SPATIAL CONSTRAINT ON MULTI-CHANNEL DEEP CLUSTERING

Masahito Togami

LINE Corporation

ABSTRACT

In this paper, a multi-channel deep clustering technique which combines two types of spatial information is proposed. The first one is an estimated direction-of-arrival (DOA) at each time-frequency point, which is utilized as an input feature of the proposed neural network. Instead of stacking embeddings of all pairs of microphones as in the conventional multi-channel deep clustering, the proposed method only requires one embedding. Therefore, the computational cost can be reduced in the inference stage. The second one is the time-frequency activity of each speech source estimated by multi-channel Wiener filtering (MWF). The MWF is inserted between two consecutive bidirectional long-short-term memory (BLSTM) layers. The estimated time-frequency activity of each speech source by the MWF is transformed into an input feature of the next BLSTM layer. The proposed MWF insertion enhances the consistency of the embedding vectors along the time-axis. Experimental results show that multi-channel deep clustering with the proposed input feature based on the estimated DOA can separate speech sources better than the conventional multi-channel deep clustering that stacks embeddings of all the pairs of the microphones. Furthermore, the proposed MWF insertion is shown to be able to reduce distortion of output signal and improve signal-to-interference ratio.

Index Terms— Deep Clustering, DOA estimation, multi-channel Wiener filtering, bidirectional long-short-term memory

1. INTRODUCTION

Blind source separation has been actively studied for a long time [1, 2, 3, 4, 5] so as to enhance speech quality in recording systems and to improve automatic speech recognition performance. Typically, blind source separation is performed in the time-frequency domain, because the mixing process can be approximated as an instantaneous mixture in the time-frequency domain. However, an inter-frequency permutation ambiguity problem occurs, and post permutation solvers are required in blind source separation in time-frequency domain [4].

Independent vector analysis (IVA) techniques which do not require any post permutation solver [6, 7, 8] have been studied. IVA utilizes a spherically symmetric multivariate distribution of a speech source, in which a speech source is assumed to have the same variance at each frequency bin. However, the assumed symmetric multivariate distribution model in the IVA is too simple to express complicated frequency characteristics of speech sources, e.g., harmonic structures. Therefore, speech source separation which is integrated with more precise spectral characteristics of speech sources is highly required.

As an alternative of IVA, multi-channel non-negative matrix factorization (MNMF) based methods have been actively studied [9, 10, 11, 12, 13]. MNMF based approaches approximate the variance of each source as the product of two non-negative matrices. The MNMF approximation is appropriate for power spectral of music sources, but it is difficult to express spectral characteristics of speech sources accurately.

Recently, deep neural network based modeling techniques have been studied as a modeling technique of complicated spectral characteristics of speech sources, e.g., an integration with local Gaussian model [14], extension of ILMA [12] with a neural network, IDLMA [15], mask-based beamforming [16, 17], and auto-encoder based methods [18, 19]. The deep learning based techniques can learn complicated power spectral of speech sources. However, the spatial information is not fully utilized in the neural network, and speech source separation is loosely coupled with the neural network.

Deep clustering (DC) is another category of the neural network based blind source separation techniques [20, 21]. DC estimates a non-linear embedding vector at each time-frequency point with multiple bidirectional long-short-term memory (BLSTM) layers. A time-frequency mask which extracts each speech source is obtained by K-means clustering after estimating the embedding vectors. The original DC was utilized for single-channel speech source separation. Multi-channel extension of deep clustering has been also proposed, multi-channel deep clustering (MDC) [21]. In MDC, estimation accuracy of the embedding vectors can be improved by utilizing phase difference between two microphones as an input feature. By stacking embedding vectors of all the pairs of the microphones, speech separation performance improves in proportional to the number of microphones, N_m. However, $\frac{N_m(N_m-1)}{2}$ more forward passes of the neural network is needed for each sample. Additionally, spatial information of the phase difference between two microphones actually is noisy and more reliable spatial information is required.

In this paper, a multi-channel deep clustering technique which combines two types of spatial information is proposed. The first one is the estimated direction-of-arrival (DOA) at each time-frequency point. DOA information is more reliable than the phase difference between two microphones, because DOA can be estimated by integrating all the microphones. From computational cost perspective, the input feature based on DOA estimation is preferable, because multiple forward passes of the neural network can be removed. Regardless of the number of the microphones, the proposed method performs only one-time forward calculation for each sample with the estimated DOA as an input feature. The second spatial information is the time-frequency activity of each speech source estimated by the multi-channel Wiener filtering (MWF). The MWF is inserted between two consecutive BLSTM layers. The estimated time-frequency activity of each speech source is transformed into the input feature of the next BLSTM layer. The dth input feature represents the time-frequency activity of the same speech source along time-axis. Therefore, the proposed MWF insertion can enhance consistency of the embedding vector along the time-axis. From another perspective, the multi-channel speech source separation is tightly coupled with the neural network in the proposed method. The proposed method optimizes the neural network parameters based on a

cost function which is calculated through the multi-channel speech source separation part. Experimental results show that the proposed method can separate speech sources with less distortion and less computational cost than the conventional MDC.

2. PROBLEM STATEMENT

2.1. Microphone input signal model

In this paper, speech source separation problems in the time-frequency domain are discussed. The microphone input signal is defined in the time-frequency domain as follows:

$$\boldsymbol{x}_{l,k} = \sum_{i=0}^{N_s-1} \boldsymbol{s}_{i,l,k}, \tag{1}$$

where $\boldsymbol{x}_{l,k} \in \mathbb{C}^{N_m}$ (l is the frame index and k is the frequency index) is the multi-channel microphone input signal (N_m is the number of the microphones), N_s is the number of the sources, and $\boldsymbol{s}_{i,l,k}$ is the ith speech source signal. The objective of the speech source separation is to extract $\boldsymbol{s}_{i,l,k}$ from the microphone input signal $\boldsymbol{x}_{l,k}$.

2.2. Conventional multi-channel deep clustering

The conventional deep clustering (DC) based methods [20, 21] separate multiple speech sources by clustering non-linear embedding vectors estimated at each time-frequency bin. The non-linear embedding vectors are estimated via a neural network. The cost function for parameter optimization of the neural network is defined as follows:

$$\mathcal{L}_{DC}(V,Y) = \|\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{Y}\boldsymbol{Y}^T\|_F^2, \tag{2}$$

where T is the transpose operator of a matrix/vector, $\|\cdot\|_F$is the Frobenius norm, $\boldsymbol{V} \in \mathbb{R}^{L_T K \times D}$ (L_T , K, and D are the length of time-frames, the size of frequency bins, and the dimension of each embedding vector, respectively) is the output embedding matrix that contains the embedding vector at each time-frequency bin, and $\boldsymbol{Y} \in \mathbb{R}^{L_T K \times N_s}$ is the target matrix. Only one element of each row of $\boldsymbol{Y}$ takes 1, and the other elements take 0. If the sth speech source is the dominant speech source at the frame l and the frequency k, $Y_{lk,s} = 1$, otherwise, $Y_{lk,s} = 0$. The difference between single-channel DC and multi-channel deep clustering (MDC) is definition of the input features. In single-channel DC, only log-magnitude spectral of microphone input signal $X_{DNN} = \{\log|\boldsymbol{x}_{l,k}|\}$ is utilized as an input feature. In the MDC, the phase difference between two microphones, $\beta_{r,l,k}$ (r is the index of a microphone pair), is utilized as an additional input feature as follows:

$$X_{DNN} = \{\log|\boldsymbol{x}_{l,k}|, \cos\beta_{r,l,k}, \sin\beta_{r,l,k}\}. \tag{3}$$

$\beta_{r,l,k}$ reflects the spatial location of the dominant speech source at the frame l and the frequency k. Therefore, by using $\beta_{r,l,k}$ as an additional feature, time-frequency points in which the spatial location is the same is gathered into one cluster.

$\boldsymbol{V}$ is estimated via stacked Bidirectional Long-Short Term Memory (BLSTM) layers. In Fig. 1 (a), the block diagram of the conventional MDC is shown. $\boldsymbol{v}_{l,k}$ is corresponding with the lkth row of $\boldsymbol{V}$. By applying K-means clustering for the estimated embedding vectors, the time-frequency mask of each speech source can be obtained. When the number of the microphones is more than two, $\boldsymbol{V}$ is calculated for each microphone pair, and the K-means clustering is applied for the stacked $\boldsymbol{V}$ of all microphone pairs. However, stacking of the embedding vectors requires for multiple-times forward calculation of the neural network, and computational cost is proportional to the number of the microphone pairs, $\frac{N_m(N_m-1)}{2}$ in the inference stage.

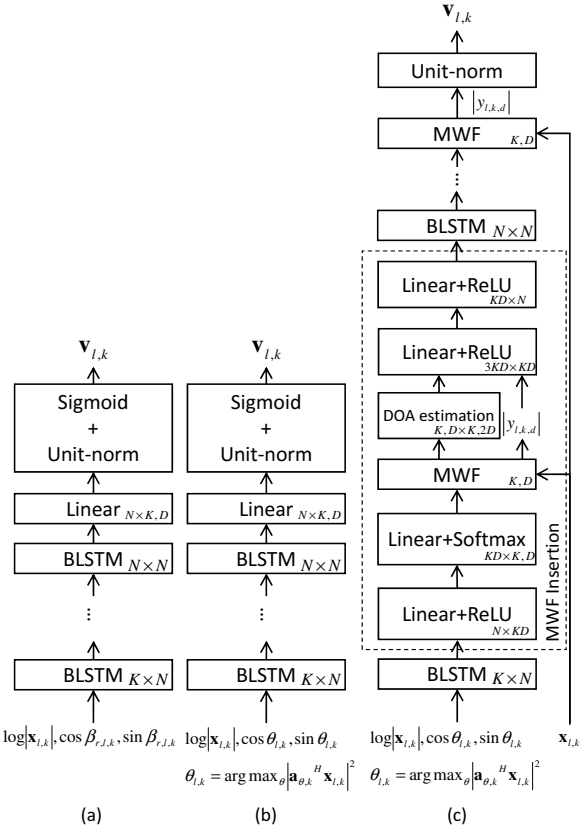

Fig. 1. Block diagrams: (a) Multi-channel Deep Clustering (MDC), phase (b) MDC, DOA, (c) MDC, DOA, MWF insertion

3. PROPOSED METHOD

3.1. Overview of proposed method

The proposed method introduces two types of spatial information into the MDC framework so as to reduce computation cost in the inference stage and to enhance estimation accuracy of the embedding vectors. In Fig. 1 (b) and (c), two types of the proposed method are shown. In Fig. 1 (b), instead of utilizing the phase difference between microphones, the estimated direction-of-arrival (DOA) $\theta_{l,k}$ is utilized as an alternative spatial feature. On contrary to the phase difference between microphones, it is not needed to perform stacking of embedding vectors of all microphone pairs. Even when N_m is more than two, it is needed to calculate only one embedding vector. Additionally, the estimated DOA is more reliable than the estimated phase difference between two microphones $\beta_{r,l,k}$, because the DOA is estimated by combining all the microphones. The second spatial information is the time-varying activity of each speech source estimated by spatial beamformer, multi-channel Wiener filtering (MWF). The MWF is inserted between two consecutive BLSTM layers. In Fig. 1 (c), the block diagram of the proposed method with the MWF insertion is shown. The MWF insertion structure is similar

to a low-rank approximation, e.g., auto-encoder [22]. The embedding space is constrained on the estimated time-frequency activity of each speech source by the MWF. By the MWF insertion, consistency of the embedding vectors along the time-axis is enhanced.

3.2. Spatial feature based on DOA estimation

Under the assumption that there is only one speech source at each time-frequency point and that the spatial location of each speech source is time-invariant, the microphone input signal can be approximated as follows:

$$\boldsymbol{x}_{l,k} \approx \boldsymbol{s}_{i_{l,k},l,k} = s_{i_{l,k},l,k}\boldsymbol{a}_{\theta_{i_{l,k}},k}, \tag{4}$$

where $\boldsymbol{a}_{\theta_i,k}$ is the steering vector of the ith speech source, and θ_i is the DOA of the ith speech source, and $i_{l,k}$ is the dominant source index at (l,k). Without loss of generality, $|\boldsymbol{a}_{\theta_{i_{l,k}},k}|$ can be assumed to be 1. The DOA of the active speech source at each time-frequency point is estimated as follows:

$$\theta_{l,k} = \arg\max_{\theta} |\boldsymbol{a}_{\theta,k}^H \boldsymbol{x}_{l,k}|^2, \tag{5}$$

where H is the Hermite transpose operator of a matrix/vector and $\boldsymbol{a}_{\theta,k}$ is the steering vector in which the DOA is assumed to be θ. For simplicity, all of the sound sources are assumed to be in the same horizontal plane. Therefore, $\theta_{l,k}$ is regarded as azimuth. The DOA based input feature for the neural network is defined as follows:

$$X_{DNN} = \{\log|\boldsymbol{x}_{l,k}|, \cos\theta_{l,k}, \sin\theta_{l,k}\}. \tag{6}$$

3.3. Multi-channel Wiener filtering (MWF) insertion between two BLSTM layers

The multi-channel Wiener filtering (MWF) is inserted between two BLSTM layers. The output signal of each BLSTM layer is transformed into a time-frequency mask of each source as follows:

$$M_{l,k,d} = \text{DNN}(\{o_{l,n}\}_{l,n}), \tag{7}$$

where $o_{l,n}$ is the output signal of the BLSTM layer and n is the feature index of the feature vector. $\text{DNN}(\cdot)$ is set to two fully-connected layers with ReLU and Softmax activations, respectively. The proposed method associates each dimension of the embedding vector with each virtual speech source. $M_{l,k,d}$ is interpreted as a time-frequency mask which extracts the dth virtual speech source. By using $M_{l,k,d}$, the spatial covariance matrix of the dth virtual speech source is estimated as follows:

$$\boldsymbol{R}_{k,d} = \frac{1}{\sum_l M_{l,k,d}} \sum_l M_{l,k,d}\boldsymbol{x}_{l,k}\boldsymbol{x}_{l,k}^H. \tag{8}$$

The MWF that extracts the dth virtual speech source is obtained as follows:

$$\boldsymbol{W}_{k,d} = \boldsymbol{R}_{k,d}\Big(\sum_i \boldsymbol{R}_{k,i}\Big)^{-1}. \tag{9}$$

The output signal of the MWF is obtained as follows:

$$\boldsymbol{y}_{l,k,d} = \boldsymbol{W}_{k,d}\boldsymbol{x}_{l,k}. \tag{10}$$

The time-frequency activity of the dth virtual speech source, $|\boldsymbol{y}_{l,k,d}|$, is utilized as one of the input features for the successive BLSTM layer. The dth input feature represents the time-frequency activity of the same virtual speech source along time-axis. Therefore, it can be said that consistency of the embedding vectors along time-axis is enhanced. In addition to the time-frequency activity, the estimated DOA for $\boldsymbol{y}_{l,k,d}$ is also inserted into the input feature of the successive BLSTM. The input feature for the $b+1$th LSTM layer, $\boldsymbol{f}_{b+1}$, is defined as follows:

$$\boldsymbol{f}_{b+1} = \{|\boldsymbol{y}_{l,k,d}|, \cos\theta_{l,k,d}, \sin\theta_{l,k,d}\}, \tag{11}$$

where $\theta_{l,k,d}$ is the estimated DOA of $\boldsymbol{y}_{l,k,d}$, which is defined as follows:

$$\theta_{l,k,d} = \arg\max_{\theta} |\boldsymbol{a}_{\theta,k}^H \boldsymbol{y}_{l,k,d}|^2. \tag{12}$$

In the back-propagation stage, the gradient for $\theta_{l,k,d}$ is not calculated. However, estimation accuracy of $\theta_{l,k,d}$ will be improved by updating $\boldsymbol{y}_{l,k,d}$. After the final BLSTM layer, the output embedding vector $\boldsymbol{v}_{l,k}$ is defined as $\left[\begin{array}{ccc} \frac{|\boldsymbol{y}_{l,k,1}|}{\sqrt{\sum_i |\boldsymbol{y}_{l,k,i}|^2}} & \cdots & \frac{|\boldsymbol{y}_{l,k,D}|}{\sqrt{\sum_i |\boldsymbol{y}_{l,k,i}|^2}} \end{array}\right]^T$.

4. EXPERIMENT

4.1. Experimental setup

Table 1. Details of dataset

	N_m	Spacing (cm)	Speech corpus	Number of utterances
Train	4	3-8-3	Train	2000
Eval1	4	4-8-4	Test	1000
Eval2	8	4-4-4-8-4-4-4	Test	1000

Speech source separation performance of the proposed method was evaluated by using measured impulse responses in Multi-channel Impulse Response Database (MIRD) [23] and TIMIT speech corpus [24]. The reverberation time RT_{60} was set to 0.16 (sec). Sampling rate was set to 8000 Hz. Therefore, sampling rate of the original speech corpus and the original impulse responses were downsampled. The number of the speech sources was set to 2. Frame size was 256 pt. Frame shift was 64 pt. The number of frequency bins was 129. The parameters of the neural network were trained by using four-microphones dataset. The impulse responses of the third, fourth, fifth, sixth microphone were extracted from the original eight-microphones impulse responses. The distance between a speech source and a microphone was set to 1 m. Azimuth of each talker is randomly selected for each utterance. In Table 1, details of the training dataset and the evaluation dataset are shown. Microphone positions were assumed to be known in advance, but impulse responses were unknown. A different microphone array was utilized in the evaluation dataset from the training dataset. Speech source separation performance with unknown impulse responses was evaluated.

In the training phase, mini-batch size was set to 128. Each utterance was splitted in every 100-frames segment. Therefore, length of each data was 100 (frame). Neural network parameters were updated by 10000 times. Adam optimizer [25] (learning rate was 0.001) with gradient clipping was utilized. If the sth speech source is the dominant source at the frame l and the frequency k, $Y_{lk,s}$ was set to 1, otherwise, $Y_{lk,s}$ was set to 0. The proposed architecture contains complex-valued gradient calculation. Tensorflow [26] was utilized for complex-valued gradient calculation.

4.2. Comparative methods

The following three time-frequency mask estimation methods were evaluated.

Table 2. Evaluation results of time-frequency embedding performance

Approaches	Eval1 $N_m = 4$				Eval2 $N_m = 8$			
	SIR (dB)	SDR (dB)	ΔMFCC	seg. SNR (dB)	SIR (dB)	SDR (dB)	ΔMFCC	seg SNR (dB)
MDC, phase ($D = 10$)	17.77	10.61	-1.33	9.35	17.57	10.47	-1.72	9.17
MDC, phase ($D = 20$)	17.94	10.77	-1.25	**9.42**	17.65	10.56	-1.71	9.19
MDC, DOA($D = 10$)	**17.98**	10.84	-0.92	9.37	17.89	10.72	-1.20	9.29
MDC, DOA ($D = 20$)	17.88	10.74	-0.98	9.34	17.86	10.68	-1.26	9.27
MDC, DOA, MWF insertion	17.21	**11.50**	**3.52**	8.49	**18.45**	**12.14**	**3.23**	**9.54**

Table 3. Evaluation results of post multi-channel Wiener filtering results

Front-end approaches	Eval1 $N_m = 4$				Eval2 $N_m = 8$			
	SIR (dB)	SDR (dB)	ΔMFCC	seg. SNR (dB)	SIR (dB)	SDR (dB)	ΔMFCC	seg SNR (dB)
MDC, phase ($D = 10$)	14.89	13.72	5.00	7.70	17.13	15.38	5.15	9.30
MDC, phase ($D = 20$)	15.00	13.82	5.02	7.77	17.23	15.47	5.13	9.36
MDC, DOA ($D = 10$)	15.04	13.87	5.03	7.84	17.45	15.73	5.47	9.57
MDC, DOA ($D = 20$)	14.99	13.82	5.02	7.81	17.40	15.68	5.47	9.54
MDC, DOA, MWF insertion	**15.36**	**14.06**	**5.18**	**8.05**	**18.14**	**16.33**	**5.76**	**10.11**

- MDC, phase (The conventional MDC [21]): Embedding vectors of all microphone pairs are stacked into one embedding vector. Time-frequency masks are obtained by K-means clustering of the stacked embedding vector.
- MDC, DOA: Instead of the phase difference between two microphones, DOA estimation results are utilized in MDC. K-means clustering is applied for no-stacked embedding vector.
- MDC, DOA, MWF insertion: In addition to the DOA estimation based input feature, MWF is inserted between two BLSTM layers.

In each method, the time-frequency mask $M_{s,l,k}$(s is the source index) is estimated via K-means clustering of the estimated embedding vector $\boldsymbol{v}_{l,k}$. The output signal, $\boldsymbol{y}_{s,l,k}$, is obtained as follows:

$$\boldsymbol{y}_{s,l,k} = M_{s,l,k}\boldsymbol{x}_{l,k}, \tag{13}$$

where $M_{s,l,k}$ is the estimated time-frequency mask by K-means clustering and $\boldsymbol{y}_{s,l,k}$ is the separated sth speech source. In "MDC, phase" and "MDC, DOA", the dimension of the embedding vector, D, was set to 10 or 20. in "MDC, DOA", MWF insertion, D was set to 10. The number of the BLSTM layers was 4. The number of the units in each BLSTM layer was set to 600.

Time-frequency masking is non-linear filtering, and it produces distortion in the output signal. To remove distortion of the output signal, multi-channel spatial filtering is effective as a post filtering [27]. Therefore, the time-frequency mask based post MWF was also evaluated. Mask based covariance matrix, MWF, and the output signal are obtained by the same form of the inserted MWF into two BLSTM layers, Eq. 8, Eq. 9, and Eq. 10, respectively.

4.3. Evaluation measures

Evaluation measures were set to SDR, SIR, Mel Frequency Cepstrum Coefficients (MFCC) distance improvement, and segmental Signal-to-Noise Ratio (seg. SNR). SDR and SIR were calculated by using BSS_EVAL [28]. MFCC distance improvement, ΔMFCC is defined as $\text{MFCC}_{input} - \text{MFCC}_{output}$. 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 was defined as follows:

$$\text{seg. SNR} = \frac{1}{L}\sum_{\tau=0}^{L} -10\log_{10}\frac{\sum_{p=0}^{P-1}\|s_{P\tau+p}\|^2}{\sum_{p=0}^{P-1}\|s_{P\tau+p} - \hat{s}_{P\tau+p}\|^2}, \tag{14}$$

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 result was calculated as average of 1000 utterances.

4.4. Experimental results

At first, time-frequency embedding performance was evaluated. In this evaluation, the output signal is obtained by the time-frequency masking. Evaluation results were shown in Table 2. It can be said that by using DOA estimation results as input features, less distorted embedding vectors can be obtained. In addition to that by inserting the MWF, distortion can be removed more than the "MDC, DOA" in 4ch and 8 ch cases. Therefore, the MWF insertion is shown to be effective. Secondly, the time-frequency mask based post MWF results are shown in Table 3. It is said that by using the post MWF, speech distortion in the output signal can be reduced more than time-frequency masking. The post MWF with the proposed method achieved the best performance. Therefore, the proposed method is shown to be effective.

5. CONCLUSION

In this paper, a multi-channel deep clustering based method with two types of spatial information was proposed, i.e., 1)The estimated direction-of-arrival (DOA) at each time-frequency point , 2)Insertion of the multi-channel Wiener filtering between two BLSTM layers so as to enhance embedding consistency along time-axis. Experimental results showed that MDC with the estimated DOA achieved better embedding performance than the conventional multi-channel Deep clustering. Furthermore, by the proposed MWF insertion, less distorted and more separated speech signals are available compared to the conventional method.

6. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley-Interscience, 2001.

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

[3] S. Makino, *Audio Source Separation*, Springer Publishing Company, Incorporated, 2018.

[4] 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.

[5] 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.

[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] T. Kim, H.T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, Jan. 2007.

[9] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, March 2010.

[10] 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.

[11] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, Sept 2016.

[12] 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.

[13] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 31–35.

[14] 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.

[15] 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.

[16] 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.

[17] H. Erdogan, J.R. Hershey, S. Watanabe, M.I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *Interspeech*, 2016, pp. 1981–1985.

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

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

[20] 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.

[21] 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.

[22] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504 – 507, 2006.

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

[24] 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.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[26] 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.

[27] T. Higuchi, K. Kinoshita, M. Delcroix, K. Zmolikova, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources," in *Proc. Interspeech 2017*, 2017, pp. 1183–1187.

[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.