INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Gated Recurrent Fusion of Spatial and Spectral Features for Multi-channel Speech Separation with Deep Embedding Representations

Cunhang Fan[1,3], *Jianhua Tao*[1,3], *Bin Liu*[1], *Jiangyan Yi*[1], *Zhengqi Wen*[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing
{cunhang.fan, jhtao, liubin, jiangyan.yi, zqwen}nlpr.ia.ac.cn

Abstract

Multi-channel deep clustering (MDC) has acquired a good performance for speech separation. However, MDC only applies the spatial features as the additional information, which does not fuse them with the spectral features very well. So it is difficult to learn mutual relationship between spatial and spectral features. Besides, the training objective of MDC is defined at embedding vectors, rather than real separated sources, which may damage the separation performance. In this work, we deal with spatial and spectral features as two different modalities. We propose the gated recurrent fusion (GRF) method to adaptively select and fuse the relevant information from spectral and spatial features by making use of the gate and memory modules. In addition, to solve the training objective problem of MDC, the real separated sources are used as the training objectives. Specifically, we apply the deep clustering network to extract deep embedding features. Instead of using the unsupervised K-means clustering to estimate binary masks, another supervised network is utilized to learn soft masks from these deep embedding features. Our experiments are conducted on a spatialized reverberant version of WSJ0-2mix dataset. Experimental results show that the proposed method outperforms MDC baseline and even better than the oracle ideal binary mask (IBM).

Index Terms: Multi-channel deep clustering, speech separation, deep attention fusion, deep embedding features

1. Introduction

Speech separation is known as the cocktail party problem [1], which aims to estimate the target sources from a noisy mixture. To address this problem, there are many works have been done and made significant advances, such as deep clustering (DC) [2, 3], permutation invariant training (PIT) [4, 5], Conv-TasNet [6] and end-to-end post-filter with deep attention fusion features [7, 8]. They all do not use the spatial information because they are monaural speech separation methods. As for the multiple microphones, they contain the directional information of each source. Therefore, the spatial features can be leveraged to the multi-channel speech separation. In this work, we focus on the multi-channel speech separation.

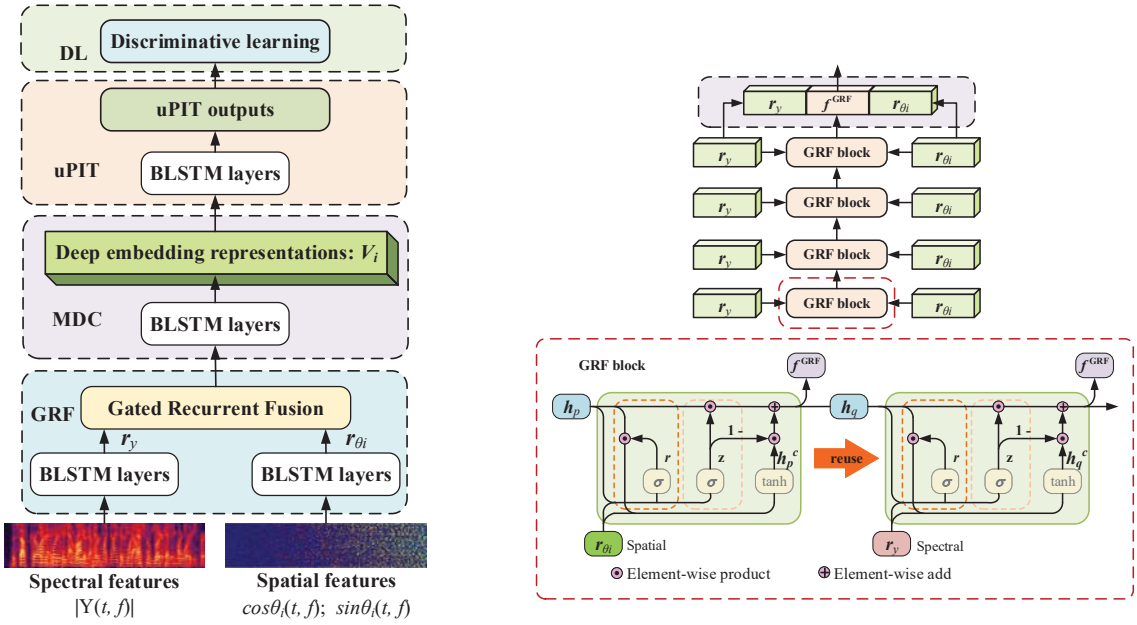
Recently, to utilize the spatial information, many works have been done for multi-channel speech separation [9, 10, 11, 12, 13]. Multi-channel deep clustering (MDC) [14] extends the DC to multi-channel. DC [2] is a single channel speech separation technique. It trains a bidirectional long-short term memory (BLSTM) network to map the mixed spectrogram into an embedding space. At testing stage, the embedding vector of each time-frequency (T-F) bin is clustered by K-means to obtain binary masks. Different from DC, MDC uses the interchannel phase differences (IPDs) [15] as the additional spatial features to the separation model. In other words, MDC applies not only spectral but also the spatial features as the input for better separation. Although MDC can separate the mixture well, there are still two limitations. Firstly, MDC only uses the spatial features as the additional information, which is difficult to learn mutual relationship between spatial and spectral features. Secondly, the training objective of MDC is defined at the embedding vectors, rather than real separated sources. These embedding vectors do not necessarily imply the perfect separation of sources in signal space.

Motivated by [16] and our previous work [17], we propose a spatial and spectral features gated recurrent fusion (GRF) method for multi-channel speech separation using deep embedding representations. Different from MDC only using the IPDs as the additional features, we utilize the GRF algorithm to fuse the spectral and spatial features as two different modalities. Therefore, the GRF could learn to adaptively select and fuse the relevant information from spectral and spatial features by making use of the gate and memory modules. In addition, to address the training objective problem of MDC, motivated by our previous work [17], we apply the deep embedding representations for multi-channel speech separation. Specifically, the MDC network is utilized to extract deep embedding representations. Instead of using the unsupervised K-means clustering algorithm to estimate binary masks, the supervised utterance-level PIT (uPIT) [5] network is applied to learn soft masks from these deep embedding representations. Therefore, the separation model can use the real separated sources as the training objective. Finally, to reduce the distance between the same speakers and increase the distance between different speakers, the discriminative learning [18, 19, 20] is utilized to fine-tune the separation model.

To summarize, the main contribution of this paper is two-fold. Firstly, we deal with the spectral and spatial features as two different modalities and apply the gated recurrent fusion algorithm to fuse them deeply. Secondly, the MDC is applied to extract deep embedding representations. And another supervised uPIT network with discriminative learning is used to learn target masks instead of the unsupervised K-means clustering. Therefore, the separation model can use the real separated sources as the training objective.

The rest of this paper is organized as follows. Section 2 presents the multi-channel deep clustering. The proposed method is stated in section 3. Section 4 shows detailed experiments and results. Section 5 draws conclusions.

http://dx.doi.org/10.21437/Interspeech.2020-1548



(a) Overall schematic diagram

(b) Gated Recurrent Fusion

Figure 1: *(a) Overall schematic diagram of our proposed method for multi-channel speech separation. (b) Schematic diagram of Gated Recurrent Fusion, which fuses the spatial and spectral features as two different modalities.*

2. Multi-channel deep clustering

The aim of single-channel deep clustering (DC) [2, 3] is to map the mixture spectrogram into a high-dimensional embedding V for each T-F bin by a deep neural network (DNN). The loss function of DC is defined as follows:

$$\begin{aligned} J_{DC} &= ||VV^T - BB^T||_F^2 \\ &= ||VV^T||_F^2 - 2||V^TB||_F^2 + ||BB^T||_F^2 \end{aligned} \tag{1}$$

where $B \in \mathbb{R}^{TF \times S}$ is the source membership function for each T-F bin, i.e.,$B_{tf,s} = 1$, if source s has the highest energy at time t and frequency f compared to the other sources. Otherwise, $B_{tf,s} = 0$. S is the number of sources. $||*||_F^2$ is the squared Frobenius norm.

The difference between single-channel DC and multi-channel DC (MDC) is the input features. As for the single-channel DC, only the mixture spectrogram $|Y(t,f)|$ is used as the input feature: $\zeta_{DNN} = \{|Y(t,f)|\}$. As for the MDC, the phase difference between two microphones, $\theta_i(t,f)$ (i is the index of a microphone pair), is applied as an additional input feature as follows:

$$\zeta_{DNN} = \{|Y(t,f)|; cos\theta_i(t,f); sin\theta_i(t,f)\} \tag{2}$$

Besides, when the number of microphone $N_m > 2$, MDC firstly chooses a reference microphone and each pair $\theta_i(t,f)$ is computed between a reference and non-reference microphone. Therefore, there will be $N_m - 1$ embeddings. When these embeddings are stacked at each T-F bin, the K-means clustering is applied to estimate binary masks. Finally, these masks are utilized to the reference microphone signal for separation.

3. The proposed separation method

In this section, we present our proposed spatial and spectral features with gated recurrent fusion (GRF) algorithm for speech separation using deep embedding representations, which is shown in Fig. 1(a). Instead of simply stacking the spatial and spectral features, the GRF algorithm is utilized to combine them deeply, which deals with them as two different modalities. Therefore, the GRF could learn to adaptively select and fuse the relevant information from spectral and spatial features by making use of the gate and memory modules. In addition, to address the training objective problem of MDC, the real separated sources are used as the training objectives, which uses the deep embedding representations for multi-channel speech separation. Finally, the discriminative learning is used to educe the distance between the same speakers and increase the distance between different speakers.

3.1. Gated Recurrent Fusion

As shown in Fig. 1(a), the spectral features $|Y(t,f)|$ and spatial IPDs $\{cos\theta_i(t,f); sin\theta_i(t,f)\}$ are firstly processed by the BLSTM network to acquire the deep representations. These spectral and spatial deep representations are denoted by $\boldsymbol{r}_y$ and $\boldsymbol{r}_{\theta_i}$, respectively. We apply the GRF to fuse the spectral and spatial features as two different modalities. The gate structure in gated recurrent unit (GRU) [21] enables the selective fusion of multi-modal features. In this paper, we extend the GRU as our GRF block and modify it to fit the feature fusion.

As shown in Fig. 1(b), at fist step (left), GRF block takes one of the spatial features as input. The outputs of this step will be used as the hidden state. Then, in the second step (right), GRF fusion block takes the spectral features as input. These two steps reuse the GRF block and share the same set of parameters. The GRF block consists of reset gate, update gate, adaptive memory and selective fusion.

As for the reset gate, at step p, the hidden state h_p and the input $\boldsymbol{r}_{\theta_i}$ together to decide the status of the reset gate $\boldsymbol{r}$ by

$$\boldsymbol{r} = \sigma(W_r, (\boldsymbol{r}_{\theta_i}, h_p)) \tag{3}$$

where σ denotes the sigmoid function, W_r is the weight of reset gate.

The update gate $\boldsymbol{z}$ is also decided by the hidden state h_p and input features $\boldsymbol{r}_{\theta_i}$.

$$\boldsymbol{z} = \sigma(W_z, (\boldsymbol{r}_{\theta_i}, h_p)) \tag{4}$$

where W_z is the weight of update gate.

As for adaptive memory, through the element-wise product $\odot$, the reset gate $\boldsymbol{r}$ determines how much information in the past needs to be memorized.

$$h_p^{'} = \boldsymbol{r} \odot h_p \tag{5}$$

$$h_p^c = tanh(W_h(\boldsymbol{r}_{\theta_i}, h_p^{'})) \tag{6}$$

where h_p^c acts similarly to the memory cell in the LSTM and helps the GRF block to remember long term information within the multi-stage fusion.

The selective fusion aims to combine the $\boldsymbol{r}_{\theta_i}$ and h_p. The fusion result at step p is

$$h_q = \boldsymbol{z} \odot h_p + (1 - \boldsymbol{z}) \odot h_p^c \tag{7}$$

In this way, the forget gate $\boldsymbol{z}$ and the input gate $(1 - \boldsymbol{z})$ are linked. That is, if the previous information is ignored with a weight of $\boldsymbol{z}$, then the information for the current input h_p^c would be selected with a weight of $(1 - \boldsymbol{z})$.

In this paper, we use the 4-stage GRF. The first stage, we randomly initialize the hidden state h_p. After the 4-stage GRF, we can acquire the fusion features: $f^{GRF} = h_p$. Finally, the fusion features f^{GRF} and these spectral and spatial deep representations ($\boldsymbol{r}_y$ and $\boldsymbol{r}_{\theta_i}$) are used as the GRF features. They are applied to extract the deep embedding representations.

3.2. Deep Embedding Representations for Separation

Clusters in the embedding space of MDC can represent the inferred spectral masking patterns of individual sources. In this paper, we utilize the MDC network to extract deep embedding representations, which contain the information of each source and are conducive to speech separation. The D-dimensional deep embedding representations V_i (i is the index of a microphone pair) can be extracted as follows:

$$V_i = \xi_{BLSTM}\{\boldsymbol{r}_y; f^{GRF}; \boldsymbol{r}_{\theta_i}\}\ (i = 1, 2, ..., N_m - 1) \tag{8}$$

where ξ_{BLSTM} denotes the mapping of BLSTM network.

In order to address the training objective problem of MDC, instead of using the unsupervised K-means clustering, the supervised uPIT network is applied to estimate soft masks from these deep embedding representations. When the $V_1, V_2, ..., V_{N_m-1}$ are stacked, they are sent to the uPIT network: $\zeta_{uPIT} = \{V_1, V_2, ..., V_{N_m-1}\}$. The uPIT network computes the mean square error (MSE) for all possible speaker permutations at utterance-level. Then the minimum cost among all permutations (P) is chosen as the optimal assignment.

$$\phi^* = \underset{P}{arg\,min} \sum_{s=1}^{S} |||Y| \odot \widetilde{M}_s - |X_s| cos(\theta_y - \theta_s)||_F^2 \tag{9}$$

where the number of all permutations P is $S!$ (! denotes the factorial symbol). $\widetilde{M}_s$ is the estimated phase sensitive mask (PSM) [22] of source s. θ_y and θ_s are the reference microphone phase of mixture speech and target source s. $|X_s|$ is the spectrogram of target source s.

3.3. Discriminative Learning and Joint Training

To reduce the distance between the same speakers and increase the distance between different speakers, discriminative learning (DL) is applied to our proposed model. The DL loss function can be defined as follows:

$$J_{DL} = \phi^* - \sum_{\phi \neq \phi^*, \phi \in P} \alpha\phi \tag{10}$$

where ϕ is a permutation from P but not ϕ^*, $\alpha \geq 0$ is the regularization parameter of ϕ. When $\alpha = 0$, the loss function is same as the ϕ^* in Eq. 9. It means without DL.

To extract embedding features effectively, we apply the joint training framework to the proposed system. The loss function of joint training is defined as follows:

$$\begin{aligned} J &= \lambda J_{DC} + (1 - \lambda) J_{DL} \\ &= \lambda J_{DC} + (1 - \lambda)(\phi^* - \sum_{\phi \neq \phi^*, \phi \in P} \alpha\phi) \end{aligned} \tag{11}$$

where $\lambda \in [0, 1]$ controls the weight of J_{DC} and J_{DL}.

4. Experiments and Results

4.1. Dataset

The room impulse response (RIR) generator [1] is used to spatialize the WSJ0-2mix dataset [2]. The dataset consists of three sets: training set (20,000 utterances about 30 hours), validation set (5,000 utterances about 10 hours) and test set (3,000 utterances about 5 hours). Specifically, the training and validation sets are generated by randomly selecting utterances from WSJ0 training set (`si_tr_s`). Similar as generating training and validation set, the test set is created by mixing the utterances from the WSJ0 development set (`si_dt_05`) and evaluation set (`si_et_05`).

The RIR is generated using the image-source method [23] with a linear microphone array with 4 microphones. The reverberation time RT_{60} is set to 0.16s. The distances between 4 microphones are 4-8-4 cm. For any two speakers, we constrain them to be at least $45°$ apart. We mix the images of two speakers with signal-to-noise ratios (SNRs) between -5dB and 5dB. The average distance between a source and array center is set to 1m.

4.2. Experimental setup

The first channel is used as the reference microphone. The sampling rate of all generated data is 8 kHz. The short-time Fourier transform (STFT) has 32 ms length hamming window and 8 ms window shift.

The proposed method contains 4 BLSTM layers, each with 600 units in each direction. More specifically, as for the deep attention fusion module, there is only one BLSTM layer for spectral and spatial, respectively. As for the DC network, there is also only one BLSTM layer. As for the uPIT network, there are two BLSTM layers. The dimension D of the embeddings is set to 20 per T-F bin [14]. The regularization parameter α of discriminative learning is set to 0.1. And the joint training weight λ is set to 0.01.

In this work, the models are evaluated on the signal-to-distortion ratio (SDR) [24], the perceptual evaluation of speech quality (PESQ) [25] measure and the short-time objective intelligibility (STOI) measure [26].

[1] Available online at https://github.com/ehabets/RIR-Generator

Table 1: *The results of SDR, PESQ and STOI for different separation methods for different gender combinations. The "MDC+GRF" means that the MDC applies the GRF for spatial and spectral features. The "MDC+GRF+uPIT" means that the uPIT loss is used in MDC+GRF. The "MDC+GRF+uPIT+DL" means that the discriminative learning is used in MDC+GRF+uPIT.*

Methods	Male-Female			Female-Female			Male-Male			AVG.		
	SDR(dB)	PESQ	STOI(%)	SDR(dB)	PESQ	STOI(%)	SDR(dB)	PESQ	STOI(%)	SDR(dB)	PESQ	STOI(%)
Mixture	0.15	1.32	61.26	0.16	1.37	62.38	0.15	1.30	61.74	0.15	1.33	61.59
MDC(baseline)	12.7	2.70	89.75	13.0	2.81	90.95	11.9	2.55	89.67	12.5	2.68	89.94
MDC+GRF	13.0	2.74	90.03	13.4	2.84	91.21	12.1	2.59	89.96	12.8	2.71	90.22
MDC+GRF+uPIT	14.5	3.39	93.36	15.0	3.44	95.08	13.6	3.32	93.49	14.4	3.38	93.70
MDC+GRF+uPIT+DL	**14.7**	**3.40**	**93.58**	**15.2**	**3.45**	**95.24**	**13.8**	**3.33**	**93.73**	**14.5**	**3.39**	**93.92**
IBM	13.7	3.29	91.77	14.2	3.36	93.75	12.8	3.17	91.05	13.5	3.26	91.91
IAM	13.0	3.80	95.44	13.4	3.79	96.38	12.0	3.82	95.19	12.8	3.80	95.53
IPSM	16.7	4.05	96.40	17.1	4.05	97.15	15.7	4.04	95.71	16.5	4.05	96.33

We apply the MDC [14] as our baseline and re-implement it with our experimental setup. More specifically, the MDC has 4 BLSTM layers with 600 units, which is same as the proposed method. As for the test, unsupervised K-means clustering is always performed on the entire utterance to acquire a binary mask for speech separation.

All models contain random dropouts with a dropout rate 0.5. Each minibatch contains 8 randomly selected utterances. The number of epoch is 20. The learning rate is initialized as 0.00001. Our models are implemented using Pytorch deep learning framework [2].

4.3. Experimental results

Table 1 shows the results of SDR, PESQ and STOI for different separation methods and different gender combinations. The "MDC+GRF" means that the MDC applies the GRF for spatial and spectral features. The "MDC+GRF+uPIT" means that the uPIT loss is used in MDC+GRF. The "MDC+GRF+uPIT+DL" means that the discriminative learning is used in MDC+GRF+uPIT. In addition, the last three rows present the results of the ideal binary mask (IBM), ideal amplitude mask (IAM) [5] and the ideal PSM (IPSM), which are oracle masks.

4.3.1. Evaluation of gated recurrent fusion

From Table 1 we can find that when the gated recurrent fusion is applied to the MDC, the performance of speech separation can be improved no matter what gender combinations for these three evaluation metrics. For example, the average SDR of "MDC+GRF" can be increased from 12.5 to 12.8 compared with the MDC method. This result indicates that this GRF algorithm is effective for speech separation. The reason is that the GRF fuses the spectral and spatial features as two different modalities. Therefore, the GRF could learn to adaptively select and fuse the relevant information from spectral and spatial features by making use of the gate and memory modules. In other words, it combines the spectral and spatial features deeply. So compared with the MDC method, the proposed "MDC+GRF" method can acquire a better speech separation performance.

4.3.2. The effectiveness of our proposed method

From Table 1 we can make several observations. Firstly, compared with the MDC baseline method, the performance of the proposed methods can be significantly improved. More specifically, compared with the MDC, the proposed "MDC+GRF+uPIT+DL" method obtains 16.0%, 26.5% and 4.4% relative improvements in SDR, PESQ, and STOI, respectively. Secondly, we surprisingly find that the results of the proposed "MDC+GRF+uPIT+DL" method are better than using the oracle mask of IBM. Note that these IBM results are the limit results of MDC baseline method. These results indicate the effectiveness of the proposed method. Thirdly, when the deep embedding representations ("MDC+GRF+uPIT") are applied, the separation performance can be improved. This is because that these deep embedding representations contain the potential information of each target source so that they can effectively estimate the masks of target sources. Therefore, these deep embedding features are discriminative features for speech separation. In addition, the separation model can use the real separated sources as the training objective. Finally, no matter what gender combinations, our proposed speech separation method can acquire better results than the baseline method. These results reveal that our proposed method has an effective ability to reconstruct target sources for all of the gender combinations.

[2] Available online at https://pytorch.org/

5. Conclusions

In this paper, we propose a spatial and spectral gated recurrent fusion method for multi-channel speech separation using deep embedding representations. The GRF fuses the spectral and spatial features as two different modalities. In addition, to address the training objective problem of MDC, the MDC is applied to extract deep embedding representations. Instead of utilizing the unsupervised K-means clustering, the supervised uPIT network is used to learn soft target masks. Results show that the proposed method outperforms MDC baseline, with relative improvements of 16.0%, 26.5% and 4.4% in SDR, PESQ, and STOI, respectively. Besides, the proposed method is even better than the oracle IBM. In the future, we will explore phase enhancement based on the proposed method.

6. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002802), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61901473, No.61771472, No.61773379) and Inria-CAS Joint Research Project (No.173211KYSB20170061 and No.173211KYSB20190049)

7. References

[1] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinncunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral Cortex*, vol. 25, no. 7, p. 1697, 2015.

[2] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 31–35.

[3] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *Interspeech 2016*, pp. 545–549, 2016.

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

[5] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[6] 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 (TASLP)*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen, and X. Liu, "End-to-end post-filter for speech separation with deep attention fusion features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1303–1314, 2020.

[8] ——, "Deep attention fusion feature for speech separation with end-to-end post-filter method," *arXiv preprint arXiv:2002.01626*, 2020.

[9] Z.-Q. Wang and D. Wang, "Integrating spectral and spatial features for multi-channel speaker separation." in *Interspeech*, 2018, pp. 2718–2722.

[10] ——, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2018.

[11] M. Togami, "Spatial constraint on multi-channel deep clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 531–535.

[12] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," *Proc. Interspeech 2019*, pp. 4290–4294, 2019.

[13] C. Fan, B. Liu, J. Tao, J. Yi, and Z. Wen, "Spatial and spectral deep attention fusion for multi-channel speech separation using deep embedding features," *arXiv preprint arXiv:2002.01626*, 2020.

[14] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[15] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[16] Y. Liu, J. Li, Q. Yan, X. Yuan, C. Zhao, I. Reid, and C. Cadena, "3d gated recurrent fusion for semantic scene completion," *arXiv preprint arXiv:2002.07269*, 2020.

[17] C. Fan, B. Liu, J. Tao, J. Yi, and Z. Wen, "Discriminative learning for monaural speech separation using deep embedding features," *Proc. Interspeech 2019*, pp. 4599–4603, 2019.

[18] C. Fan, B. Liu, J. Tao, Z. Wen, J. Yi, and Y. Bai, "Utterance-level permutation invariant training with discriminative learning for single channel speech separation," in *ISCSLP*. IEEE, 2018, pp. 26–30.

[19] E. M. Grais, G. Roma, A. J. Simpson, and M. Plumbley, "Combining mask estimates for single channel audio source separation using deep neural networks," *Interspeech2016 Proceedings*, 2016.

[20] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks." in *ISMIR*, 2014, pp. 477–482.

[21] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[22] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.