Deep Multi-channel Speech Source Separation with Time-frequency Masking for Spatially Filtered Microphone Input Signal

Masahito Togami
LINE Corporation, Tokyo, Japan
masahito.togami@linecorp.com

***Abstract*—In this paper, we propose a multi-channel speech source separation technique which connects an unsupervised spatial filtering without a deep neural network (DNN) to a DNN-based speech source separation in a cascade manner. In the speech source separation technique, estimation of a covariance matrix is a highly important part. Recent studies showed that it is effective to estimate the covariance matrix by multiplying cross-correlation of microphone input signal with a time-frequency mask (TFM) inferred by the DNN. However, this assumption is not valid actually and overlapping of multiple speech sources lead to degradation of estimation accuracy of the multi-channel covariance matrix. Instead, we propose a multi-channel covariance matrix estimation technique which estimates the covariance matrix by a TFM for the separated speech signal by the unsupervised spatial filtering. Pre-filtered signal can reduce overlapping of multiple speech sources and increase estimation accuracy of the covariance matrix. Experimental results show that the proposed estimation technique of the multi-channel covariance matrix is effective.**

***Index Terms*—Speech source separation, time-frequency masking, deep neural network, multi-channel covariance matrix estimation**

I. Introduction

Speech source separation techniques have been actively studied for improvement of speech quality of hands-free communication systems and automatic speech recognition systems. Thanks to high expression capability of the deep neural network (DNN) for complicated speech spectrum, the DNN-based speech source separation has evolved so much. A popular approach of the DNN-based speech source separation is a time-frequency mask based approach [1]–[5]. Under the assumption that speech sources rarely overlap in the time-frequency domain, single-channel speech source separation can be performed by multiplying the inferred time-frequency mask and the microphone input signal. However, since multiple speech sources overlap actually in the same time-frequency point, time-frequency masking suppresses overlapped speech sources and generates distortion in the output signal.

In the multi-channel speech source separation context, the time-frequency masks inferred by the DNN are utilized for estimation of a multi-channel covariance matrix of each speech source. The multi-channel covariance matrix is estimated by averaging cross-correlation of time-frequency masked microphone input signal along time-axis [6]–[10]. Since the final output signal is generated by a multi-channel spatial filtering, the distortion of the output signal of the multi-channel approach is less than that of the single-channel approach. However, the DNN for time-frequency mask estimation does not optimize multi-channel speech source separation performance. To optimize the multi-channel speech source separation performance, we propose a DNN-based speech source separation in which the output signal after multi-channel speech source separation is evaluated [11], [12]. However, this method also estimates the multi-channel covariance matrix with time-frequency masking for microphone input signal under the assumption that speech sources rarely overlap in the time-frequency domain.

In this paper, we propose a novel covariance matrix estimation method with a DNN. Instead of time-frequency masking for the microphone input signal, the proposed method infers the time-frequency mask for a separated multi-channel speech signal which is estimated by an unsupervised spatial filtering. Since overlapping of multiple speech sources in the estimated multi-channel speech source signal is reduced, the proposed method is less affected by the overlapping problem of multiple speech sources in time-frequency domain. In this setting, a loss function which evaluates the inter-mediate time-frequency mask is not applicable, because we cannot define an oracle time-frequency mask for the spatial filtered signal. Instead, the proposed method adopts a loss function which evaluates the output signal after speech source separation [11], [12]. Although the oracle signal of the time-frequency mask for the spatial filtered signal cannot be defined, the proposed loss function can train the DNN by evaluating the speech quality of the output signal. Experimental results show that the proposed covariance matrix estimation for the spatial filtered signal is more effective than the covariance matrix estimation for the microphone input signal.

II. Signal model

A. Microphone input signal

Microphone input signal is modeled as an instantaneous mixture in time-frequency domain 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, and N_s is the number of the speech sources. The objective of multi-channel speech source separation is to estimate $\boldsymbol{c}_{i,l,k}$ from the observed microphone input signal $\boldsymbol{x}_{l,k}$.

B. Probabilistic modeling based on local Gaussian modeling

The local Gaussian modeling (LGM) based speech source separation method [13] is a common approach for multi-channel speech source separation 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:

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

where $v_{i,l,k}$ is the time-frequency activity of the i-th speech source, $\boldsymbol{R}_{i,k}$ is the time-invariant multi-channel covariance matrix of the i-th speech source, ϕ_k is a model parameter which is defined as $\{\{v_{i,l,k}\}_{i,l}, \{\boldsymbol{R}_{i,k}\}_i\}$. Based on the prior PDF of each speech source, the posterior PDF of each speech source is estimated under the condition that the microphone input signal $\boldsymbol{x}_{l,k}$ is given as follows:

$$p(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k}, \phi_k) = \mathcal{N}(\boldsymbol{c}_{i,l,k}|\boldsymbol{\mu}_{i,l,k}, \boldsymbol{V}_{i,l,k}), \tag{3}$$

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

$$\boldsymbol{\mu}_{i,l,k} = \boldsymbol{W}_{i,l,k}\boldsymbol{x}_{i,l,k}, \tag{4}$$

$$\boldsymbol{V}_{i,l,k} = (\boldsymbol{I} - \boldsymbol{W}_{i,l,k})\boldsymbol{R}_{i,l,k}, \tag{5}$$

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:

$$\boldsymbol{W}_{i,l,k} = \boldsymbol{R}_{i,l,k}\left(\sum_{\acute{i}=0}^{N_s-1}\boldsymbol{R}_{\acute{i},l,k}\right)^{-1}. \tag{6}$$

After estimating the parameter ϕ_k, we can obtain the spatially filtered signal by Eq. 4.

III. Conventional speech source separation methods

A. Unsupervised speech source separation for local Gaussian modeling

Unsupervised speech source separation approaches [13], [14] update the parameter of each frequency ϕ_k so as to minimize the negative log likelihood function $\mathcal{F}_k(\phi_k) = \sum_l -\log p(\boldsymbol{x}_{l,k}|\phi_k)$. By using the prior PDF of each speech source defined in Eq. 2, $\mathcal{F}_k(\phi_k)$ can be calculated as follows:

$$\mathcal{F}_k(\phi_k) = \sum_{l=1}^{L_T}\boldsymbol{x}_{l,k}^H\boldsymbol{R}_{\boldsymbol{x},l,k}^{-1}\boldsymbol{x}_{l,k} + \log\det\boldsymbol{R}_{\boldsymbol{x},l,k} + const. \tag{7}$$

where H is the Hermitian transpose operator of a matrix/vector and $\boldsymbol{R}_{\boldsymbol{x},l,k} = \sum_i v_{i,l,k}\boldsymbol{R}_{i,k}$ is the multi-channel covariance matrix of the microphone input signal. It is difficult to minimize $\mathcal{F}_k(\phi_k)$ w.r.t. ϕ_k directly. Instead, an auxiliary function approach [14], [15] is adopted. An auxiliary function can be obtained as follows:

$$\begin{aligned}\mathcal{F}_k^+(\phi_k, \gamma_k) = &\sum_{l=1}^{L_T}\frac{\mathrm{tr}(\tilde{\boldsymbol{R}}_{\boldsymbol{x},l,k}\boldsymbol{Q}_{i,l,k}^H\boldsymbol{R}_{i,k}^{-1}\boldsymbol{Q}_{i,l,k})}{v_{i,l,k}} \\ &+ \log\det\boldsymbol{U}_{l,k} + \mathrm{tr}(\boldsymbol{R}_{\boldsymbol{x},l,k}\boldsymbol{U}_{l,k}^{-1}) - N_m,\end{aligned} \tag{8}$$

where $\tilde{\boldsymbol{R}}_{\boldsymbol{x},l,k} = \boldsymbol{x}_{l,k}\boldsymbol{x}_{l,k}^H$ and the auxiliary variable γ_k is $\{\boldsymbol{Q}_{i,l,k}, \boldsymbol{U}_{l,k}\}$. The auxiliary function satisfied the following condition:

$$\mathcal{F}_k(\phi_k) \leq \mathcal{F}_k^+(\phi_k, \gamma_k), \tag{9}$$

$$\mathcal{F}_k(\phi_k) = \min_{\gamma_k}\mathcal{F}_k^+(\phi_k, \gamma_k). \tag{10}$$

We can decrease $\mathcal{F}_k(\phi_k)$ monotonically by minimizing $\mathcal{F}_k^+(\phi_k, \gamma_k)$ w.r.t. ϕ_k and γ_k alternately, because

$$\begin{aligned}\mathcal{F}(\phi_k^{(t+1)}) &\leq \mathcal{F}^+(\phi_k^{(t+1)}, \gamma_k^{(t+1)}) \\ &\leq \mathcal{F}^+(\phi_k^{(t)}, \gamma_k^{(t+1)}) \\ &= \mathcal{F}(\phi_k^{(t)}),\end{aligned} \tag{11}$$

where t is the iteration index. $\phi_k^{(t+1)}$ can be obtained as follows:

$$\tilde{\boldsymbol{R}}_{i,k}^{(t+1)} = \boldsymbol{G}_{i,k}^{(t),-1}\#(\boldsymbol{R}_{i,k}^{(t)}\boldsymbol{J}_{i,k}^{(t)}\boldsymbol{R}_{i,k}^{(t)}), \tag{12}$$

$$v_{i,l,k}^{(t+1)} = v_{i,l,k}^{(t)}\sqrt{\frac{\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{x},l,k}^{(t),-1}\tilde{\boldsymbol{R}}_{\boldsymbol{x},l,k}\boldsymbol{R}_{\boldsymbol{x},l,k}^{(t),-1}\boldsymbol{R}_{i,k}^{(t+1)})}{\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{x},l,k}^{(t),-1}\boldsymbol{R}_{i,k}^{(t+1)})}}, \tag{13}$$

where $\boldsymbol{G}_{i,k} = \sum_l v_{i,l,k}\boldsymbol{R}_{\boldsymbol{x},l,k}^{-1}$, $\#$ is the geometric mean operator [16], and $\boldsymbol{J}_{i,k} = \sum_l v_{i,l,k}\boldsymbol{R}_{\boldsymbol{x},l,k}^{-1}\tilde{\boldsymbol{R}}_{\boldsymbol{x},l,k}\boldsymbol{R}_{\boldsymbol{x},l,k}^{-1}$. After the parameter optimization in each frequency bin, the inter-frequency permutation problem is solved by an external permutation solver, e.g., a power-spectral correlation based method [17].

B. Deep neural network based speech source separation

Recently, multi-channel speech source separation techniques on the basis of deep neural network (DNN) have been proposed. The covariance matrix of each speech source $\boldsymbol{R}_{i,k}$ is estimated with time-frequency masking which is inferred by a DNN [6]. The DNN-based methods separate speech sources without an external inter-frequency permutation solver based on high expression capability of speech spectrum of the DNN. Under the assumption that multiple speech sources rarely overlap with each other, a multi-channel covariance matrix of each speech source is estimated as follows:

$$\boldsymbol{R}_{i,k} = \frac{1}{\sum_l \mathcal{M}_{i,l,k}}\sum_l \mathcal{M}_{i,l,k}\boldsymbol{x}_{l,k}\boldsymbol{x}_{l,k}^H, \tag{14}$$

where $\mathcal{M}_{i,l,k}$ is the time-frequency mask for the i-th speech source. However, speech sources are frequently overlap. Let the microphone input signal $\boldsymbol{x}_{l,k}$ be $\boldsymbol{c}_{i,l,k} + \boldsymbol{r}_{i,l,k}$, where $\boldsymbol{r}_{i,l,k} = \sum_{j\neq i}\boldsymbol{c}_{j,l,k}$ and $\boldsymbol{r}_{i,l,k}$ is the summation of the other signals except of the i-th speech source signal. Under the

assumption that multiple speech sources are independent of each other, $\boldsymbol{R}_{i,k}$ can be approximated as follows:

$$\begin{aligned}\boldsymbol{R}_{i,k} \approx & \frac{1}{\sum_l \mathcal{M}_{i,l,k}} \sum_l \mathcal{M}_{i,l,k} \boldsymbol{c}_{i,l,k} \boldsymbol{c}_{i,l,k}^H \\ & + \frac{1}{\sum_l \mathcal{M}_{i,l,k}} \sum_l \mathcal{M}_{i,l,k} \boldsymbol{r}_{i,l,k} \boldsymbol{r}_{i,l,k}^H,\end{aligned} \tag{15}$$

where the first term of Eq. 15 is corresponding with the correct covariance matrix of the i-th speech source and the second term is corresponding with the covariance matrix of the other speech sources. These two matrices are non-negative definite matrices. Therefore, the output $\boldsymbol{R}_{i,k}$ always over-estimates the correct covariance matrix.

IV. Proposed method

A. Overview of proposed method

Block diagram of the proposed method is shown in Fig. 1. In the proposed method, prior to the DNN-based parameter estimation, the LGM-based speech source separation is performed. The separation parameter ϕ_k is updated based on Eq. 12 and Eq. 13. The MWF $\boldsymbol{W}_{uns,i,l,k}$ is obtained with $\phi_k^{(L_t)}$ (L_t is the number of iterations) by Eq. 6. The spatially filtered speech source $\boldsymbol{y}_{i,l,k}$ is estimated as $\boldsymbol{y}_{i,l,k} = \boldsymbol{W}_{uns,i,l,k} \boldsymbol{x}_{l,k}$. The input feature for the successive neural network is defined as log spectral of the filtered signal $\boldsymbol{y}_{i,l,k}$, $\cos\theta_{\boldsymbol{y}_{i,l,k}}$, and $\sin\theta_{\boldsymbol{y}_{i,l,k}}$, where $\theta_{\boldsymbol{y}_{i,l,k}}$ is the phase difference between microphones in $\boldsymbol{y}_{i,l,k}$. The neural network estimates time-frequency masks for $\boldsymbol{y}_{i,l,k}$. The neural network structure is shown in Fig. 2. The neural network consists of four bidirectional long short term memory (BLSTM) layers with 1200 hidden units and three dense layers. The neural network infers the time-frequency mask $\mathcal{M}_{i,j,l,k}$ that estimates the ratio of the i-th speech source in the j-th output signal $\boldsymbol{y}_{j,l,k}$ and the time-frequency activity of each source $v_{i,l,k}$. The multi-channel covariance matrix of each speech source is obtained by time-frequency masking for $\boldsymbol{y}_{j,l,k}$. A posterior PDF of each speech source is obtained via the estimated multi-channel covariance matrix. The loss function of the proposed method evaluates the posterior PDF of each speech source directly regarding each speech source as a probabilistic variable [11] , which is shown to be more effective than the l_2 loss function based method.

B. Time-frequency masking for filtered microphone input signal

Instead of utilizing time-frequency masking with microphone input signal, the proposed method utilizes time-frequency masking with the multi-channel spatially filtered signal by the unsupervised speech source separation without the DNN so as to overcome the over-estimation problem of the covariance matrix. We assume that there are some permutation errors in $\boldsymbol{y}_{i,l,k}$ due to the inter-frequency permutation solver. The proposed time-frequency mask also reduces the inter-frequency permutation errors by using supervised data. The covariance matrix of the proposed method is estimated as follows:

$$\boldsymbol{R}_{i,k} = \frac{1}{\sum_{l,j} \mathcal{M}_{i,j,l,k}} \sum_{l,j} \mathcal{M}_{i,j,l,k} \boldsymbol{y}_{j,l,k} \boldsymbol{y}_{j,l,k}^H, \tag{16}$$

where $\mathcal{M}_{i,j,l,k}$ estimates ratio of the i-th speech source in the j-th output signal $\boldsymbol{y}_{j,l,k}$ at each time-frequency point. If $\boldsymbol{W}_{uns,j,l,k}$ extracts the j-th speech source completely and reduces the other speech sources, $\boldsymbol{y}_{j,l,k}$ can be approximated as follows:

$$\boldsymbol{y}_{j,l,k} \approx \boldsymbol{c}_{j,l,k} + \alpha_{j,l,k} \boldsymbol{r}_{j,l,k}, \tag{17}$$

where $\alpha_{j,l,k}$ is less than 1. Under the mutual independence assumption of the speech sources, $\boldsymbol{R}_{i,k}$ can be approximated as follows:

$$\begin{aligned}\boldsymbol{R}_{i,k} \approx & \frac{1}{\sum_{j,l} \mathcal{M}_{i,j,l,k}} \sum_{j,l} \mathcal{M}_{i,j,l,k} \boldsymbol{c}_{j,l,k} \boldsymbol{c}_{j,l,k}^H \\ & + \frac{1}{\sum_{j,l} \mathcal{M}_{i,j,l,k}} \sum_{j,l} \mathcal{M}_{i,j,l,k} |\alpha_{j,l,k}|^2 \boldsymbol{r}_{j,l,k} \boldsymbol{r}_{j,l,k}^H.\end{aligned} \tag{18}$$

We further assume that the neural network completely solves the permutation problem completely and

$$\mathcal{M}_{i,j,l,k} = \begin{cases} \mathcal{M}_{i,l,k} \text{ if } i = g(j,k) \\ 0 \text{ otherwise} \end{cases}, \tag{19}$$

where g is the permutation function. In this case, $\boldsymbol{R}_{i,k}$ can be written as follows:

$$\begin{aligned}\boldsymbol{R}_{i,k} = & \frac{1}{\sum_l \mathcal{M}_{i,l,k}} \sum_l \mathcal{M}_{i,l,k} \boldsymbol{c}_{i,l,k} \boldsymbol{c}_{i,l,k}^H \\ & + \frac{1}{\sum_l \mathcal{M}_{i,l,k}} \sum_l \mathcal{M}_{i,l,k} |\alpha_{i,l,k}|^2 \boldsymbol{r}_{i,l,k} \boldsymbol{r}_{i,l,k}^H.\end{aligned} \tag{20}$$

Since $|\alpha_{i,l,k}|^2 \leq 1$, the amount of the over-estimation in $\boldsymbol{R}_{i,k}$ is reduced in Eq. 20 by comparing with Eq. 15. Therefore, the amount of the upper-bound of the overestimation of $\boldsymbol{R}_{i,k}$ can be reduced by using time-frequency masking for a multi-channel spatial filtered signal.

C. Loss function with permutation invariant training

When the time-frequency masking for the microphone input signal is utilized, distance between the oracle time-frequency mask and the inferred time-frequency mask is typically utilized as a loss function of the DNN. However, it is difficult to define an oracle time-frequency mask for the spatially filtered signal. Instead, the proposed method utilizes a loss function which evaluates the output signal after the DNN-based speech source separation. The proposed method infers the parameter of the prior PDF, i.e., $\boldsymbol{R}_{i,k}$ and $v_{i,k}$, via the DNN. By using the inferred parameter, the negative log posterior PDF, $-\log p(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k}, \phi_k^{(L_t)})$, is obtained. The proposed method utilizes the negative log posterior PDF as the loss function $\mathcal{C}$. To calculate the loss function, the utterance-level permutation

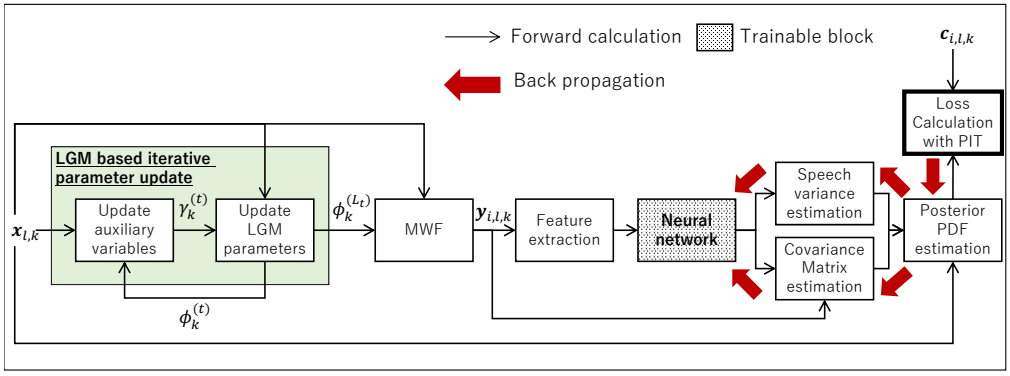

Fig. 1. Block diagram of proposed method

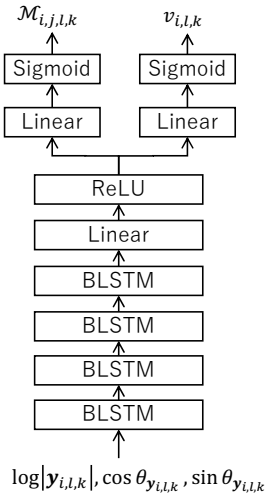

Fig. 2. Neural network structure

invariant training (PIT) [4] is utilized similarly to the conventional supervised speech source separation [5], [11], Π is a set of possible permutations, and $\mathcal{C}$ is obtained as follows:

$$\mathcal{C} = \min_{f\in\Pi} \sum_{i,l,k} \left(\boldsymbol{c}_{f(i),l,k} - \boldsymbol{\mu}_{i,l,k}\right)^H \boldsymbol{V}_{i,l,k}^{-1} \left(\boldsymbol{c}_{f(i),l,k} - \boldsymbol{\mu}_{i,l,k}\right) + \log|\boldsymbol{V}_{i,l,k}|. \quad (21)$$

V. Experiment

A. Experimental setup

Speech source separation performance of the proposed method was evaluated. The dataset was made by convolving measured impulse responses in Multi-channel Impulse Response Database (MIRD) [18] with the clean speech sources in TIMIT speech corpus [19]. 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 indices 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 and a 8-8-8-8-8-8-8 spacing (cm) microphone array were utilized, the distance between speech sources and microphones was set to 1 m. In the evaluation phase, a 4-4-4-8-4-4-4 spacing (cm) microphone array was utilized, and the distance between speech sources and microphones was set to 1 m or 2m. 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. 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 split in every 100-frames segment. Therefore, length of each data was 100 (frame). Adam optimizer [20] (learning rate was 0.001) with gradient clipping was utilized. The proposed architecture contains complex-valued gradient calculation. Tensorflow [21] was utilized for complex-valued gradient calculation. Evaluation measures were SDR, SIR from BSS_EVAL [22], Cepstrum distance (CD), Frequency-weighted segmental SNR (FWseg.SNR), and PESQ. The proposed method was compared with three unsupervised speech source separation methods, i.e., 1) Auxiliary-function-based IVA (Aux IVA) [23]; 2) ILRMA [24], [25]; and 3) local Gaussian modeling (LGM) [13]. We utilized Aux IVA and ILRMA implemented in [26]. We also evaluated two supervised methods, i.e., 1) Baseline: A time-frequency mask for a multi-channel microphone input signal is estimated via a DNN. The input feature is log spectral of the microphone input signal $\boldsymbol{x}_{l,k}$ and the phase difference between two microphone input signals; 2)TFM-Input: A time-frequency mask for a multi-channel microphone input signal is estimated via a DNN. The input feature is the same as the proposed method. Output signal is estimated by multi-channel spatial filtering. In the unsupervised speech source separation methods, the number of the iteration for the separation parameter update was 20. In each supervised method, the DNN parameter was updated by 10000 times.

B. Experimental results

Experimental results when the distance between the speech sources and the microphones is 1 m are shown in Table I. Experimental results when the distance between the speech sources and the microphones is 2 m are also shown in Table II. It is shown that the proposed method achieved the

TABLE I
EVALUATION RESULTS: DISTANCE IS 1 M

Approaches	SDR diff.	SIR diff.	CD diff.	FWseg.SNR diff.	PESQ diff.
AuxIVA	6.63	8.77	-1.13	4.96	0.51
ILRMA	6.73	8.92	-1.15	5.01	0.55
LGM	7.45	9.86	-1.43	5.52	0.64
Baseline	8.61	10.49	-1.66	5.88	0.73
TFM-Input	8.80	10.96	-1.76	6.26	0.76
Proposed	**8.94**	**11.13**	**-1.80**	**6.29**	**0.77**

TABLE II
EVALUATION RESULTS: DISTANCE IS 2 M

Approaches	SDR diff.	SIR diff.	CD diff.	FWseg.SNR diff.	PESQ diff.
AuxIVA	5.30	6.77	-0.85	3.29	0.34
ILRMA	5.45	6.98	-0.87	3.47	0.37
LGM	6.20	7.83	-1.12	4.18	0.46
Baseline	7.46	8.81	-1.37	4.55	0.55
TFM-Input	7.68	9.31	-1.54	4.98	0.59
Proposed	**7.78**	**9.52**	**-1.56**	**5.09**	**0.60**

best performance. The proposed method outperformed the other supervised methods. Especially, the proposed method outperformed the TFM-Input method. This result confirmed that the proposed time-frequency masking for the spatial filtered signal is effective.

VI. CONCLUSIONS

In this paper, we propose a deep neural network (DNN) based speech source separation technique. The proposed method estimates a time-frequency mask for estimating a multi-channel covariance matrix of a spatial filtered signal. On contrary to the conventional time-frequency masking for the microphone input signal, the proposed method is robust against the over-estimation problem of the multi-channel covariance matrix. Experimental results show that the proposed method is effective.

REFERENCES

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

[2] J. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP 2016*, 2016, pp. 31–35.

[3] Z. Wang, J. L. Roux, and J. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP 2018*, 2018, pp. 1–5.

[4] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP 2017*, March 2017, pp. 241–245.

[5] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *ICASSP 2018*, April 2018, pp. 5739–5743.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP 2016*, 2016, pp. 196–200.

[7] Y. Zhou and Y. Qian, "Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming," in *ICASSP 2018*, April 2018, pp. 536–540.

[8] H. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *ICASSP 2018*, April 2018, pp. 531–535.

[9] Z. Wang and D. Wang, "Mask weighted stft ratios for relative transfer function estimation and its application to robust asr," in *ICASSP 2018*, April 2018, pp. 5619–5623.

[10] 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 *ICASSP 2018*, April 2018, pp. 6717–6721.

[11] M. Togami, "Multi-channel Itakura Saito distance minimization with deep neural network," in *ICASSP 2019*, May 2019, pp. 536–540.

[12] Y. Masuyama, M. Togami, and T. Komatsu, "Multichannel Loss Function for Supervised Speech Source Separation by Mask-Based Beamforming," in *Proc. Interspeech 2019*, 2019, pp. 2708–2712. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1289

[13] N. 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.

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

[15] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," *30th International Conference on Machine Learning, ICML 2013*, pp. 1613–1621, 01 2013.

[16] K. Yoshii, K. Kitamura, Y. Bando, E. Nakamura, and T. Kawahara, "Independent low-rank tensor analysis for audio source separation," in *EUSIPCO 2018*, Sep. 2018, pp. 1657–1661.

[17] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1 – 24, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231200003453

[18] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," *IWAENC 2014*, pp. 313–317, 2014.

[19] 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.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*, 2015.

[21] 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. [Online]. Available: https://www.tensorflow.org/

[22] 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.

[23] 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.

[24] 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.

[25] ——, *Determined Blind Source separation with Independent Low-Rank Matrix Analysis*. Springer Publishing Company, Incorporated, 2018, ch. 6, pp. 125–155.

[26] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *ICASSP*, 2018, pp. 351–355.