we propose a recurrent attention with parallel branches (RAPB) in this paper. The proposed RAPB is used to first fully exploit the contextual information contained in the T-F features. Then, this information is further modeled by the recurrent modules in a conventional manner. Specifically, the proposed RAPB uses two attention modules stacked sequentially. Each attention module has two parallel branches of self-attention to model dependencies along two axes and one convolutional layer for feature fusion. Thus, the contextual information contained in the T-F features can be fully exploited and further modeled by the subsequent recurrent modules. Experimental results showed the effectiveness of our proposed method.

The rest of this paper is organized as follows. The details of the proposed method are described in Section 2. The experimental setup is presented in Section 3. The experimental results are shown and discussed in Section 4. The conclusions are given in Section 5.

2. Proposed Method

2.1. Problem formulation

The purpose of speech separation is to extract the speech signal of each speaker from the mixed signal, namely:

$$\hat{x}_j = \mathcal{M}(y) = \mathcal{M}\left(\sum_{k=1}^{S} z_k + n\right), \quad j = 1,...,S. \tag{1}$$

where $\hat{x}_{j,\ j=1,\ ...,\ S}$ are the separated signals, $\mathcal{M}(\cdot)$ represents the mapping function, y is the mixed signal captured by the microphone, S is the number of speakers in the mixed signal, z_k is the source image of the k^{th} speaker, n is the additive noise. Note that the separated signals are clean signals for both reverberant scenario and anechoic scenario in this paper.

2.2. Recurrent attention with parallel branches

In many speech separation methods, the attention module is usually concatenated with the recurrent module to jointly exploit the contextual information along the same specific axis. However, how to combine these two powerful approaches more effectively still needs to be explored. Inspired by [19-21], we propose a RAPB to first exploit the contextual information contained in the whole T-F features and this information is further modeled by the subsequent recurrent modules. The details of RAPB are depicted in Figure 1.

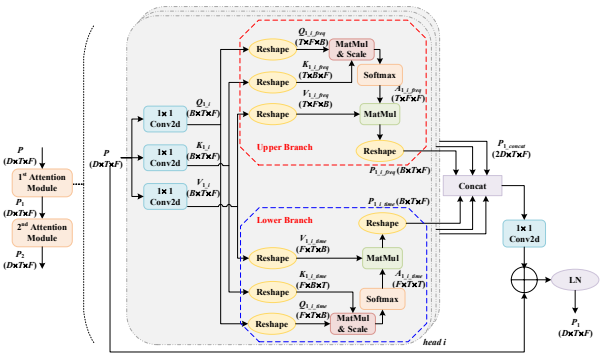

Figure 1: *The details of our proposed RAPB.*

As shown in the left part of Figure 1, the input $P \in \mathbb{R}^{D\times T\times F}$ is processed sequentially by two attention modules. Note that D is the channel dimension, T is the number of frames and F is the number of frequencies. These two modules have the same architecture and the first one is detailed in the right part of Figure 1. The number of heads is denoted as H and for head $i \in [1, H]$ of the first attention module, the input P is mapped into three feature tensors, i.e., query Q_{1_i}, key K_{1_i} and value V_{1_i}. These mapped feature tensors have the channel dimension $B=D/H$ and are utilized to perform two self-attentions separately. In the upper branch, the self-attention is done along the frequency axis. These three tensors are first reshaped and the resulting tensor dimensions are given in the parentheses. The reshaped tensors $Q_{1\_i\_freq}$ and $K_{1\_i\_freq}$ are first multiplied and scaled to calculate similarity. Then the attention map $A_{1\_i\_freq}$ is obtained by performing a softmax function along the last dimension. Finally, the reshaped value $V_{1\_i\_freq}$ is multiplied by the attention map $A_{1\_i\_freq}$ and reshaped back to obtain feature tensor $P_{1\_i\_freq}$. The same procedure is done in the lower branch along the time axis and the feature tensor $P_{1\_i\_time}$ is obtained. These feature tensors $P_{1\_i\_freq}$ and $P_{1\_i\_time}$ of each head are concatenated to form $P_{1\_concat} \in \mathbb{R}^{2D\times T\times F}$, which is further mapped through a convolutional layer to change the channel dimension back to D. The output $P_1 \in \mathbb{R}^{D\times T\times F}$ of the first attention module is obtained by skip connection and layer normalization (LN). This procedure can be written as:

$$P_{1\_i\_freq} = \text{Reshape}\left(Softmax\left(\frac{Q_{1\_i\_freq} K_{1\_i\_freq}}{\sqrt{B}} \right) V_{1\_i\_freq} \right) \tag{2}$$

$$P_{1\_i\_time} = \text{Reshape}\left(Softmax\left(\frac{Q_{1\_i\_time} K_{1\_i\_time}}{\sqrt{B}} \right) V_{1\_i\_time} \right) \tag{3}$$

$$P_{1\_concat} = \text{Concat}\left[..., P_{1\_i\_freq}, P_{1\_i\_time}, ... \right], \quad i = 1,...,H. \tag{4}$$

$$P_1 = LN\left(P + Conv2d\left(P_{1\_concat} \right) \right) \tag{5}$$

where Reshape$(\cdot)$ represents the reshape operation, $Softmax(\cdot)$ is the softmax function, Concat$[\cdot]$ represents the concatenation, $LN(\cdot)$ is the layer normalization and $Conv2d(\cdot)$ represents the 2-D convolution. With P_1 as input, the second attention module is executed in the same way as the first module and $P_2 \in \mathbb{R}^{D\times T\times F}$ denotes the final output of RAPB. Using two attention modules sequentially, RAPB can exploit the contextual information contained in the whole T-F features.

2.3. Proposed network architecture

In this paper, the proposed network is named as RAPBNet and can be divided into three parts as in [9], i.e., the encoder, the separator and the decoder, as shown in Figure 2(a).

The mixed signal y serves as the input of the encoder. The STFT is first used to obtained the complex spectrum Y. Note that the STFT is realized through 1-D convolution to enable E2E training. Then, the dynamic range compression (DRC) [22] is performed on the magnitude. The real and imaginary parts of the compressed spectrum is concatenated to form a feature tensor $Y^c \in \mathbb{R}^{2\times T\times F}$:

$$Y^c = \text{Concat}\left[|Y|^\alpha \cos\theta_Y, |Y|^\alpha \sin\theta_Y \right] \tag{6}$$

where α is the compression factor in the range (0, 1]. The magnitude and phase of Y are denoted as $|Y|$ and θ_Y, respectively. The stacked tensor Y^c is fed into a 7×7 convolutional layer and

a rectified linear unit (ReLU) to obtain the non-negative high-dimensional feature tensor $E \in R^{U \times T \times F}$.

In the separator, the feature tensor E is first layer normalized and its channel dimension is reduced through a convolutional layer. The obtained feature tensor $P \in R^{D \times T \times F}$ is processed by N RAPB blocks. Note that each RAPB block consists of one RAPB and two recurrent modules, as shown in Figure 2(b). The contextual information contained in the whole T-F features is exploited by the RAPB and is provided to the two recurrent modules. The recurrent modules model the dependencies along frequency axis and time axis through bi-directional long short-term memory (BLSTM) [23]. The fully connected (FC) layer, skip connection and LN are used as well. The feature tensor $R \in R^{D \times T \times F}$ in Figure 2(a) is further fed into a convolutional layer and a ReLU to obtain the mask for each speaker.

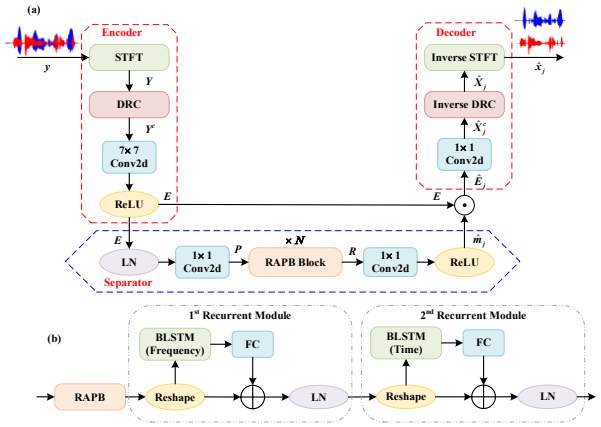

Figure 2: (a)*The proposed RAPBNet*. (b)*The details of RAPB block.*

The estimated masks are element-wise multiplied with the output of the encoder E to obtain the separated feature tensors $\hat{E}_{j,j=1,\ldots,S} \in R^{U \times T \times F}$. In the decoder, a convolutional layer is first used to change the channel dimension from U to 2. Subsequently, the inverse DRC and inverse STFT are performed to obtain the separated speech signals $\hat{x}_{j,\ j=1,\ \ldots,\ S}$. Note that the inverse STFT is realized through 1-D transposed convolution to enable E2E training.

3. Experimental setup

3.1. Datasets

In this paper, four common datasets are used to evaluate our proposed method. The first dataset is WSJ0-2Mix [3], in which each mixed signal is generated by randomly selecting utterances from different speakers in Wall Street Journal dataset (WSJ0). These selected utterances are further mixed at a random signal-to-noise ratio (SNR) between -5dB and 5dB. The second dataset is an open-source dataset Libri2Mix [24]. In this dataset, the train-360, dev and test set are used as the training, validation and test set, respectively. The third dataset is WSJ0 Hipster Ambient Mixtures (WHAM!) [25], which is an extension of WSJ0-2Mix towards more complex acoustic scenarios. In this dataset, the clean mixtures from WSJ0-2Mix are further mixed with real ambient noise, which is collected in coffee shops, restaurants and bars. The last dataset is WHAMR! [26], which attempts to simulate real-world scenarios and further includes reverberation. For each dataset, the speech signals are down-sampled to 8 kHz for reducing computational complexity and memory consumption.

3.2. Model configurations

In our proposed method, the Hanning widow is used to split speech signal into frames. The window length and the hop size are 32ms and 16ms, respectively. To perform speech separation in T-F domain, a 256-point STFT is used. Thus, 129 frequency bins are obtained. The compression factor α used in DRC is set to 0.5. For the convolutional layer in the encoder, the kernel size is set to (7, 7) for extracting local information. Except for this convolutional layer, the kernel sizes for other convolutional layers are all set to (1, 1). As for hyperparameters U and D, they are set to 256 and 64, respectively. In the RAPB block, the number of attention head H is 4 and the number of hidden units in each BLSTM is 128. Besides, the number N of RAPB blocks is set to 6 in this paper.

3.3. Training details

All models are trained on 4s long speech signals for 120 epochs. The optimization algorithm used is Adam [27] optimizer and the initial learning rate is set to 0.0005. In the first 100 epochs, the learning rate is multiplied by a factor of 0.98 for every two epochs. Then, the multiplied factor is reduced to 0.9. The gradient clipping is applied with a maximum L_2-norm of 1 when training the models on WHAMR! dataset and 5 on other datasets. Using utterance-level permutation invariant training (uPIT) [6], our training objective is to maximize the scale-invariant signal-to-distortion ratio (SI-SDR) [28].

4. Results and discussions

To verify the effectiveness of proposed method, the separation performance is evaluated through several well-known metrics, including SI-SDR improvement (SI-SDRi) and SDR improvement (SDRi) [29]. In addition, the computational complexities[1] of various methods are presented as well.

4.1. Comparison with baseline methods

To demonstrate the effectiveness of better exploiting contextual information, the proposed RAPBNet is compared with two baseline methods in this subsection. These three methods share the same framework and use different basic blocks in the separator. The proposed method takes RAPB block as the basic block where the contextual information contained in the whole T-F features is provided for the recurrent modules. For comparison, the basic block of Baseline 1 only consists of recurrent modules and there is no attention mechanism used. The basic block of Baseline 2 uses axial attention before the subsequent recurrent module, which means that the contextual information is exploited sequentially. For clarity, the basic blocks of these baseline methods are shown in Figure 3.

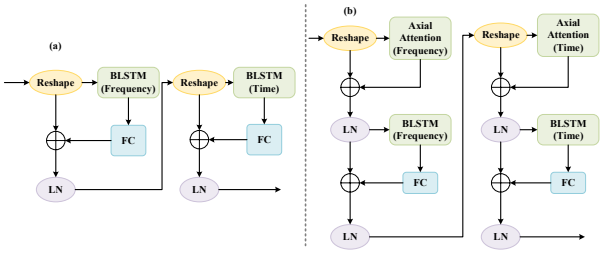

Figure 3: *The basic block of* (a) *Baseline 1 and* (b) *Baseline 2.*

[1]https://github.com/sovrasov/flops-counter.pytorch

Table 1: *Comparison with baseline methods on WSJ0-2Mix.*

Methods	SI-SDRi (dB)	SDRi (dB)	Parameter size ($\times 10^6$)	MACs ($\times 10^9$)
Baseline 1	20.0	20.1	2.66	22.2
Baseline 2	20.7	20.9	2.86	25.1
RAPBNet	21.3	21.4	2.91	24.2

The experiment was conducted on WSJ0-2Mix and the results are given in Table 1. Only using recurrent modules as the basic block, Baseline 1 can already achieve 20.0dB and 20.1dB on SI-SDRi and SDRi, respectively. Comparing these two baseline methods, Baseline 2 can achieve additional 0.7dB and 0.8dB improvements over SI-SDR and SDR by providing contextual information sequentially using axial attention. With RAPB, the proposed method can achieve 21.3dB and 21.4dB on SI-SDRi and SDRi, respectively. Compared with our proposed RAPBNet, Baseline 2 adopts axial attention and exploits the contextual information sequentially, leading to suboptimal performance. That is to say, the improvement over Baseline 2 indicates the effectiveness of better exploiting contextual information provided by the whole T-F features. The last two columns show the parameter size and the multiply-accumulate operations (MACs). Although the proposed RAPBNet uses two attention modules in one basic block, the increase in parameter size and MACs is small.

4.2. Results on WSJ0-2Mix and Libri2Mix

Several methods and their types are listed in Table 2 and compared with our proposed RAPBNet. The E2E methods listed here mainly exploit the contextual information along time axis. These E2E methods (except for DPRNN [10], Gated DPRNN [14] and Wavesplit [15]) use axial attention in their basic blocks and exploit the contextual information sequentially. On WSJ0-2Mix, several methods (from DPRNN [10] to Wavesplit [15]) perform worse than TFPSNet [18] and RAPBNet, which are separation methods performed in T-F domain. MTDS(DPTNet) [13] uses an additional time-delay sampling network to further model dependencies at different scales and RAPBNet performs slightly worse with fewer parameters. SFSRNet [16] utilizes the super resolution technique and QPDN [17] uses temporal convolutional layers to achieve better performance. However, the parameter size of these two methods is more than ten times of the proposed method. TFPSNet [18] uses three kinds of path scanning with axial attention to exploit the contextual information sequentially. Our proposed RAPBNet achieves higher performance by first exploiting the contextual information with RAPB, which is further processed with recurrent modules. Similarly, RAPBNet achieves competitive performance on Libri2Mix.

4.3. Results on WHAM! and WHAMR!

To further verify the applicability of the proposed method in complex acoustic scenarios, the experimental results are given in Table 3 on WHAM! and WHAMR!. The four comparison methods are all E2E methods and mainly model the contextual information along time axis. Also note that the Gated DPRNN [14] and Wavesplit [15] use additional identity information. Compared with these methods, RAPBNet achieves much better results on both datasets and shows more robustness for noisy and/or reverberant scenarios. This can be attributed to better modeling of the contextual information contained in the whole T-F features. These results prove the effectiveness and robustness of the proposed method.

Table 3: *Experimental results on WHAM! and WHAMR!. Results with superscript "†" were reported in [14].*

Methods	WHAM!		WHAMR!	
	SI-SDRi (dB)	SDRi (dB)	SI-SDRi (dB)	SDRi (dB)
Conv-TasNet [9]	12.7†	—	8.3†	—
DPRNN [10]	13.9†	—	10.3†	—
Gated DPRNN [14]	15.2	—	12.2	—
Wavesplit [15]	15.4	15.8	12.0	11.1
RAPBNet	16.3	16.5	15.6	14.0

5. Conclusions

In this paper, a monaural speech separation method was proposed based on recurrent attention with parallel branches. The proposed RAPB is used to first fully exploit the contextual information contained in the T-F features. Then, this information is further modeled by the recurrent modules in a conventional manner. Specifically, the proposed RAPB uses two attention modules stacked sequentially. Each attention module has two parallel branches of self-attention to model dependencies along two axes and one convolutional layer for feature fusion. Thus, the contextual information contained in the T-F features can be fully exploited and further modeled by the subsequent recurrent modules. Compared with several E2E methods and TFPSNet [18], our proposed RAPBNet achieved competitive performance on WSJ0-2Mix and Libri2Mix with a small number of parameters. For more complex acoustic scenarios, the proposed RAPBNet achieved higher performance, proving the effectiveness of better exploiting the contextual information contained in the whole T-F features with RAPB.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants 61831019 and 62006010).

Table 2: *Experimental results on WSJ0-2Mix and Libri2Mix. Results with superscript "†" were reported in [18].*

Methods	Type	Parameter size ($\times 10^6$)	WSJ0-2Mix		Libri2Mix	
			SI-SDRi (dB)	SDRi (dB)	SI-SDRi (dB)	SDRi (dB)
DPRNN [10]	E2E	2.6	18.8	19.0	16.5†	16.8†
Gated DPRNN [14]	E2E	7.5	20.1	—	—	—
DPTNet [11]	E2E	2.7	20.2	20.6	18.2†	18.4†
Sandglasset [12]	E2E	2.3	20.8	21.0	—	—
Wavesplit [15]	E2E	29	21.0	21.2	19.5	20.0
MTDS(DPTNet) [13]	E2E	4.0	21.5	21.7	—	—
SFSRNet [16]	E2E	59	22.0	22.1	21.1	21.4
QPDN [17]	E2E	200	22.1	—	—	—
TFPSNet [18]	T-F	2.7	21.1	21.3	19.7	19.9
RAPBNet	T-F	2.9	21.3	21.4	20.1	20.4

7. References

[1] E. C. Cherry, *On Human Communication*, Cambridge, MA, USA: MIT Press, 1957.

[2] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, Oct. 2018.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31-35.

[4] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-Channel Multi-Speaker Separation Using Deep Clustering," in *Proc. Interspeech 2016*, pp. 545-549.

[5] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241-245.

[6] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker Speech Separation Wth Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901-1913, Oct. 2017.

[7] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-Independent Speech Separation With Deep Attractor Network," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787-796, April 2018.

[8] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Alternative Objective Functions for Deep Clustering," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 686-690.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, Aug. 2019.

[10] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46-50.

[11] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *Proc. Interspeech 2020*, pp. 2642-2646.

[12] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A Light Multi-Granularity Self-Attentive Network for Time-Domain Speech Separation," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5759-5763.

[13] S. Qian, L. Gao, H. Jia, and Q. Mao, "Efficient Monaural Speech Separation with Multiscale Time-Delay Sampling," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6847-6851.

[14] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *International Conference on Machine Learning 2020*, pp. 7164-7175.

[15] N. Zeghidour and D. Grangier, "Wavesplit: End-to-End Speech Separation by Speaker Clustering," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840-2849, 2021.

[16] J. Rixen and M. Renz, "SFSRNet: Super-Resolution for Single-Channel Audio Source Separation," in *Proc. AAAI 2022*, pp. 11220-11228.

[17] J. Rixen and M. Renz, "QDPN - Quasi-dual-path Network for single-channel Speech Separation," in *Proc. Interspeech 2022*, pp. 5353-5357.

[18] L. Yang, W. Liu, and W. Wang, "TFPSNet: Time-Frequency Domain Path Scanning Network for Speech Separation," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6842-6846.

[19] F. Xiong, W. Chen, P. Wang, X. Li, and J. Feng, "Spectro-Temporal SubNet for Real-Time Monaural Speech Denoising and Dereverberation", in *Proc. Interspeech 2022*, pp. 931-935.

[20] Q. Zhang, X. Qian, Z. Ni, A. Nicolson, E. Ambikairajah, and H. Li, "A Time-Frequency Attention Module for Neural Speech Enhancement," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 462-475, 2023.

[21] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. ICCV 2019*, pp. 603-612.

[22] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Express Letters*, vol. 1, no. 1, pp. 014802, 2021.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[24] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv*:2005.11262, 2020.

[25] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech 2019*, pp. 1368-1372.

[26] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 696-700.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[28] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 626-630.

[29] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462-1469, July 2006.