MULTI-CHANNEL NARROW-BAND DEEP SPEECH SEPARATION WITH FULL-BAND PERMUTATION INVARIANT TRAINING

Changsheng Quan [1,2], *Xiaofei Li* [2,*]

[1] Zhejiang University, Hangzhou, China
[2] Westlake University & Westlake Institute for Advanced Study, Hangzhou, China

ABSTRACT

This paper addresses the problem of multi-channel multi-speech separation based on deep learning techniques. In the short time Fourier transform domain, we propose an end-to-end narrow-band network that directly takes as input the multi-channel mixture signals of one frequency, and outputs the separated signals of this frequency. In narrow-band, the spatial information (or inter-channel difference) can well discriminate between speakers at different positions. This information is intensively used in many narrow-band speech separation methods, such as beamforming and clustering of spatial vectors. The proposed network is trained to learn a rule to automatically exploit this information and perform speech separation. Such a rule should be valid for any frequency, thence the network is shared by all frequencies. In addition, a full-band permutation invariant training criterion is proposed to solve the frequency permutation problem encountered by most narrow-band methods. Experiments show that, by focusing on deeply learning the narrow-band information, the proposed method outperforms the oracle beamforming method and the state-of-the-art deep learning based method.

Index Terms— Speech separation, deep learning, narrow-band, multi-channel, reverberant environments

1. INTRODUCTION

This work addresses the multi-channel speech source separation problem leveraging deep learning techniques. Traditional speech source separation methods are often designed in the short-time Fourier transform (STFT) domain. In [1], the authors introduced the W-disjoint orthogonality assumption based on the time-frequency (TF) sparsity of speech, namely each TF bin of the mixed speech signals is assumed to be dominated by one speech source. Consequently, speech sources can be separated by clustering the TF bins. The authors of [2] proposed to estimate the mixing matrix using hierarchical clustering of the TF-wise mixing vectors. Beamforming conducts speech separation by applying spatial filtering. The widely used minimum variance distortionless response (MVDR) beamformer preserves the speech with desired steering vector while suppresses others [3]. The recently proposed Guided Source Separation (GSS) [4] method combines the techniques of TF-bin clustering and MVDR beamforming, which achieves excellent speech separation performance, and thus is extensively adopted by the participants of the CHiME-6 multispeaker speech recognition challenge [5]. The traditional methods mentioned above [2, 3, 4] are all performed in narrow-band, namely processing the STFT frequencies separately. This leads to the well-known frequency permutation problem that the separated signals at different frequencies should be assigned to the same source.

Deep learning has been first used for single-channel speech separation by learning the spectral pattern of speech. Deep clustering (DC) [6] first learns an embedding for each TF bin, then uses k-means to cluster these embeddings and separate the TF bins. Permutation invariant training (PIT) [7] directly predicts the TF mask of each source, and then the separated signals can be obtained by multiplying the masks with the mixture signal. PIT uses the minimum loss of all possible source permutations for training. When multiple microphones are available, in addition to the spectral pattern of speech, the spatial information of speakers can be employed. The single-channel DC method is extended to the multi-channel case in [8] by integrating the spatial features, such as inter-channel phase difference, into the network input. In [9], it is proposed to automatically learn spatial features with neural network. An end-to-end filter-and-sum network (FaSNet) [10] is proposed to directly predict the spatial filters from the time-domain mixture signals. It is further combined with the transform-average-concatenate (TAC) paradigm in [11], which eventually achieves excellent speech separation performance. Another popular technique is to estimate the beamformer parameters using the speech separation results of deep learning based methods, such as in [12, 13].

In this work, we propose an end-to-end narrow-band speech separation network. A long short-term memory (LSTM) network is designed to take as input the STFT coefficients of multi-channel mixture signals for one frequency, and predict the STFT coefficients of multiple speech sources for the same frequency. By analyzing the traditional narrow-

[*] corresponding author

band methods [2, 3, 4], we can find that one STFT frequency includes rich informations to separate speech sources, such as the spectral sparsity of speech and the inter-channel differences of multiple sources. The proposed network is trained to learn a function/rule to automatically exploit these informations, and to perform end-to-end narrow-band speech separation. As is the case for traditional methods, one unique function/rule should be learned for all frequencies, thus the proposed network is designed to be shared by all frequencies.

The proposed narrow-band method also suffers from the frequency permutation problem. To solve this problem, the spectral correlation of adjacent frequencies, or the spatial consistency of multiple frequencies for the same speaker, can be used [14, 15]. In this paper, inspired by utterance-level PIT (uPIT) [16], we propose a training criterion called full-band PIT (fPIT) to solve the frequency permutation problem. It requires the network to output the separated signals of all frequencies belonging to the same speaker at the same output position. This scheme can be implemented in an end-to-end manner within the training process.

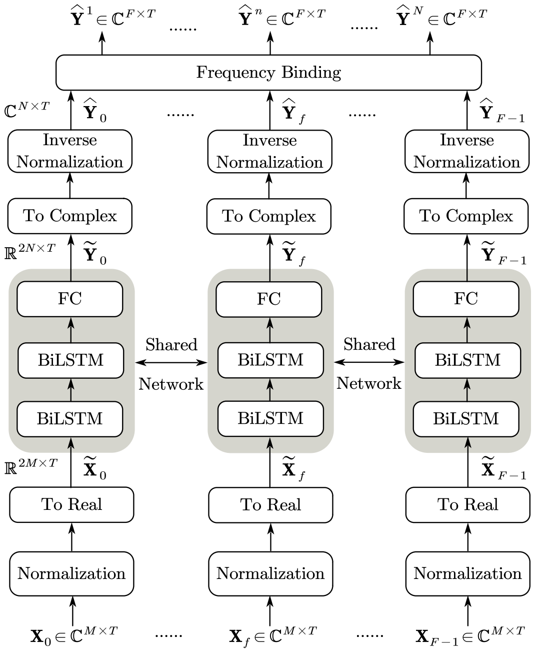
This paper is a continuous work of our previous papers [17, 18], in which a narrow-band network was proposed for speech denoising by exploiting the differences between speech and noise, as speech is non-stationary and directional while noise is stationary and spatially diffuse. Narrow-band speech separation is a different task, in the sense that it mainly exploits the inter-channel cues of different speakers. Compared to the full-band multi-channel methods, such as FaSNet [10, 11], even though our proposed method does not learn the spectral pattern of speech at all, it employs a powerful network dedicated to fully leverage the narrow-band information. Experiments show that the proposed method achieves better performance than FaSNet. The code for our proposed method can be found at[1].

2. METHOD

We consider multichannel signals in the STFT domain: $\mathrm{X}^m_{f,t} = \sum_{n=1}^{N} \mathrm{Y}^{n,m}_{f,t}$, where $f \in \{0, ..., F-1\}$, $t \in \{1, ..., T\}$, $m \in \{1, ..., M\}$ and $n \in \{1, ...N\}$ denote the indices of frequency, time frame, microphone channel and speaker, respectively, $\mathrm{X}^m_{f,t}$ and $\mathrm{Y}^{n,m}_{f,t}$ are the complex-valued STFT coefficients of the microphone signals and reverberant spatial image of speech sources, respectively. This work focuses on the separation task, and our target is to recover the reverberant spatial image of multiple speakers at a reference channel, e.g. $\mathrm{Y}^{n,r}_{f,t}$, where r denotes the index of reference channel.

2.1. Narrow-band Deep Speech Separation

The flowchart of the proposed method is shown in Fig.1. The basic idea is to perform speech separation independently for each frequency.



$\widehat{\mathbf{Y}}^1 \in \mathbb{C}^{F\times T}$
......
$\widehat{\mathbf{Y}}^n \in \mathbb{C}^{F\times T}$
......
$\widehat{\mathbf{Y}}^N \in \mathbb{C}^{F\times T}$
Frequency Binding
$\mathbb{C}^{N\times T}$
$\widehat{\mathbf{Y}}_0$
$\widehat{\mathbf{Y}}_f$
$\widehat{\mathbf{Y}}_{F-1}$
Inverse Normalization
To Complex
$\mathbb{R}^{2N\times T}$
$\widetilde{\mathbf{Y}}_0$
$\widetilde{\mathbf{Y}}_f$
$\widetilde{\mathbf{Y}}_{F-1}$
FC
BiLSTM
BiLSTM
Shared Network
$\mathbb{R}^{2M\times T}$
$\widetilde{\mathbf{X}}_0$
$\widetilde{\mathbf{X}}_f$
$\widetilde{\mathbf{X}}_{F-1}$
To Real
Normalization
$\mathbf{X}_0 \in \mathbb{C}^{M\times T}$
$\mathbf{X}_f \in \mathbb{C}^{M\times T}$
$\mathbf{X}_{F-1} \in \mathbb{C}^{M\times T}$

Fig. 1: The flowchart of narrow-band deep speech separation. Each frequency is processed individually. The dimension of each intermediate layer is on its leftmost, e.g. $\widetilde{\mathbf{X}}_f \in \mathbb{R}^{2M\times T}$.

The STFT coefficient of multi-channel mixture signals are directly taken as the input feature of the network. For one TF bin, the STFT coefficients are concatenated along channels:

$$\mathbf{X}_{f,t} = [\mathrm{X}^1_{f,t}, \ldots, \mathrm{X}^M_{f,t}]^T \in \mathbb{C}^{M\times 1} \tag{1}$$

where ^{T} denotes vector transpose. Then, the time sequence of $\mathbf{X}_{f,t}$ for one frequency is taken as the input sequence of the RNN network:

$$\mathbf{X}_f = (\mathbf{X}_{f,1}, \ldots, \mathbf{X}_{f,T}) \in \mathbb{C}^{M\times T} \tag{2}$$

To facilitate the network training, all input sequences are normalized to have the magnitude mean for the reference channel being one. The normalization is performed as $\mathbf{X}_f/\overline{\mathrm{X}}_f$, where $\overline{\mathrm{X}}_f = \sum_{t=1}^{T} |\mathrm{X}^r_{f,t}|/T$. As the network can only process real numbers, the complex-valued input sequence should be converted into real-valued sequence. This is done by simply replacing the complex number with its real part and imaginary part. The real-valued sequence is denoted by $\widetilde{\mathbf{X}}_f \in R^{2M\times T}$.

The outputs of the network are the separated signals for this frequency. In the literature, the magnitude spectra, TF mask or complex mask are often chosen as the training target [19]. However, it was demonstrated in our previous work [18] that, for the speech denoising task, the STFT coefficients of clean speech can be directly predicted by the multi-channel narrow-band network. We follow and extend this principle in this work to the speech separation task, namely predicting the STFT coefficients of multiple speech signals. More specifically, the output $\widetilde{\mathbf{Y}}_f \in R^{2N\times T}$ consists of the real part

[1] https://github.com/Audio-WestlakeU/NBSS

and imaginary part of the (normalized) spatial image at the reference channel of N speakers.

The STFT coefficients at the original level of each separated speech signal, can be obtained from the network output $\widetilde{\mathbf{Y}}_f$ by first constructing the complex numbers, and then multiplying with the normalization factor: $\widehat{\mathbf{Y}}_f^n = (\widetilde{\mathbf{Y}}_f^{2n-1} + i\widetilde{\mathbf{Y}}_f^{2n})\overline{\mathrm{X}}_f \in \mathbb{C}^{1\times T}, n = 1, \dots, N$.

As shown in Fig.1, this paper uses a network which has two layers of bidirectional LSTM (BiLSTM) and one fully connected (FC) layer. The network is shared by all frequencies, thus each frequency is processed independently.

2.2. Full-band Permutation Invariant Training

The training process of the proposed network has the label permutation problem, which can be solved by the PIT technique. Applying PIT for each frequency separately, although the speech signals can be well separated at each frequency, it still suffers from the frequency permutation problem, as is for other narrow-band methods [2, 3, 4].

To solve the frequency permutation problem, we propose a full-band PIT (fPIT) technique, which forces the network to produce predictions of all frequencies with an identical speaker label permutation. Specifically, the predictions at the same output position of all frequencies, i.e. $\widehat{\mathbf{Y}}^n = [\widehat{\mathbf{Y}}_0^n; \dots; \widehat{\mathbf{Y}}_{F-1}^n] \in \mathbb{C}^{F\times T}$, are forced to belong to the same speaker, and binded together to form the complete spectra of this speaker. Thus, the best permutation for the N bindings can be regarded as the permutation for all frequencies. The loss of all frequencies can then be calculated in a fPIT way:

$$\text{fPIT}(\widehat{\mathbf{Y}}^1, \dots, \widehat{\mathbf{Y}}^N, \mathbf{Y}^1, \dots, \mathbf{Y}^N) = \underset{p\in\mathcal{P}}{min} \frac{1}{N} \sum_n \mathcal{L}(\mathbf{Y}^n, \widehat{\mathbf{Y}}^{p(n)}) \tag{3}$$

where $\mathbf{Y}^n \in \mathbb{C}^{F\times T}$ is a matrix consisting of all the STFT coefficients (of the spatial image at the reference channel) of the n-th speaker, i.e. $\mathrm{Y}_{f,t}^{n,r}, f = 0, \dots, F-1; t = 1, \dots, T$. $\mathcal{P}$ denotes the set of all possible permutations, and p denotes a permutation in $\mathcal{P}$ which maps labels of ground truth to labels of predictions. $\mathcal{L}$ denotes a loss function.

The frequency binding in fPIT requires the network to not only separate the speech signals for each individual frequency, but also output the predictions of all frequencies for one speaker at the same position, even though the network processes frequencies separately. The former task relies on learning narrow-band spatial information to discriminate multiple speakers. The latter task possibly uses the (partial) narrow-band spatial information that are consistent along frequencies to determine the output position, such as the inter-channel cues related to the speaker direction.

fPIT uses the negative SI-SDR [20] as the loss function:

$$\mathcal{L}(\mathbf{Y}^n, \widehat{\mathbf{Y}}^{p(n)}) = -10\log_{10}\frac{\|\alpha\mathbf{y}^n\|^2}{\left\|\alpha\mathbf{y}^n - \widehat{\mathbf{y}}^{p(n)}\right\|^2} \tag{4}$$

where $\alpha = (\widehat{\mathbf{y}}^{p(n)})^T\mathbf{y}^n / \|\mathbf{y}^n\|^2$, $\mathbf{y}^n$ and $\widehat{\mathbf{y}}^{p(n)}$ are the inverse STFT of $\mathbf{Y}^n$ and $\widehat{\mathbf{Y}}^{p(n)}$, respectively.

3. EXPERIMENTS

3.1. Experimental Setup

Dataset. The proposed method is evaluated on a spatialized version of the WSJ0-2mix dataset [6]. The WSJ0-2mix dataset contains 20000, 5000 and 3000 speech pairs with varying lengths for training, validation and test respectively. In this experiment, the speech pairs are overlapped in the manner used in [11], in which the tail part of one signal is overlapped with the head part of the other signal. The overlap ratio is uniformly sampled in the range of [10%, 100%]. The resulting mixed utterances are all set to four-second long. Room impulse responses are simulated using the gpuRIR package [21], which is a GPU based implementation of the image method [22]. The length, width and height of the simulated rooms are uniformly sampled in the range of [3 m, 8 m], [3 m, 8 m] and [3 m, 4 m], respectively. The RT60 of each room is uniformly sampled in [0.1 s, 1.0 s]. An 8-channel circular microphone array with a radius of 5 cm is used. The center of the array is randomly sampled in a square area (length is 1 m) in the center of the room at a height of 1.5 m. The speaker locations are randomly sampled in the room with a height of 1.5 m and with the difference between the direction of two speakers randomly sampled from 0° to 180°. The speakers are located at least 0.5 m away from the walls.

Training Configurations. The sampling rate is 16 kHz. STFT is performed using a hanning window of length 512 samples (32ms) with a hop size of 256 samples. The numbers of hidden units of the first and second BiLSTM layers are set to 256 and 128 respectively. The sizes of the input and output of the FC layer are 256 and 4 respectively. The network is trained with 30 utterances per mini-batch, thus the batch size of frequencies is 7710 (30 × 257). The Adam [23] optimizer is used with an initial learning rate of 0.001. When the validation loss does not decrease in 10 consecutive epochs, the learning rate is halved until it reaches a given minimum of 0.0001. Gradient clipping is applied with a threshold of 5.

Performance Metrics and Comparison Methods. To evaluate the speech separation performance, three metrics are used: i) PESQ [24], ii) SDR [25], iii) SI-SDR [20]. We compare the proposed method with two baselines: i) Oracle MVDR beamformer[2], for which the steering vector of desired speech and the covariance matrix of undesired signals are computed using the true desired speech and undesired signals, respectively; ii) The recently proposed FaSNet with the TAC mechanism [11], referred to as FaSNet-TAC.

[2] https://github.com/Enny1991/beamformers

Table 1: Speech separation results

Model	SDR	SI-SDR	NB-PESQ	WB-PESQ
Mixture	0.18	0.00	2.05	1.6
Oracle MVDR	12.19	11.70	3.21	2.68
FaSNet-TAC [11]	12.81	12.26	2.92	2.49
prop. with corr	12.59	11.09	3.14	2.68
prop.	**13.89**	**13.26**	**3.31**	**2.87**

3.2. Experimental Results

Table 1 shows the results. Compared with the oracle MVDR, FaSNet-TAC achieves higher SDR and SI-SDR scores, but lower NB-PESQ and WB-PESQ scores. This means that FasNet-TAC is able to better separate the speech signals relying on the capability of the DNN, but the separated signals are possibly distorted and have a worse perceptual quality. The proposed method achieves the best performance in terms of all metrics. This indicates that narrow-band indeed involves rich information to discriminate between speakers, and the proposed network is able to exploit such information to well separate the signals. The possible reason for the high PESQ scores and low speech distortion is that the proposed narrow-band method can better leverage the spatial constraints of speakers to preserve the signals, as is also done by MVDR. The difference of perceptual quality between different methods is audible when listening to the separated sounds. Some sound examples are available in our webpage [3].

To evaluate the effectiveness of fPIT, the experiment using the correlation method in [14] for solving the frequency permutation problem is also conducted, in which fPIT is not used. The speech separation results are shown in the line of "prop. with corr" in Table 1. Without fPIT, all performance measures significantly drop, which shows the effectiveness of fPIT for solving the frequency permutation problem.

Fig. 2 shows the performance improvement as a function of RT60, difference between speaker directions and speech overlap ratio. Not surprisingly, the performance of all three methods degrades with the increasing of RT60, since the reverberation distorts the spatial cues used for separation. Regrading the difference between speaker directions, FaSNet-TAC and the proposed method achieve better performance when the difference increases. A larger direction difference leads to a larger difference between the spatial cues of speakers, relying on which FaSNet-TAC and the proposed method separate the speakers. By contrast, the influence of the direction difference for MVDR depends on array's beampattern.

Along with the increase of speech overlap ratio, the performance of all the methods degrade, since non-overlapped signal is easier to separate than overlapped signal. Moreover, in the non-overlapped signals, each speaker presents solely, which provides some useful information, such as the speaker embeddings and the clean spatial cues, that can be used for separating the overlapped signals. The performance of the proposed method significantly drops when the overlap ratio gets extremely high (higher than 95%). This indicates that the proposed method highly relies on the non-overlapped part to extract some information about the clean spatial cues, and much information can be provided even by a small portion (larger than 5%) of non-overlapped signals.

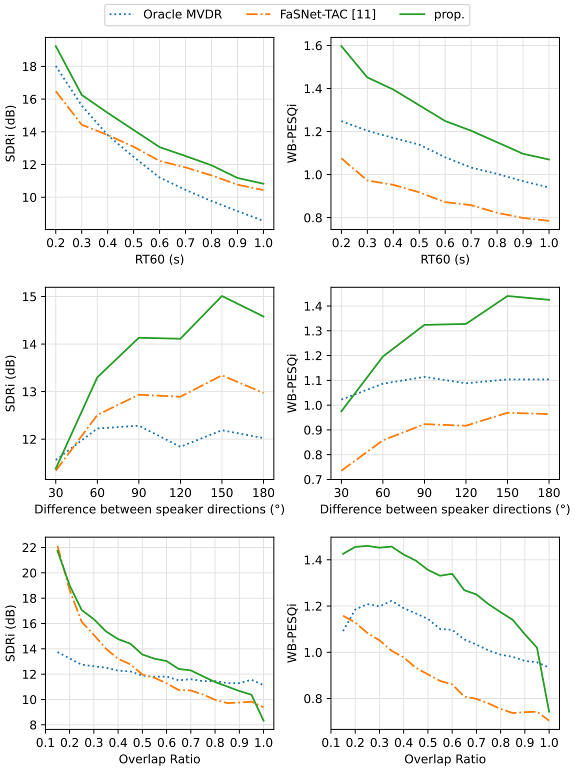


Fig. 2: The influence of (a) RT60, (b) difference between speaker directions, and (c) speech overlap ratio in terms of SDRi (SDR improvement) and WB-PESQi (WB-PESQ improvement)

4. CONCLUSION

In this paper, we proposed a narrow-band multi-channel speech separation network, which takes a single frequency band of STFT coefficients as input, and outputs the separated STFT coefficients of the same frequency. The proposed network is trained with fPIT in an end-to-end way. This work is the first deep learning based work exploring speech separation in a frequency by frequency fashion as many traditional speech separation methods do. The experimental results show the superiority of the proposed method under most of the considered experimental conditions. It is worth to note that the narrow-band information is hopefully complementary to the full-band information like spectral patterns of speech. In the future, we will investigate how to fuse them for better speech separation.

[3] https://quancs.github.io/blog/nbss/

5. REFERENCES

[1] Ozgur Yilmaz and Scott Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[2] Stefan Winter, Walter Kellermann, Hiroshi Sawada, and Shoji Makino, "MAP-Based Underdetermined Blind Source Separation of Convolutive Mixtures by Hierarchical Clustering and l_1-Norm Minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–12, Dec. 2006.

[3] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[4] Christoph Boeddecker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME*, Sept. 2018, pp. 35–40.

[5] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, and others, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[6] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*, Shanghai, Mar. 2016, pp. 31–35.

[7] Dong Yu, Morten Kolbaek, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, New Orleans, LA, Mar. 2017, pp. 241–245.

[8] Zhong-Qiu Wang, Jonathan Le Roux, and John R. Hershey, "Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation," in *ICASSP*, Calgary, AB, Apr. 2018, pp. 1–5.

[9] Rongzhi Gu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu, "Enhancing End-to-End Multi-Channel Speech Separation Via Spatial Feature Learning," in *ICASSP*, Barcelona, Spain, May 2020, pp. 7319–7323.

[10] Yi Luo, Cong Han, Nima Mesgarani, Enea Ceolini, and Shih-Chii Liu, "FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing," in *ASRU*, Singapore, 2019, pp. 260–267.

[11] Yi Luo, Zhuo Chen, Nima Mesgarani, and Takuya Yoshioka, "End-to-end Microphone Permutation and Number Invariant Multi-channel Speech Separation," in *ICASSP*, Barcelona, Spain, May 2020, pp. 6394–6398.

[12] Tsubasa Ochiai, Marc Delcroix, Rintaro Ikeshita, Keisuke Kinoshita, Tomohiro Nakatani, and Shoko Araki, "Beam-TasNet: Time-domain Audio Separation Network Meets Frequency-domain Beamformer," in *ICASSP*, Barcelona, Spain, May 2020, pp. 6384–6388.

[13] Jahn Heymann, Lukas Drude, Aleksej Chinaev, and Reinhold Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3RD CHiME challenge," in *ASRU*, 2015, pp. 444–451.

[14] Hiroshi Sawada, Ryo Mukai, Shoko Araki, and Shoji Makino, "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[15] Radoslaw Mazur and Alfred Mertins, "An Approach for Solving the Permutation Problem of Convolutive Blind Source Separation Based on Statistical Signal Models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 117–126, Jan. 2009.

[16] Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[17] Xiaofei Li and Radu Horaud, "Multichannel Speech Enhancement Based On Time-Frequency Masking Using Subband Long Short-Term Memory," in *WASPAA*, New Paltz, NY, USA, Oct. 2019, pp. 298–302.

[18] Xiaofei Li and Radu Horaud, "Narrow-band Deep Filtering for Multichannel Speech Enhancement," *arXiv preprint arXiv:1911.10791*, 2019.

[19] DeLiang Wang and Jitong 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, Oct. 2018.

[20] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR – Half-baked or Well Done?," in *ICASSP*, Brighton, United Kingdom, May 2019, pp. 626–630.

[21] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, Feb. 2021.

[22] Jont B. Allen and David 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, Apr. 1979.

[23] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[24] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.

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