INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Time-frequency Domain Filter-and-sum Network for Multi-channel Speech Separation

Zhewen Deng, Yi Zhou, Hongqing Liu

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing (CQUPT), China
Chongqing Key Laboratory of Signal and Information Processing, CQUPT, Chongqing, China
Intelligent Speech and Audio Research Lab (ISARL), CQUPT, Chongqing, China

jonathan.dzw@gmail.com, zhouy@cqupt.edu.cn, hongqingliu@cqupt.edu.cn

Abstract

Learning-based methods have made impressive strides in speech separation, and the implicit filter-and-sum network (iFaSNet) stands out as a reliable multi-channel solution. Meanwhile, the TF-GridNet has achieved state-of-the-art performance on the WSJ0-2mix dataset, indicating the underlying capability of time-frequency (T-F) domain speech separation methods. This paper investigates the possibility of constructing a T-F domain filter-and-sum network that improves upon the iFaSNet. In addition to optimizing the separation module, we develop a narrow-band spatial feature as a cross-channel feature and a convolution module for context decoding. With these enhancements, we redesign each module under the iFaSNet architecture, which entirely operates in the T-F domain. Thus, the proposed method is referred to as the TF-FaSNet. Experimental results on fixed microphone array geometries show that the TF-FaSNet outperforms the standard iFaSNet under all conditions with similar model complexity.

Index Terms: Speech separation, multi-channel, time-frequency domain, network model

1. Introduction

In recent years, end-to-end time-domain models incorporating learned encoder-decoder modules as a direct replacement of the short-time Fourier transform (STFT) have gradually dominated speech separation under anechoic conditions since the invention of the time-domain audio separation network (TasNet) [1]. Since 2019, most research advances in single-channel speech separation have been made using the time-domain model, such as the convolutional time-domain audio separation network (Conv-TaSNet) and its variants [2–12]. One of the latest time-domain models [4] reports a substantial scale-invariant signal-to-distortion ratio improvement (SI-SDRi) of 22.1 dB on the WSJ0-2mix dataset [13]. Given the popularity of time-domain models, a recent study has shown that T-F domain models are beginning to demonstrate their advantages. TF-GridNet [14], which operates in the T-F domain, achieves an impressive 23.4 dB SI-SDRi on the WSJ0-2mix dataset, which significantly outperforms all existing time-domain models. Following the dual-path recurrent neural network (DPRNN) [3] and the time-frequency domain path scanning network (TFPSNet) [15], it proposes a novel multi-path architecture where each block consists of an intra-frame spectral module, a sub-band temporal module, and a full-band self-attention module. By stacking multiple multi-path blocks, it learns the patterns of the speech spectrogram in a grid-like manner and utilizes local and global spectro-temporal information for separation. This well-designed model illustrates the potential of T-F domain approaches.

Inspired by the conventional filter-and-sum beamformer, the original filter-and-sum network (FaSNet) [16, 17] estimated a set of beamforming filters with a neural network and later performed filter-and-sum beamforming in the time domain. In the consecutive work, the iFaSNet [18] adopted the encoder-decoder architecture to estimate the filter in a learnable latent space. Both the vanilla FaSNet and its implicit variant use a stack of dual-path recurrent neural network blocks with the transform-average-concatenate module (DPRNN-TAC) as the separation module. In this work, we go one step further to explore how to integrate the multi-path architecture into the iFaSNet architecture adequately. The ultimate goal is to transform it into a T-F domain model since the potential of the T-F domain model has been shown in recent studies. To that aim, three modifications to iFaSNet are investigated: (1) The proposed method leverages a multi-path separation module to perform complex spectral mapping [19–21] in the T-F domain as a substitute for DPRNN-TAC. In addition, a 2D positional encoding is added to aid attention modules in learning spectro-temporal information. (2) The proposed method employs narrow-band feature extraction to exploit the inter-channel cues of different speakers since the original design of TF-GridNet lacks cross-channel information to address multi-channel conditions adequately. (3) Since the multi-path separation module mainly consists of long short-term memory (LSTM) modules and attention modules, it is good at capturing content-based global interactions. To add more local interactions, a convolution module is added at the end of the separation module to exploit local features effectively. Experimental results show that the proposed method achieves better results than iFaSNet under various data configurations while maintaining the same model size.

2. Proposed method

This work is based on the iFaSNet and aims to improve its performance. Specifically, the DPRNN-TAC architecture, which serves as its separation module, is replaced with a more powerful multi-path architecture. To ensure compatibility with the iFaSNet structure, the model is divided into five parts: encoding, feature extraction, separation, context decoding, and decoding, and other compatible components are designed to ensure the multi-path network fits seamlessly into the five-part design.

As shown in Figure 1, five processing stages are indicated with different colors from left to right. Figure 2 shows the structure of the multi-path network used in the proposed network, and Figure 3 depicts the detailed structure of the rest processing stages. The same network block is highlighted with the same color for each processing stage.

The five-part design of the original iFaSNet is introduced in Section 2.1. Then, Section 2.2 presents the proposed method

10.21437/Interspeech.2023-165

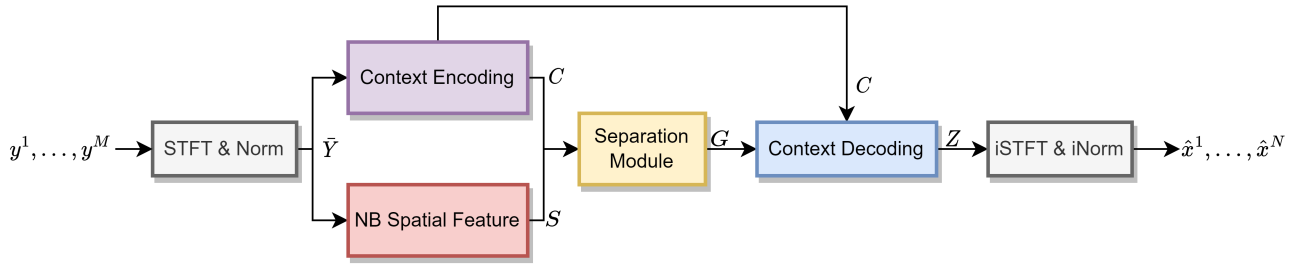
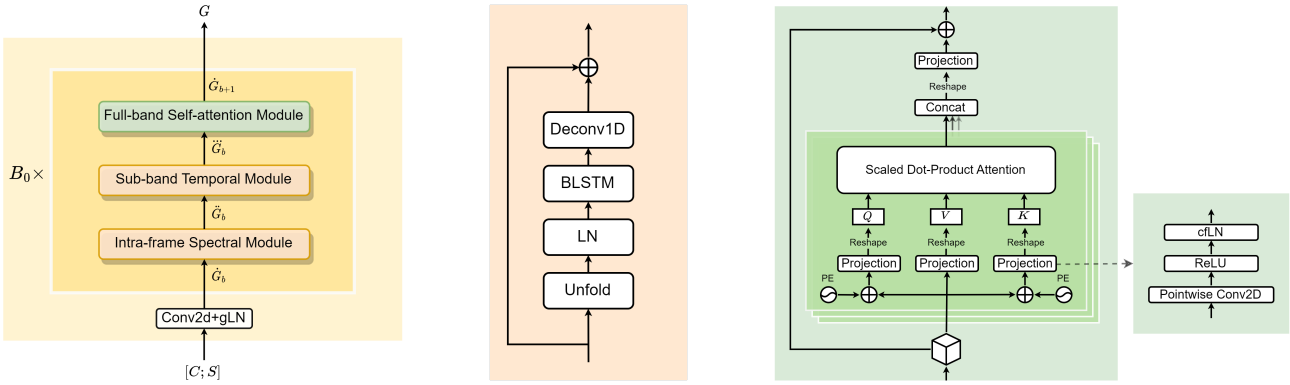


Figure 1: *The block diagram of the TF-FaSNet system.*



(a) *Separation module* (b) *Spectral & temporal module* (c) *Self-attention module*

Figure 2: *An overview of the separation module.*

and three major transformations made to the original iFaSNet.

2.1. Implicit filter-and-sum network recap

For an N-speaker mixture signal recorded by an M-microphone array in a noisy-reverberant setting, the i^{th} frame of the input mixtures from the M channels is denoted as $\{\boldsymbol{y}_i^m\}_{m=1}^M \in \mathbb{R}^{1\times L}$. The iFaSNet consists of five steps: (1) An encoder transforms time-domain waveform $\boldsymbol{y}_i^m$ into a feature $\boldsymbol{c}_i^m \in \mathbb{R}^{1\times H_0}$ in the latent space. Preceding and succeeding E feature vectors are stacked before and after the center feature $\boldsymbol{c}_i^m$ to form the context feature $\boldsymbol{C}_i^m \in \mathbb{R}^{(1+2E)\times H_0}$. (2) To extract the channel-wise and cross-channel features, two feature extraction modules receive the context feature $\boldsymbol{C}_i^m$ to produce the compressed context feature $\hat{\boldsymbol{c}}_i^m \in \mathbb{R}^{1\times H_0}$ and feature-level normalized cross correlation (fNCC) feature $\hat{\boldsymbol{s}}_i^m \in \mathbb{R}^{1\times(1+2E)^2}$. (3) The channel-wise and cross-channel features are passed to the DPRNN-TAC separation module to generate a feature vector $\boldsymbol{g}_i^1 \in \mathbb{R}^{1\times H_0}$ as the separation cue. (4) A context decoder receives the context feature $\boldsymbol{C}_i^m$ and the feature vector $\boldsymbol{g}_i^1$ to estimate the filters $[\hat{\boldsymbol{h}}_{i-E}^{1,n}, \ldots, \hat{\boldsymbol{h}}_i^{1,n}, \ldots, \hat{\boldsymbol{h}}_{i+E}^{1,n}] \in \mathbb{R}^{(1+2E)\times H_0}$, where $\hat{\boldsymbol{h}}_i^{1,n}$ denotes the i^{th} frame estimated filter for the n^{th} speaker. The filters are then applied to the encoder outputs, and mean-pooling is applied across $1+2E$ context vectors:

$$\boldsymbol{z}_i^{1,n} = \frac{1}{1+2E}\sum_{j=0}^{2E} \boldsymbol{c}_{i-E+j}^{1,n} \odot \hat{\boldsymbol{h}}_{i-E+j}^{1,n} \tag{1}$$

where $\odot$ denotes the Hadamard product, and $\boldsymbol{z}_i^{1,n}$ represents the filtering result of the i^{th} frame for the n^{th} speaker in the latent space. (5) A transposed 1D convolution layer transforms the latent feature back to the waveform. Thus, the separated speeches for each speaker $\{\hat{x}^n\}_{n=1}^N$ are estimated by iFaSNet.

2.2. Time-frequency domain filter-and-sum network

2.2.1. Complex spectral mapping with separation module

Figure 2a shows the flowchart of the customized multi-path architecture used in the proposed network. The concatenated feature denoted by $[C; S] \in \mathbb{R}^{2D\times T\times F}$ is first fed into a 2D convolution layer followed by global layer normalization (gLN) to obtain a $D \times T \times F$ tensor. The tensor is subsequently sent to B_0 blocks of the separation module to refine the T-F embedding progressively. In each block, the input tensor of the b^{th} block, denoted as $\dot{G}_b \in R^{D\times T\times F}$, is first passed into the intra-frame spectral module and the sub-band temporal module to explore spectro-temporal information. Both modules have the same structure, as shown in Figure 2b. The output of the sub-band temporal module, denoted as $\ddot{G}_b \in \mathbb{R}^{D\times T\times F}$, is then fed into the full-band self-attention module. In this module, the procedure is almost the same as that in the multi-head attention module proposed in [22]. However, instead of applying linear projections, it utilizes a 2D convolution layer, a parametric rectified linear unit (PreLU) activation function, and a layer normalization among the channel and frequency dimensions to transform 3D inputs into 2D queries, keys, and values. After the attention outputs are concatenated, they are reshaped to a $D \times T \times F$ tensor and passed to the same projection again. The tensor is then added to input $\ddot{G}_b$ through a residual connection, producing a tensor $\dot{G}_{b+1}$ as the input to the next block. In addition to the original design of the full-band self-attention module, a 2D positional encoding proposed in [23] is applied to inject the spatial information into queries and keys to produce

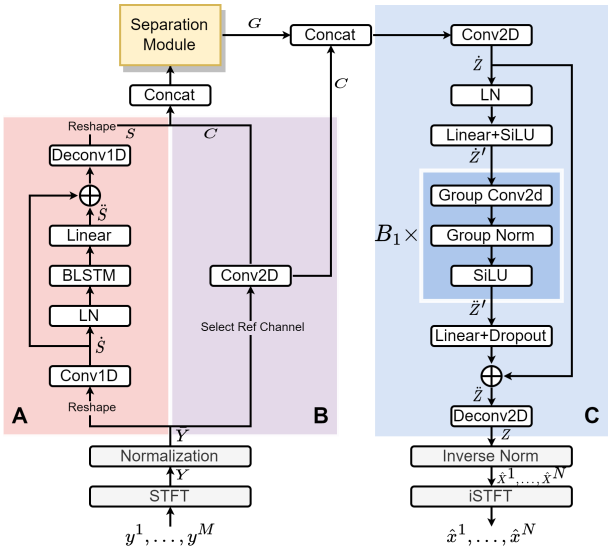


Figure 3: *A flowchart of the proposed method. Three network blocks are: (A) Narrow-band spatial feature extraction (B) Context encoding (C) Context decoding.*

location-aware attention scores. As shown in Figure 2c, the input $\ddot{G}_b$ is a $D \times T \times F$ tensor, and the 2D positional encoding is added to queries and keys before projections. We believe it can facilitate the training process by providing the spatial relationship among all complex components in different directions.

In the proposed network, since the multi-path architecture is employed as the separation module, the way to generate estimation results should be reconsidered. Experiments in [14] show that using TF-GridNet with mapping provides better results than masking. Hence, the proposed method directly performs complex spectral mapping with a context decoder rather than estimating filters to perform masking. As a result, the context decoder output Z represents the predicted real and imaginary (RI) components of each speaker.

2.2.2. *Inter-channel narrow-band spatial features*

For channel-wise features, extracting the contextual information between the RI components within each channel is the main focus, especially for the reference channel. Based on this intuition, only the RI components of the reference channel denoted as $\bar{Y}_{ref} \in \mathbb{R}^{2\times T\times F}$ are passed through a 2D convolution layer to generate context feature $C \in \mathbb{R}^{D\times T\times F}$. Although this leads to the sacrifice of inter- and intra-channel information of other channels, it preserves the most critical contextual information of the reference channel and, more importantly, facilitates subsequent context decoding. As shown in Figure 3 Block B, the context decoder receives the context feature C as one of the inputs rather than getting the RI components of the reference channel like in iFaSNet. We assume that it brings richer context information to the context decoder, which benefits the decoder in getting better results.

The task of extracting cross-channel information is assigned to another feature extraction module in parallel. With the narrow-band technique introduced in [24–26], the module can focus on exploiting the inter-channel cues of different speakers. As shown in Figure 3 Block A, the input $\bar{Y} \in \mathbb{R}^{2M\times T\times F}$ is viewed as F narrow-band sequences with sequence length T and the number of features $2M$. Next, the F sequences are passed to a 1D convolution layer to compute H_1 dimensional embedding for each sequence, obtaining $\dot{S} \in \mathbb{R}^{F\times T\times H_1}$. A layer of bidirectional LSTM and a fully-connected layer are used to extract spatial information within each narrow-band sequence, and layer normalization is applied to maintain the stability of training. The output of the linear layer $\ddot{S} \in \mathbb{R}^{F\times T\times H_1}$ is added to $\dot{S}$ through a residual connection, and a 1D deconvolution layer then transforms the hidden dimension of H_1 to D. Lastly, the context feature $S \in \mathbb{R}^{D\times T\times F}$ is obtained by a reshape operation.

2.2.3. *Context decoding with a convolution module*

Each block of the multi-path separation module can be divided into two parts: a dual-path network formulated by the spectral and temporal modules, and a self-attention network. Since both networks aim to capture long-range global information, a convolution module is employed to gather local information on the fine-grained features produced by the separation module as a way to add some local interactions.

As shown in Figure 3 Block C, the concatenation of the context feature C and the separation cue G is first passed to a 2D convolution layer with $2D$ output channels, obtaining $\dot{Z} \in \mathbb{R}^{2D\times T\times F}$. Based on [27,28], a structure that utilizes prenorm residual units and two linear layers combined with the sigmoid linear unit (SiLU) activation and dropout is created. This structure resembles the feed forward module in Conformer. To incorporate local spectro-temporal information, B_1 blocks of group convolution are sandwiched between the two linear layers. Each group convolution consists of a 2D group convolution (GroupConv2D) layer, a group normalization (GN) layer, and a SiLU activation. The b^{th} convolution block can be formulated as:

$$\ddot{Z}'_b = \text{SiLU}(\text{GN}(\text{GroupConv2D}(\dot{Z}'_b))) \in \mathbb{R}^{H_2\times T\times F} \quad (2)$$

where $\dot{Z}'_b$ and $\ddot{Z}'_b$ denote the input and output of the b^{th} convolution block, respectively. H_2 is the number of hidden units between two linear layers, which is restored to $2D$ by the second linear layer. At last, a 2D deconvolution layer with $2N$ output channels is applied to $\ddot{Z}$ for complex spectral mapping. The inverse normalization and inverse STFT (iSTFT) are subsequently applied to context decoding result $Z \in \mathbb{R}^{2N\times T\times F}$ for signal re-synthesis.

3. Experiment configurations

3.1. Dataset

The proposed method is evaluated on a simulated multi-channel two-speaker noisy speech dataset with fixed geometry microphone arrays, the same as in [17]. The simulated dataset contains 20000, 5000, and 3000 4-second long utterances sampled at 16 kHz for training, validation, and test sets, respectively. For each utterance, two speakers and one nonspeech noise are randomly selected from the 100-hour Librispeech dataset [29] and the 100 Nonspeech Corpus [30], respectively. An overlap ratio between the two speakers is uniformly sampled between 0% and 100% such that the average overlap ratio across the dataset is 50%. The two speech signals are adjusted in time and scaled to a random signal-to-noise ratio (SNR) between 0 and 5 dB. The relative SNR between the combined power of the two clean speech signals and the noise is randomly selected between 10 and 20 dB. For more specific settings, please refer to [17].

Table 1: *Comparison with other methods on the simulated 6-mic circular array dataset*

Model	# of param.	SI-SDR (dB)									PESQ
		Speaker angle				Overlap ratio				Average	
		<15°	15-45°	45-90°	>90°	<25%	25-50%	50-75%	>75%		
Mixture	-	-0.5	-0.4	-0.4	-0.4	-0.4	-0.4	-0.5	-0.4	-0.4	1.35
MC-TasNet-S	1.3M	7.6	7.8	8.3	8.4	13.0	9.0	6.3	3.7	8.0	1.54
MC-TasNet-L	2.6M	8.2	8.5	8.8	9.1	13.4	9.7	7.0	4.4	8.6	1.59
FaSNet-TAC-S	2.1M	7.6	9.8	11.4	12.2	14.1	11.2	9.0	6.6	10.2	1.77
FaSNet-TAC-L	3.5M	8.3	10.4	11.8	12.6	14.6	11.7	9.4	7.3	10.8	1.81
iFaSNet-S	2.0M	7.8	8.9	9.8	9.7	13.7	10.1	7.5	4.8	9.0	1.62
iFaSNet-L	3.3M	8.2	9.7	10.5	10.4	14.2	10.6	8.2	5.7	9.7	1.67
TD-GWF-iter.1	4.0M	11.2	13.2	14.3	15.1	17.3	14.5	12.2	9.7	13.4	2.03
TD-GWF-iter.2		11.9	14.0	15.2	15.9	17.8	15.2	13.2	10.6	14.2	2.17
MC-TF-GridNet-S	2.4M	13.6	14.4	15.1	15.8	18.6	15.4	13.3	11.5	14.7	2.66
MC-TF-GridNet-L	3.6M	13.7	14.6	15.6	16.4	18.8	15.8	13.6	12.0	15.1	2.72
TF-FaSNet	2.5M	**14.8**	**15.4**	**15.8**	**16.1**	**19.3**	**16.2**	**14.2**	**12.3**	**15.5**	**2.81**

3.2. Model evaluation and training configurations

The proposed method is compared with five baselines: (a) MC-TasNet [31], a multi-channel version of a single-channel TasNet system, achieved by using extra encoders to encode each input channel signal. Its separator receives the concatenated features for further operations. (b) FaSNet-TAC [17], which is FaSNet with transform-average-concatenate (TAC) module. (c) iFaSNet [18], introduced in the previous section, is a variant of the FaSNet-TAC system. (d) TD-GWF [32], the Time-Domain Real-Valued Generalized Wiener Filter, is a novel sequential beamforming pipeline that performs iterative beamforming and separation. Both its pre-separation module and post-separation module are set to FaSNet-TAC. (e) MC-TF-GridNet [14], a multi-channel version of TF-GridNet. Similar to (a), the single-channel TF-GridNet system is extended by stacking $2M$ RI components and modifying the input channel of the first 2D convolution from 2 to $2M$, where M denotes the number of microphones. The benchmark systems consist of two settings: a small setting indicated by "-S" and a large setting indicated by "-L". However, the iterative model, TD-GWF, is labeled differently, with the number of iterations denoted as "-iter.1" or "-iter.2".

All models are trained with the Adam optimizer [33] for 100 epochs starting with an initial learning rate of 0.001. The scale-invariant signal-to-distortion ratio (SI-SDR) is used to evaluate the speech separation performance. The learning rate is halved when the validation loss does not decrease in 3 consecutive epochs. Training is stopped when it does not decrease in 10 consecutive epochs (early stop), and gradient clipping is applied with a threshold of 1.

The implementation of TF-FaSNet is available online[1].

4. Results and discussions

In Table 1, the performance of the proposed method is compared with those of others over the simulated 6-microphone circular array dataset. As shown in the table, all T-F domain methods based on multi-path architecture obtain better SI-SDRi than DPRNN-based methods (FaSNet-TAC, iFaSNet), which indicates that by integrating full-band and sub-band modeling on complex components, T-F domain models can further improve the overall performance. One observation is that iFaSNet produces a poor overall result compared to FaSNet-TAC. This is probably due to its poor performance on mixtures with a high overlap ratio since the performance gap between FaSNet-TAC and iFaSNet becomes more significant as the overlap ratio increases. However, for smaller speaker angles, the performance of iFaSNet can match or even exceed the performance of FaSNet-TAC, suggesting that the iFaSNet architecture may be able to use spatial information more efficiently to discriminate speakers from different angles. This phenomenon is also reflected in the proposed network. Compared to MC-TF-GridNet, the proposed method performs significantly better for smaller speaker angles. A possible explanation for this is that the iFaSNet architecture combined with the narrow-band method can better exploit the spatial information from different microphones to identify the spatial cues of different speakers. Another improvement is that for mixtures with different overlap ratios, TF-FaSNet has a consistent improvement in SI-SDRi compared to MC-TF-GridNet, rather than having a noticeable performance degradation as the overlap ratio gets higher, like iFaSNet. This improvement is probably due to the context decoding design, which helps the network more efficiently extract contextual information from non-overlapping signals to distinguish different speakers. This experiment shows that although iFaSNet performs poorly on the fixed array dataset, and a large portion of the overall performance improvement is contributed by the multi-path architecture, by integrating iFaSNet architecture, the proposed method manages to improve the performance with a modest number of parameters increase. As a result, the proposed method outperforms other benchmark methods with a relatively small model size in most conditions.

5. Conclusions

This paper introduces a novel approach to multi-channel speech separation that improves upon existing methods. The proposed approach involves transforming each module of the iFaSNet architecture to perform separation in the time-frequency domain, building on recent progress in monaural speech separation. Three major changes were made to the model to achieve this, including the use of a multi-path separation module with 2D positional encoding, a narrow-band spatial feature as the cross-channel feature, and a convolution module for context decoding. The experimental results indicate that the proposed method is superior under the experimental conditions.

[1] https://github.com/JonathanDZ/TF-FaSNet

6. References

[1] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[2] ——, "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.

[3] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[4] J. Rixen and M. Renz, "SFSRNet: Super-resolution for single-channel audio source separation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 220–11 228.

[5] S. Qian, L. Gao, H. Jia, and Q. Mao, "Efficient monaural speech separation with multiscale time-delay sampling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6847–6851.

[6] J. Rixen and M. Renz, "QDPN-quasi-dual-path network for single-channel speech separation," *Proc. Interspeech 2022*, pp. 5353–5357, 2022.

[7] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[8] M. W. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5759–5763.

[9] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[10] Y. Zhu, X. Zheng, X. Wu, W. Liu, L. Pi, and M. Chen, "DPTCN-ATPP: Multi-scale end-to-end modeling for single-channel speech separation," in *5th International Conference on Communication and Information Systems (ICCIS)*. IEEE, 2021, pp. 39–44.

[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*, 2020, pp. 2642–2646.

[12] D. Markovic, A. Defossez, and A. Richard, "Implicit Neural Spatial Filtering for Multichannel Source Separation in the Waveform Domain," in *Proc. Interspeech 2022*, 2022, pp. 1806–1810.

[13] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[14] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," *arXiv preprint arXiv:2209.03952*, 2022.

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

[16] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 260–267.

[17] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6394–6398.

[18] Y. Luo and N. Mesgarani, "Implicit Filter-and-Sum Network for End-to-End Multi-Channel Speech Separation," in *Proc. Interspeech 2021*, 2021, pp. 3071–3075.

[19] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[20] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *IEEE 27th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2017, pp. 1–6.

[21] 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.

[22] 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.

[23] Z. Wang and J.-C. Liu, "Translating math formula images to latex sequences using deep neural networks with sequence-level training," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 24, no. 1-2, pp. 63–75, 2021.

[24] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 298–302.

[25] ——, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 298–302.

[26] C. Quan and X. Li, "Multi-channel narrow-band deep speech separation with full-band permutation invariant training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 541–545.

[27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[28] C. Quan and X. Li, "Multichannel Speech Separation with Narrow-band Conformer," in *Proc. Interspeech 2022*, 2022, pp. 5378–5382.

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

[30] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[31] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7319–7323.

[32] Y. Luo, "A Time-Domain Real-Valued Generalized Wiener Filter for Multi-Channel Neural Separation Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3008–3019, 2022.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.