MULTI-CHANNEL ITAKURA SAITO DISTANCE MINIMIZATION WITH DEEP NEURAL NETWORK

Masahito Togami

LINE Corporation

ABSTRACT

A multi-channel speech source separation with a deep neural network which optimizes not only the time-varying variance of a speech source but also the multi-channel spatial covariance matrix jointly without any iterative optimization method is shown. Instead of a loss function which does not evaluate spatial characteristics of the output signal, the proposed method utilizes a loss function based on minimization of multi-channel Itakura-Saito Distance (MISD), which evaluates spatial characteristics of the output signal. The cost function based on MISD is calculated by the estimated posterior probability density function (PDF) of each speech source based on a time-varying Gaussian distribution model. The loss function of the neural network and the PDF of each speech source that is assumed in multi-channel speech source separation are consistent with each other. As a neural-network architecture, the proposed method utilizes multiple bidirectional long-short term memory (BLSTM) layers. The BLSTM layers and the successive complex-valued signal processing are jointly optimized in the training phase. Experimental results show that more accurately separated speech signal can be obtained with neural network parameters optimized based on the proposed MISD minimization than that with neural network parameters optimized based on loss functions without spatial covariance matrix evaluation.

Index Terms— Deep Learning, speech source separation, bidirectional long-short term memory, multi-channel Itakura-Saito distance

1. INTRODUCTION

Speech source separation techniques which separates multiple speech sources from multiple mixtures [1] are effective in human-listening devices and automatic speech recognition under multi-talkers conditions. In the speech source separation research field, multi-microphone based blind speech source separation in time-frequency domain has been actively studied such as independent component analysis (ICA) [2], sparseness based method [3], and local Gaussian modeling (LGM) based method [4]. These separation methods utilize predetermined signal models such as super-Gaussian models of speech sources, sparseness models at each time-frequency point, or Gaussian distribution models with time-varying covariance matrices. These approaches successfully separate speech sources in each frequency bin separately. However, additional models for frequency characteristics of speech sources is needed to solve the well-known inter-frequency permutation ambiguity problem [5]. Recently, blind speech source separation techniques which does not require for permutation problem solvers have been studied, e.g., Independent vector analysis (IVA) [6, 7]. However, IVA also requires for predetermined generative models for frequency characteristics of speech sources such as a spherically symmetric multivariate distribution, non-negative matrix factorization (NMF) models [8]. However, the conventional models are too simple to express frequency characteristics of speech sources precisely.

Recently, neural network based noise reduction techniques have been actively studied [9, 10, 11, 12, 13]. These approaches estimate time-frequency masks like DUET [3]. Instead of the predetermined speech source models, precise frequency characteristics of speech sources are learned via a neural network. Multi-channel beamformers with time-frequency masks learned by neural networks have been also proposed [11, 14, 13, 15, 16]. However, these techniques are noise reduction techniques, which assumes that there are only one speech source and background noise. It is difficult to apply these techniques for speech source separation directly.

Recently, speech source separation methods with neural network masking have been proposed, e.g., permutation invariant training (PIT) [17] , deep clustering [12, 18]. From the power spectral of the microphone input signal, a time frequency mask which extracts each source is estimated. Time-frequency masking based speech source separation methods assume that multiple speech sources are sparse enough in time-frequency domain. However, the sparseness assumption is not always valid, and it produces unwanted speech distortion in the output signal. Another neural network based speech source separation techniques are extensions of conventional speech source separation methods such as ILMA [8], LGM [4]. Instead of predetermined generative models for speech sources, neural-network based generative models are utilized, e.g. IDLMA [19], LGM with neural network [10], and multi-channel variational auto-encoder [20, 21]. However, in these methods, the neural-network is utilized for only estimation of time-varying variance of speech sources. Multi-channel covariance matrices are needed to be estimated in the other ways. Therefore, iterative update procedures such as IP algorithm [22], expectation-maximization (EM) algorithm [23] are still needed in these frameworks for optimization of multi-channel covariance matrices.

In this paper, a multi-channel speech source separation method with multi-channel Itakura-Saito Distance (MISD) minimization criteria is proposed. The proposed loss function evaluates not only the time-varying variance of speech sources but also the spatial covariance matrix. It is not needed to utilize additional iterative update procedure for optimization of the multi-channel covariance matrices. The multi-channel covariance matrices are estimated by using time-frequency masks which is estimated via a neural-network. The time-varying variance of speech sources is also estimated via the same neural network. The posterior probability density function (PDF) of the speech sources is estimated via a time-varying multi-channel Wiener filtering under the assumption that a PDF of a multi-channel speech source signal is a LGM with a zero-mean vector and a time-varying multi-channel covariance matrix [4]. The likelihood function of the LGM is known to be equivalent with the MISD [24]. Therefore, the likelihood function of speech source separation are

consistent with the proposed loss function for neural network optimization. Inspired by success of the deep clustering [12, 18], the proposed method utilizes multiple bidirectional long-short term memory (BLSTM) layers as the neural-network in the proposed method. The BLSTM layers and the successive TV-MWF are jointly optimized so as to minimize the MISD.

2. PROBLEM STATEMENT

2.1. Microphone input model

In this paper, multi-channel speech source separation is performed at time-frequency domain. The microphone input signal at time-frequency domain is defined as follows:

$$\boldsymbol{x}_{l,k} = \sum_{i=1}^{N_s} \boldsymbol{c}_{i,l,k}, \tag{1}$$

where $\boldsymbol{x}_{l,k}$ (l is the frame index and k is the frequency index) is the multi-channel microphone input signal at each time-frequency point, the number of the microphones is N_m, N_s is the number of the speech sources, and $\boldsymbol{c}_{i,l,k}$ is the ith speech signal. The objective of multi-channel speech source separation is to separate $\boldsymbol{c}_{i,l,k}$ from the microphone input signal, $\boldsymbol{x}_{l,k}$.

2.2. Speech source separation based on local Gaussian modeling

Local Gaussian modeling (LGM) based speech source separation methods [4] separate multiple speech sources under the assumption that a prior probability density function (PDF) of each speech source belongs to a time-varying Gaussian distribution with a zero-mean vector and a time-varying covariance matrix as follows:

$$p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}) = \mathcal{N}(\boldsymbol{c}_{i,l,k}|\boldsymbol{0}, \boldsymbol{R}_{i,l,k}), \tag{2}$$

where $\mathcal{M}$ is set to the predefined model parameter and $\boldsymbol{R}_{i,l,k}$ is the time-varying multi-channel covariance matrix of the ith speech source. So as to reduce the number of the time-varying parameters, the time-varying covariance matrix $\boldsymbol{R}_{i,l,k}$ is defined as follows:

$$\boldsymbol{R}_{i,l,k} = v_{i,l,k}\boldsymbol{R}_{i,k}, \tag{3}$$

where $v_{i,l,k}$ is the time-varying variance of the ith speech source, $\boldsymbol{R}_{i,k}$ is the multi-channel covariance matrix of the ith speech source, and $\mathcal{M}$ is defined as $\{v_{i,l,k}, \boldsymbol{R}_{i,k}\}$.

The LGM based approaches estimate the posterior PDF of the speech source $p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$ with the given microphone input signal. Under the LGM assumption, $p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$ is also a time-varying multi-channel Gaussian distribution defined as follows:

$$p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k}) = \mathcal{N}(\boldsymbol{c}_{i,l,k}|\boldsymbol{\mu}_{i,l,k}, \boldsymbol{V}_{i,l,k}), \tag{4}$$

where $\boldsymbol{\mu}_{i,l,k}$ and $\boldsymbol{V}_{i,l,k}$ are the conditional mean vector and the conditional covariance matrix of $\boldsymbol{c}_{i,l,k}$, respectively. $\boldsymbol{\mu}_{i,l,k}$ and $\boldsymbol{V}_{i,l,k}$ are calculated as follows:

$$\boldsymbol{\mu}_{i,l,k} = \boldsymbol{W}_{i,l,k}\boldsymbol{x}_{i,l,k}, \tag{5}$$

$$\boldsymbol{V}_{i,l,k} = (\boldsymbol{I} - \boldsymbol{W}_{i,l,k})\boldsymbol{R}_{i,l,k}, \tag{6}$$

where $\boldsymbol{I}$ is a $N_m \times N_m$ identity matrix and $\boldsymbol{W}_{i,l,k}$ is the multi-channel Wiener filter which is defined as follows:

$$\boldsymbol{W}_{i,l,k} = \boldsymbol{R}_{i,l,k}\left(\sum_{i=0}^{N_s-1} \boldsymbol{R}_{i,l,k}\right)^{-1}, \tag{7}$$

Therefore, the posterior PDF can be calculated by estimating the prior PDF parameters, i.e., $v_{i,l,k}$ and $\boldsymbol{R}_{i,k}$.

2.3. Iterative parameter optimization

In the conventional method [4, 24], $v_{i,l,k}$ and $\boldsymbol{R}_{i,k}$ are estimated in an iterative way. $v_{i,l,k}$ and $\boldsymbol{R}_{i,k}$ are iteratively updated which assures that the cost function of each frequency bin decreases monotonically. To solve the well-known inter-frequency permutation ambiguity problem [5], it is needed to utilize a frequency characteristics model of speech sources such as non-negative matrix factorization [24]. However, the conventional models are too simple to express precise frequency characteristics of speech sources. Recently, neural-network based frequency characteristics models have been proposed [10, 19, 20, 21]. In these neural-network methods, estimation of the time-varying speech source variance $v_{i,l,k}$ is replaced with a neural-network based method. However, estimation of multi-channel covariance matrices is still needed by using an iterative way such as the EM algorithm.

3. PROPOSED METHOD

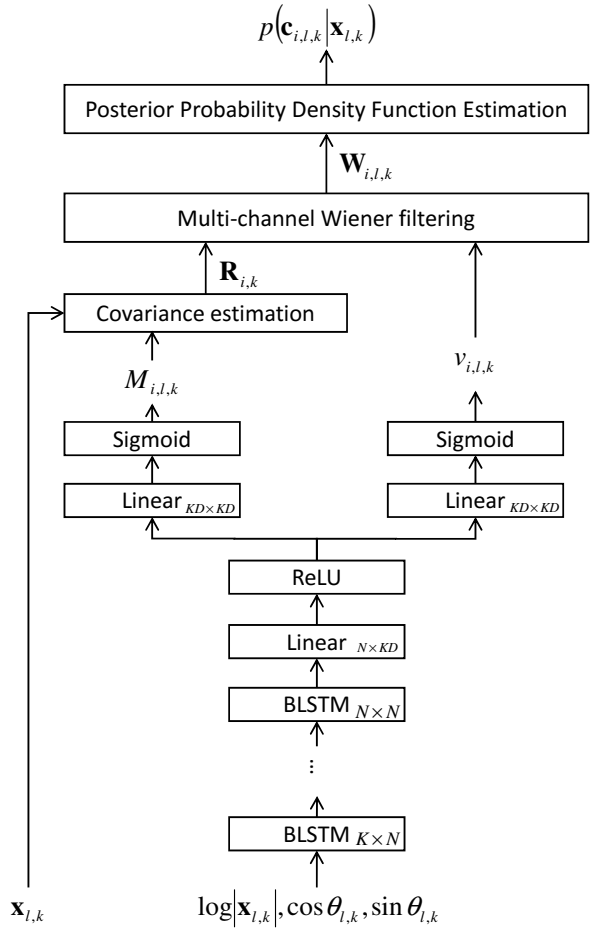

Fig. 1. Block diagram of proposed method

3.1. Overview of the proposed method

Overview of the block diagram of the proposed method is shown in Fig. 1. The proposed architecture estimates both the covariance matrix $\boldsymbol{R}_{i,k}$ and the time-varying activity of the speech sources $v_{i,l,k}$ via a neural network. The covariance matrix is updated with a time-frequency mask estimated by the neural network. Inspired by effectiveness of phase difference between microphones as an input feature in the multi-channel deep clustering [18], the proposed method also utilizes the phase difference between microphones as an input feature. In the same way as the original LGM [4], the proposed method estimates the posterior PDF of the clean speech source $p(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$ via a time-varying multi-channel Wiener filter with the estimated time-frequency mask and the estimated time-varying activity. Therefore, the posterior PDF is estimated via a real-valued neural network and a complex-valued signal proessing. In the training stage, parameters of the neural network are trained so as to minimize a loss function which is defined as the multi-channel Itakura-Saito distance (MISD). Instead of evaluating only the time-varying variance $v_{i,l,k}$, the proposed method evaluates both $v_{i,l,k}$ and $\boldsymbol{R}_{i,k}$. Therefore, both $v_{i,l,k}$ and $\boldsymbol{R}_{i,k}$ are jointly optimized in the proposed neural network structure. It is not needed to utilize an iterative optimization method such as the EM algorithm for the covariance matrix optimization. The MISD is known to be equivalent with the log-likelihood function of the time-varying Gaussian distribution, $\log p(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$. Therefore, in the proposed method, there is consistency between the loss function for the neural network optimization and the PDFs of the speech sources.

3.2. Proposed multi-channel loss function

Let $\mathcal{M}$ be the parameter of the neural network. The loss function of the proposed method for optimization of the model parameter $\mathcal{M}$ is defined as the following MISD [24]:

$$\mathcal{L}(\mathcal{M}) = \sum_{i,l,k} (\boldsymbol{c}_{i,l,k} - \boldsymbol{\mu}_{i,l,k})^H \boldsymbol{V}_{i,l,k}^{-1} (\boldsymbol{c}_{i,l,k} - \boldsymbol{\mu}_{i,l,k}) + \log |\boldsymbol{V}_{i,l,k}|. \tag{8}$$

This loss function is known to be equivalent with the log-likelihood function of the time-varying Gaussian distribution [24]. Therefore, the loss function can be calculated by estimating the posterior PDF of a clean speech source.

In the training phase, the error is back-propagated through not only $\boldsymbol{\mu}_{i,l,k}$ but also $\boldsymbol{V}_{i,l,k}$. In the first term, $\boldsymbol{V}_{i,l,k}$ acts as a time-frequency weight which normalizes the error in each time-frequency bin. The second term acts as a regularization term for a time-frequency weight. When there is reverberation, the steering vector of each source is not stationary. Therefore, there is difference between $\boldsymbol{c}_{i,l,k}$ and $\boldsymbol{\mu}_{i,l,k}$. In this case, $\boldsymbol{V}_{i,l,k}$ will be learned to reflect the amount of variance of the difference, and $\boldsymbol{R}_{i,k}$ can be trained so as to reflect the reverberation effect.

To remove permutation ambiguity, the proposed method calculates the loss function with a permutation matrix which minimizes the loss function [17]. The best permutation matrix is estimated in each parameter update step.

3.3. Parameter estimation of probability density function

The proposed method calculates the posterior PDF of the ith clean speech source, $p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$, based on the LGM. Both $v_{i,l,k}$ and $\boldsymbol{R}_{i,k}$ are estimated with no iterative way. The covariance matrix $\boldsymbol{R}_{i,k}$ is estimated like mask-based beamforming techniques [13] as follows:

$$\boldsymbol{R}_{i,k} = \sum_{l} M_{i,l,k} \boldsymbol{x}_{l,k} \boldsymbol{x}_{l,k}^H. \tag{9}$$

Both the time-frequency mask $M_{i,l,k}$ and the time-varying activity $v_{i,l,k}$ are real-valued variables, and these variables are estimated via a real-valued deep neural network. Under the assumption that location of each speech source is stationary within each utterance, spatial information is effective for the time-frequency masks estimation [18]. The proposed method utilizes the phase difference between microphones, $\theta_{l,k}$, as one of input features of the deep neural network in addition to the amplitude spectral feature, $\log |\boldsymbol{x}_{l,k}|$.

4. EVALUATION

4.1. Setup

Speech source separation performance of the proposed method was evaluated. The dataset was made by convolving measured impulse response in Multi-channel Impulse Response Database (MIRD) [25] with the clean speech sources in TIMIT speech corpus [26].

In the training phase, TIMIT train corpus was utilized. In the evaluation phase, TIMIT test corpus was utilized. Related to impulse responses, the reverberation time RT_{60} was set to 0.16 [sec]. The number of the microphone was set to 2. The number of the speech sources was set to 2 in each sample. Two microphone indeces were randomly selected for each sample both in the training phase and in the evaluation phase. In the training phase, a 3-3-3-8-3-3-3 spacing (cm) microphone array was utilized. In the evaluation phase, a 4-4-4-8-4-4-4 spacing (cm) microphone array was utilized. Therefore, a different microphone array was utilized in the evaluation phase from the training phase. Sampling rate was set to 8000 Hz. Frame size was 256 pt. Frame shift was 64 pt. The number of frequency bins was 129. The distance between speech sources and microphones was set to 1 m. Azimuth of each talker is randomly selected for each utterance. The number of total training utterances was 2000. Mini-batch size was set to 128. Each utterance was splitted in every 100-frames segment. Therefore, length of each data was 100 (frame).

4.2. Neural network architecture

The number of BLSTM layers was set to 4. The number of the units in each BLSTM layer was set to 600. Dropout was utilized in both BLSTM and dense layers in the training phase. For dense layers, batch normalization was utilized. Neural network parameters were updated by 2400 and 4800 times. Adam otpimizer (learning rate was 0.001) with gradient clipping was utilized. The proposed architecture contains complex-valued gradient calculation. Tensorflow [27] was utilized for complex-valued gradient calculation.

4.3. Evaluation measure

Evaluation measures were set to SIR, SDR, Mel Frequency Cepstrum Coefficients (MFCC) distance improvement , and segmental Signal-to-Noise Ratio (seg. SNR). SIR and SDR were calculated by using BSS_EVAL [28]. MFCC distance improvement, ΔMFCC is defined as MFCC$_{input}$ $-$ MFCC$_{output}$. MFCC$_{input}$ is the MFCC distance between the clean speech source and the microphone input signal and MFCC$_{output}$ is the MFCC distance between the clean speech source and the estimated speech source. The dimension of MFCC was set to 13. The seg. SNR is defined as follows:

$$\text{seg. SNR} = \frac{1}{L} \sum_{\tau=0}^{L} -10 \log_{10} \frac{\sum_{p=0}^{P-1} \|s_{P\tau+p}\|^2}{\sum_{p=0}^{P-1} \|s_{P\tau+p} - \hat{s}_{P\tau+p}\|^2}, \tag{10}$$

Table 1. Evaluation results of two-channel speech source separation

Approaches	$N_{iter} = 2400$				$N_{iter} = 4800$			
	SIR (dB)	SDR (dB)	ΔMFCC	seg. SNR (dB)	SIR (dB)	SDR (dB)	ΔMFCC	seg. SNR (dB)
l_2 loss	10.54	8.54	3.50	5.47	11.99	9.46	3.94	6.57
Single-channel IS	1.03	0.16	0.52	0.80	1.39	0.05	0.34	1.06
MMSE (Full)	9.76	7.94	3.24	5.14	10.98	8.88	3.66	5.92
MMSE (Diag)	9.38	7.67	3.13	4.91	10.71	8.47	3.46	5.73
LGM Prior	10.34	8.59	**4.39**	6.06	11.34	9.13	**4.69**	6.84
LGM Posterior (Diag)	11.27	9.29	4.19	6.78	10.80	9.10	4.37	6.44
Proposed method	**11.57**	**9.65**	4.36	**6.96**	**12.16**	**10.17**	4.66	**7.40**

where s_t is the clean speech source in time domain, $\hat{s}_t$ is the estimated one, L is the length of time-segments, and P is the length of each segment. P was set to 512. Each evaluation results were calculated as average of 1000 utterances.

4.4. Comparative loss functions

The proposed method with the MISD based cost function ($L_{MIS} = \mathcal{L}(\mathcal{M})$) which is defined in Eq. 8 was compared with the following six methods based on a different cost function:

- l_2 loss function:

$$L_{l_2} = \sum_{i,l,k} \|\boldsymbol{c}_{i,l,k} - \boldsymbol{\mu}_{i,l,k}\|^2. \tag{11}$$

- Single-channel Itakura-Saito (IS) distance:

$$L_{SIS} = \sum_{i,l,k} \frac{p_{i,l,k}}{y_{i,l,k}} - \log \frac{p_{i,l,k}}{y_{i,l,k}} - 1, \tag{12}$$

  where $p_{i,l,k}$ is the actual power spectral of the ith speech source and $y_{i,l,k}$ is the estimated one, $\|\boldsymbol{\mu}_{i,l,k}\|$. This cost function is commonly utilized for time-frequency activity estimation in conventional neural-network based blind source separation methods. To improve estimation accuracy of the spatial covariance matrices, the conventional methods utilizes iterative ways in pararel [19, 10].

- MMSE (Full):

$$\begin{aligned} L_{MMSE} &= \sum_{i,l,k} E[\|\boldsymbol{c}_{i,l,k} - \hat{c}_{i,l,k}\|\|^2]_{p_{\mathcal{M}}(\hat{c}_{i,l,k}|\boldsymbol{x}_{l,k})}, \\ &= \sum_{i,l,k} \|\boldsymbol{c}_{i,l,k} - \boldsymbol{\mu}_{i,l,k}\|^2 + \mathrm{tr}(\boldsymbol{V}_{i,l,k}). \end{aligned} \tag{13}$$

  In this equation, the second term, $\mathrm{tr}(\boldsymbol{V}_{i,l,k})$, is also regarded as one of regularizers. However, on contrary to L_{MIS}, the regularizer does not affect the l_2 loss term directly in the MMSE based cost function case.

- MMSE (Diag): The covariance matrix of the posterior PDF of a speech source is approximated as a diagonal matrix as $\boldsymbol{V}_{i,l,k} \approx \mathrm{diag}(\boldsymbol{V}_{i,l,k})$ in the MMSE loss function.

- LGM Prior: Instead of estimating $p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$, $p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k})$ is utilized in the proposed loss function.

- LGM Posterior (Diag): The covariance matrix of the posterior PDF of a speech source is approximated as a diagonal matrix as $\boldsymbol{V}_{i,l,k} \approx \mathrm{diag}(\boldsymbol{V}_{i,l,k})$ in the proposed loss function.

4.5. Experimental results

Experimental results are shown in Table 1 for the number of the iterations $N_{iter} = 2400, 4800$. It is shown that the proposed method achieved the best performance except for ΔMFCC. From comparision between the proposed method and l_2 loss, it can be said that the estimated covariance matrix of the posterior PDF of a speech source works well as a regularization term. Additionally, when the number of N_{iter} increases, SIR was improved in l_2 loss. However, speech distortion was improved less than the proposed method.

It is also shown that evaluation of spatial information in the loss function is effective by comparing the proposed method with LGM Posterior (Diag). Single-channel IS does not separate speech sources sufficiently, because in this method, only the amplitude spectral of speech sources are evaluated, and spatial information is not evaluated. Therefore, when there is no iterative way, it is highly difficult to separate speech sources with the single-channel IS based loss function. In MMSE cases, the loss function is inconsistent with the probabilistic model of each speech source. From comparison of the proposed method with MMSE cases, it can be said that a loss function which is consistent with a probabilistic model of a speech source increases speech source separation performance.

5. CONCLUSION

In this paper, a deep neural network based multi-channel speech source separation technique was proposed. The loss function of the proposed method is based on minimization of Multi-channel Itakura-Saito Distance (MISD). The proposed loss function is consisitent with the probability density function of each speech source that is assumed in multi-channel speech source separation. Not only the time-varying variance of a speech source but also the spatial covariance matrix of a speech source can be jointly estimated in the proposed framework. Therefore, additional iteration is not required in the proposed method so as to estimate the spatial covariance matrix. Experimental results showed that the proposed method with the MISD loss function can separate two speech sources more clearly than the other methods.

6. REFERENCES

[1] S. Makino, T.W. Lee, and H. Sawada, *Blind Speech Separation*, Springer Publishing Company, Incorporated, 2007.

[2] P. Common, "Independent component analysis, a new concept ?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mix-

tures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[4] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[5] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, March 2011.

[6] A. Hiroe, "Solution of permutation problem in frequency domain ica using multivariate probability density functions," in *Proceedings ICA*, Mar. 2006, pp. 601–608.

[7] T. Kim, H.T. Attias, S.-Y. Lee, and T.-W. Lee, "Independent vector analysis: an extension of ica to multivariate components," in *Proceedings ICA*, Mar. 2006, pp. 165–172.

[8] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, *Determined Blind Source separation with Independent Low-Rank Matrix Analysis*, chapter 6, pp. 125–155, Springer Publishing Company, Incorporated, 2018.

[9] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 116–120.

[10] A.A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[11] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

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

[13] H. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 531–535.

[14] Y. Zhou and Y. Qian, "Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 536–540.

[15] Z. Wang and D. Wang, "Mask weighted stft ratios for relative transfer function estimation and its application to robust asr," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5619–5623.

[16] Y. Liu, A. Ganguly, K. Kamath, and T. Kristijansson, "Neural network based time-frequency masking and steering vector estimation for two-channel mvdr beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 6717–6721.

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

[18] 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)*, 2018, pp. 1–5.

[19] D. Kitamura N. Takamune S. Takamichi H. Saruwatari S. Mogami, H. Sumino and N. Ono, "Independent deeply learned matrix analysis for multichannel audio source separation," in *18th European Signal Processing Conference (EUSIPCO 2018)*, Sep. 2018, pp. 1571–1575.

[20] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," 2018.

[21] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," 2018.

[22] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2011, pp. 189–192.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[24] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, May 2013.

[25] "Multi-Channel Impulse Response Database," https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[27] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y.Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[28] 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.