Computer-Resource-Aware Deep Speech Separation with a Run-Time-Specified Number of BLSTM Layers

Masahito Togami* and Yoshiki Masuyama† and Tatsuya Komatsu* and Kazuyoshi Yoshii‡ and Tatsuya Kawahara‡
* LINE Corporation, Tokyo, Japan
E-mail: masahito.togami@linecorp.com
† Waseda University, Tokyo, Japan
‡ Kyoto University, Kyoto, Japan

***Abstract*—Recently, deep neural networks (DNNs) with multiple bidirectional long short term memory (BLSTM) layers have been successfully applied to supervised multi-channel speech separation. When it is applied for industrial products, one shortage is that the number of the BLSTM layers is not variable according to the available computational resource once the DNN is trained. Since available computational resource varies from device to device, it is preferable that the number of the BLSTM layers can be changed for optimal performance. In this paper, we propose a DNN based speech separation, in which each BLSTM layer is connected with a signal processing layer. It can output a separated speech signal, which can also be fed into the successive BLSTM layer. The proposed method trains two types of BLSTM layers. The first one is utilized for initialization of speech separation. The second one is utilized for enhancing separation performance. The proposed method can increase the number of the BLSTM layers by stacking the second type of the BLSTM layer to improve separation performance. Experimental results show that the proposed method is effective.**

I. Introduction

Separation of multiple speech sources accurately is one of the important issues in speech signal processing community. Recently, many speech separation techniques have been proposed [1], [2]. Generally speaking, there is a trade-off between computational cost and speech separation performance, and its acceptable computational cost heavily depends on the computational resource of a front-end device. Therefore, a speech separation framework which can flexibly control the amount of the computational cost is highly required in industrial applications.

Speech separation techniques are categorized into 1) unsupervised approaches which do not require any training dataset and 2) supervised approaches which require a training dataset. Unsupervised approaches are commonly called as blind source separation (BSS). BSS which separates speech sources with a generative model of a speech source has been actively studied, e.g., independent component analysis (ICA) [3], independent vector analysis (IVA) [4], [5], independent low-rank matrix analysis (ILRMA) [6], [7], local Gaussian modeling (LGM) [8], and multi-channel non-negative matrix factorization (MNMF) [9]–[13]. BSS updates a spatial model and a source model iteratively to increase separation performance. However, a source model which is used in BSS is not sufficient to capture complicated spectral characteristics of a speech source.

Recently, deep neural network (DNN) based approaches have been widely studied [14]–[23]. Supervised approaches can learn the complicated spectral characteristics of a speech source on the background of powerful expression capability of DNN. Especially, bidirectional long short term memory (BLSTM) based approaches have been actively studied e.g., permutation invariant training (PIT) [21], [22], deep clustering (DC) [24], [25], and multi-channel Itakura-Saito Divergence minimization (MISD-M) [23]. Current BLSTM based approaches assume that the number of the BLSTM layers is fixed, which limits applications. To expand applications of the BLSTM based approaches, it is valuable that the number of the BLSTM layers can be changed depending on the computational resource.

In this paper, we propose a BLSTM based speech separation method which can change the number of the BLSTM layers flexibly. Each BLSTM layer in the proposed method is connected with a signal processing layer which outputs separated speech signals. Each BLSTM layer is trained sequentially so as to minimize distance between the output signal and the oracle speech signal. Therefore, the proposed method can output a separated signal from any BLSTM layer. The proposed method trains two types of BLSTM layers. The first one is utilized for only initialization, and it is trained so as to output a clean signal from noisy microphone input signal. The second one is trained so as to increase separation performance, and it utilizes the output signal of the previous layer as the input signal. Stacking of the first type is not preferable because the first type assumes an input signal that is not well separated. Instead of the first type, the proposed method increases the number of the BLSTM layers by stacking the second type of the BLSTM layer while increasing separation performance. Stacking of variable number of DNNs are popular in the image processing research field [26] and the phase reconstruction research field [27]. To the best of authors' knowledge, it is

978-988-14768-8-3/20/$31.00 ©2020 APSIPA

the first time to apply variable number of BLSTM layers for supervised speech separation.

II. Problem statement and baseline

A. Input signal model

Let $\boldsymbol{x}_{l,k} \in \mathbb{C}^{N_m}$ be a multi-channel microphone input signal which is defined in a time-frequency domain (l is the frame index and k is the frequency index) as follows:

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

where $\boldsymbol{c}_{i,l,k}$ is the i-th speech source signal and N_s is the number of sources. The objective of speech separation is to separate $\boldsymbol{c}_{i,l,k}$ from the observed microphone input signal $\boldsymbol{x}_{l,k}$.

B. Supervised speech separation with multi-channel Itakura-Saito divergence minimization

One of the authors proposed multi-channel Itakura-Saito Divergence Minimization based supervised speech separation (MISD-M) [23] which utilizes a deep neural network (DNN). DNN based time-frequency mask estimation techniques typically train the DNN to minimize the mask estimation error [28], [29], which does not always lead to increasing multi-channel separation performance in the inference. Instead, the MISD-M trains the DNN to minimize estimation error of the separated signal, which directly leads to increasing multi-channel separation performance in the inference. Additionally, the time-frequency mask estimation techniques do not infer multiple separation parameters that are needed for a time-varying spatial filtering. On the other hand, the MISD-M can infer these multiple separation parameters simultaneously. Generally speaking, time-varying spatial filtering is more preferable than time-invariant spatial filtering.

In the training stage, the MISD-M evaluates a loss function based on speech quality of a multi-channel separated signal after multi-channel Wiener filtering (MWF). The output signal of the MISD-M suffers from the utterance-level permutation-problem. Thus, the MISD-M utilizes the utterance-level permutation-invariant-training (PIT) [21]. The loss function is set to the negative-log posterior-probability-distribution with the utterance-level PIT as follows:

$$\mathcal{L}_{\text{MISD}} = \min_{f \in \Pi} -\sum_{i,l,k} \log p_\theta(\boldsymbol{c}_{f(i),l,k} | \boldsymbol{x}_{l,k}), \tag{2}$$

where $\{\boldsymbol{c}_{i,l,k}, \boldsymbol{x}_{l,k}\}$ is a pairwise training data, Π is a set of possible permutations, and θ is a neural network parameter. The input of the MISD-M is set to a multi-channel observed microphone signal $\boldsymbol{x}_{l,k}$ and the training target is set to a multi-channel oracle clean signal $\boldsymbol{c}_{i,l,k}$. The posterior probability distribution is set to a time-varying Gaussian distribution as follows:

$$p_\theta(\boldsymbol{c}_{f(i),l,k} | \boldsymbol{x}_{l,k}) \sim \mathcal{N}(\boldsymbol{c}_{f(i),l,k}; \boldsymbol{\mu}_{i,l,k}, \boldsymbol{V}_{i,l,k}), \tag{3}$$

where $\boldsymbol{\mu}_{i,l,k}$ and $\boldsymbol{V}_{i,l,k}$ are the estimated mean vector and the estimated multi-channel covariance matrix of the posterior probability distribution of the i-th source, respectively. Thus, the neural network that infers parameters of the MWF is trained via backpropagation of the loss function defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{MISD}} = \min_{f \in \Pi} \sum_{i,l,k} (\boldsymbol{c}_{f(i),l,k} - \boldsymbol{\mu}_{i,l,k})^H \boldsymbol{V}_{i,l,k}^{-1} (\boldsymbol{c}_{f(i),l,k} \\ - \boldsymbol{\mu}_{i,l,k}) + \log |\boldsymbol{V}_{i,l,k}| + const.,\end{aligned} \tag{4}$$

where H is the Hermitian transpose of a matrix/vector. The prior probability distribution of $\boldsymbol{c}_{i,l,k}$ is modeled as a multi-channel Gaussian distribution as follows:

$$p_\theta(\boldsymbol{c}_{i,l,k}) \sim \mathcal{N}(\boldsymbol{c}_{i,l,k}; \boldsymbol{0}, v_{i,l,k} \boldsymbol{R}_{i,k}), \tag{5}$$

where $v_{i,l,k}$ is the time-varying activity of the i-th speech source and $\boldsymbol{R}_{i,k}$ is the time-invariant multi-channel spatial covariance matrix (SCM). Eq. (5) is known as local Gaussian modeling (LGM) [8]. Under the LGM, $\boldsymbol{\mu}_{i,l,k}$ and $\boldsymbol{V}_{i,l,k}$ are calculated as follows:

$$\boldsymbol{\mu}_{i,l,k} = \boldsymbol{W}_{i,l,k} \boldsymbol{x}_{l,k}, \tag{6}$$

$$\boldsymbol{V}_{i,l,k} = (\boldsymbol{I}_{N_m \times N_m} - \boldsymbol{W}_{i,l,k}) v_{i,l,k} \boldsymbol{R}_{i,k}, \tag{7}$$

where $\boldsymbol{I}$ is an identity matrix and $\boldsymbol{W}_{i,l,k}$ is a time-varying MWF which is defined as follows:

$$\boldsymbol{W}_{i,l,k} = v_{i,l,k} \boldsymbol{R}_{i,k} (\sum_{i'} v_{i',l,k} \boldsymbol{R}_{i',k})^{-1}. \tag{8}$$

In the MISD-M, the SCM $\boldsymbol{R}_{i,k}$ is estimated with a time-frequency mask $M_{i,l,k}$ as follows:

$$\boldsymbol{R}_{i,k} = \frac{1}{\sum_l M_{i,l,k}} \sum_l M_{i,l,k} \boldsymbol{x}_{l,k} \boldsymbol{x}_{l,k}^H. \tag{9}$$

In Fig. 1 (a), the block diagram of the MISD-M is shown. The neural network in the MISD-M with the parameter θ infers a time-frequency activity $v_{i,l,k}$ and a time-frequency mask $M_{i,l,k}$ as follows:

$$v_{i,l,k} = v_{i,l,k}(\boldsymbol{z}; \theta), \tag{10}$$

$$M_{i,l,k} = M_{i,l,k}(\boldsymbol{z}; \theta), \tag{11}$$

where $\boldsymbol{z}$ is set to the input feature. The input feature is defined as a concatenation of the log spectral of the microphone input signal $\log|\boldsymbol{x}_{l,k}|$ and $\{\cos \eta_{l,k}, \sin \eta_{l,k}\}$, where $\eta_{l,k}$ is the phase difference between microphones. The neural network of the MISD-M consists of multiple bidirectional long short term memory (BLSTM) layers and multiple dense layers. Let $\boldsymbol{h}_n$ be the output variable of the n-th BLSTM layer. Only the output variable $\boldsymbol{h}_L$ (L is the final layer index) of the final BLSTM layer is connected with the dense layers that infer $v_{i,l,k}$ and $M_{i,l,k}$. $v_{i,l,k}$ and $M_{i,l,k}$ cannot be inferred from the output variable $\boldsymbol{h}_n$ of any intermediate BLSTM layer. Thus, the number of the BLSTM layers cannot be changed depending on available computational resource of a device. In this way, direct connection of two BLSTM layers loses interpretability of the output variables of intermediate BLSTM layers.

III. Proposed method

A. Overview of proposed method

We extend the MISD-M based supervised speech separation so that the number of the BLSTM layers is variable according to the amount of available computational resource. The proposed method does not connect multiple BLSTM layers directly. Instead, the proposed method inserts signal processing layers between two BLSTM layers. Because of this insertion, the output variable of each BLSTM layer can be converted into separated speech signals.

The proposed method trains two types of BLSTM layers. We define the word "block". Each block contains one BLSTM layer, dense layers, and signal processing layers. In Fig. 1 (b) and (c), the block diagrams of the proposed blocks are shown. The first block contains the first type of the BLSTM layer. The first type is utilized for only initialization, and it outputs a refined separation parameter $v_{i,l,k}^{(1)}, \boldsymbol{R}_{i,k}^{(1)}$ from a randomly initialized parameter $v_{i,l,k}^{(0)}, \boldsymbol{R}_{i,k}^{(0)}$. The second block and the subsequent blocks share the second type of the BLSTM layer. The second type is utilized for increasing separation performance. The $n \geq$ 2-th block outputs more refined separation parameter $v_{i,l,k}^{(n)}, \boldsymbol{R}_{i,k}^{(n)}$ from the refined parameter in the previous block $v_{i,l,k}^{(n-1)}, \boldsymbol{R}_{i,k}^{(n-1)}$. The proposed method can extract the output signal of the MWF from any block. The proposed method does not stack blocks which contain the first type of the BLSTM layer, because the first type assumes an input signal that is not well separated and separation performance degrades when the input parameter of the first type is a refined parameter by the previous layer. Thus, the proposed method stacks blocks which contain the second type of the BLSTM layer. In this paper, separation performance is also evaluated experimentally when the first type of the BLSTM layer is stacked (MISD-M 2).

B. Proposed block architecture

1) Feature extraction: Each block extracts the input feature of the BLSTM layer as the log power spectral and the phase difference of the separated speech signal. The separated speech signal is generated by the MWF with the output parameter by the previous block.

2) Training blocks with neural network: The neural network structure of each block is the same as the MISD-M [23]. The number of the BLSTM layers is set to 1. As the same way to MISD-M, the output variable of the BLSTM is converted into $v_{i,l,k}$ and $M_{i,l,k}$.

3) SCM estimation and MWF: The SCM of each speech source $\boldsymbol{R}_{i,k}$ is estimated by Eq. 9 from $M_{i,l,k}$. The MWF is adapted by using the estimated $v_{i,l,k}$ and $\boldsymbol{R}_{i,k}$. In the training phase, the estimated speech signal $\boldsymbol{\mu}_{i,l,k}$ and $\boldsymbol{V}_{i,l,k}$ are evaluated in the loss function.

4) SCM update: Additional SCM updates based on the LGM is performed so as to increase separation performance. In the SCM update, the multi-channel covariance matrix of each source $\boldsymbol{R}_{i,k}$ is updated iteratively based on the expectation-maximization (EM) algorithm [30] as follows:

E step: The posterior probability distribution of each source $p_\theta(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$ is estimated with the current $\boldsymbol{R}_{i,k}$ and $v_{i,l,k}$, and $\boldsymbol{\mu}_{i,l,k}$ and $\boldsymbol{V}_{i,l,k}$ are obtained by Eq. (6) and Eq. (7).
M step: $\boldsymbol{R}_{i,k}$ is updated as follows:

$$\boldsymbol{R}_{i,k} = \frac{1}{L_T} \sum_l \boldsymbol{\mu}_{i,l,k} \boldsymbol{\mu}_{i,l,k}^H + \boldsymbol{V}_{i,l,k}, \tag{12}$$

where L_T is the number of the time-frames. Since $\boldsymbol{R}_{i,k}$ estimated by Eq. (9) is an approximated SCM by time-frequency masking, Eq. (12) updates $\boldsymbol{R}_{i,k}$ to decrease the approximation error.

C. Training of deep neural network

Each type of the BLSTM layer is trained sequentially so as to minimize distance between the output signal and the oracle speech signal. For example, in the first P updates, only the first type of the BLSTM layer is trained. In the next P updates, the second type of the BLSTM layer is trained. In each training phase, the loss function is set to $\mathcal{L}_{\text{MISD}}$ defined by Eq. (4) similarly to the MISD-M.

1) Training of first type: When the first type is trained, the input parameter $\boldsymbol{R}_{i,k}^{(0)}$ and $v_{i,l,k}^{(0)}$ are randomly initialized. The training target is set to the oracle clean signal $\boldsymbol{c}_{i,l,k}$.

2) Training of second type: After training the first type, the second type is trained under the condition that the parameter of the first type is fixed. Since the input parameter $\boldsymbol{R}_{i,k}^{(n)}$ and $v_{i,l,k}^{(n)}$ $(n \geq 1)$ is a refined parameter which is the output of the previous block in the inference stage, the input separation parameter for training of the second type is also set to a refined parameter. For training of the second type, the proposed method utilizes $\boldsymbol{R}_{i,k}^{(1)}$ and $v_{i,l,k}^{(1)}$ which is inferred via the first block with the first type of the BLSTM layer as the refined input separation parameter. The training target is also set to the oracle clean signal $\boldsymbol{c}_{i,l,k}$. Because of the proposed training strategy, we can avoid an input-mismatch problem between the training stage and the inference stage.

IV. Experiments and evaluations

A. Experimental setup

Speech separation performance was evaluated. The dataset was made by convolving measured impulse responses in Multi-channel Impulse Response Database (MIRD) [31] with the clean speech sources in TIMIT speech corpus [32]. In the training phase, TIMIT train corpus was utilized. In the evaluation phase, TIMIT test corpus was utilized. We evaluated speech separation performance with an open set of speakers. The reverberation time RT_{60} was set to 0.16 [sec] or 0.36 [sec]. The number of the microphones was set to 2. The number of the speech sources was set to 2 in each utterance. Two microphone indices were randomly selected for each utterance 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. In the evaluation phase, a 4-4-4-8-4-4-4 spacing (cm) microphone array was utilized. Thus, a different

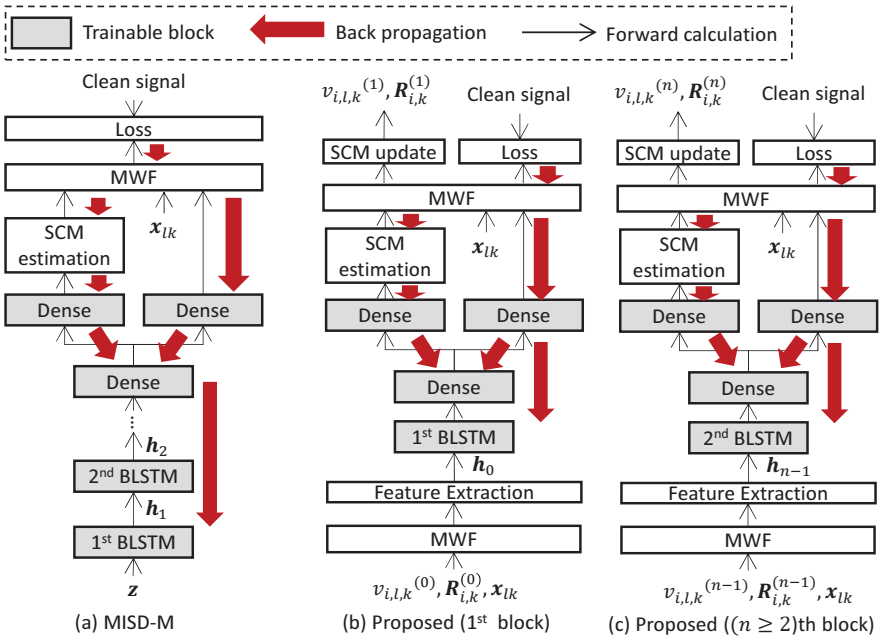

Fig. 1. Block diagrams

TABLE I
EVALUATION RESULTS OF SPEECH SEPARATION AT RT_{60}=0.16 [SEC]: EACH RESULT IS SIR/SDR

N_{sb}	Unsupervised		Supervised				
	LGM	ILRMA	LGM+DNN+KL	LGM+DNN+MISD	MISD-M 1	MISD-M 2	Proposed
1	10.66/8.90	9.71/8.50	1.39/1.13	6.85/4.55	11.72/10.08	12.41/10.47	11.90/10.08
2	-	-	1.32/0.88	8.57/5.74	-	2.71/2.01	12.78/10.71
3	-	-	1.29/0.69	9.30/6.37	-	11.90/10.17	13.10/10.77
4	-	-	1.47/0.61	9.65/6.71	-	2.88/2.18	13.13/**10.91**
5	-	-	1.61/0.49	9.80/6.88	-	11.88/10.14	**13.19**/10.79
6	-	-	1.65/0.36	9.84/6.95	-	3.07/2.32	13.18/**10.91**

microphone array was utilized in the evaluation phase from the training phase so as to evaluate separation performance in a microphone-array open condition. Sampling rate was set to 8000 Hz. Frame size was 256 sample. Frame shift was 64 sample. The number of the frequency bins K was 129. The number of the units in each BLSTM layer was set to 600. 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 10000. Mini-batch size was set to 128. Each utterance was split in every 100-frames segment in the training stage. Evaluation measures were set to SIR and SDR calculated by BSS_EVAL [33]. The DNN of each block was updated $P = 3000$ times. In the training phase, we utilized the permutation invariant training (PIT) [21]. Adam optimizer [34] (learning rate was 0.001) with gradient clipping was utilized.

B. Compared methods

We compared unsupervised speech separation methods and supervised speech separation methods. As unsupervised methods, two methods were evaluated, i.e., 1) LGM: Blind speech separation based on LGM [8] and an auxiliary function based parameter optimization [9], [10]; 2) ILRMA [6], [7]: We utilized an implementation in Pyroomacoustics [35]. The number of the basis functions in ILRMA was set to 2. The number of iterations for LGM and ILRMA was set to 20. As the supervised methods, the following methods were evaluated:

- LGM+DNN with Kulback Leibler divergence (KLD) minimization (LGM+DNN+KL) [14]: Although LGM+DNN was not proposed for speech separation, we evaluated LGM+DNN as a baseline method, because it has a variable number of blocks. In each block, the SCM is updated one-time based on the EM algorithm [8] and the time-varying activity is also updated one-time with the DNN similarly to the proposed method. The loss function is defined as a KLD which achieved the best performance in [14]. The DNN structure is the same as the proposed method.
- LGM+DNN with multi-channel Itakura-Saito Divergence

TABLE II
EVALUATION RESULTS OF SPEECH SEPARATION AT RT_{60}=0.36 [SEC]: EACH RESULT IS SIR/SDR

N_{sb}	Unsupervised		Supervised				
	LGM	ILRMA	LGM+DNN+KL	LGM+DNN+MISD	MISD-M 1	MISD-M 2	Proposed
1	6.58/4.79	6.30/4.59	1.40/1.07	6.43/4.05	8.13/6.61	8.62/6.84	8.24/6.56
2	-	-	1.42/0.91	6.69/4.39	-	2.12/1.52	9.07/7.05
3	-	-	1.28/0.64	6.98/4.68	-	8.24/6.65	9.67/7.33
4	-	-	1.35/0.52	7.13/4.86	-	2.34/1.73	9.64/**7.36**
5	-	-	1.33/0.26	7.19/4.95	-	8.14/6.58	**9.80**/7.34
6	-	-	1.35/0.11	7.22/4.99	-	2.60/1.95	9.69/7.27

minimization (LGM+DNN+MISD): The loss function is changed from the KLD to the MISD [23].

- MISD-M 1 [23]: The block diagram is shown in Fig. 1 (a). The number of the BLSTM layers was set to 2. The input feature $\boldsymbol{z}$ is extracted from the microphone input signal. The DNN is trained so as to minimize $\mathcal{L}_{\text{MISD}}$.
- MISD-M 2: The neural network structure is the same as that of the MISD-M 1 except for the input feature. The DNN is trained so as to minimize $\mathcal{L}_{\text{MISD}}$. The input feature is the same as that of the 1st BLSTM layer in the proposed method. Thus, the output signal of this method can be converted into the input feature of this method. We also evaluates this method by changing the number of the stacked blocks.
- Proposed: The number of the BLSTM layers in each block was 1. The loss function was MISD. The total number of BLSTM layers is the same as MISD-M 1. Two DNNs were trained in the proposed sequential way.

Importantly, the total number of the BLSTM layers are the same in all supervised methods.

C. Experimental results

The number of the stacked blocks N_{sb} was changed from 1 to 6. The experimental results are shown in Table I and Table II for RT_{60}=0.16 [sec] and 0.36 [sec], respectively. Average of 200 results is shown. The proposed method outperformed the other methods when $N_{sb} \geq 2$. The MISD-M 2 outperformed the MISD-M 1 because of difference of the input features. The MISD-M 2 with $N_{sb} = 1$ also outperformed the proposed method with $N_{sb} = 1$, because the number of the BLSTM layers in the MISD-M 2 with $N_{sb} = 1$ is more than that in the proposed method with $N_{sb} = 1$. When the number of the BLSTM layers is the same, e.g., the proposed method with $N_{sb} = 2$ and the MISD-M 2 with $N_{sb} = 1$, the proposed method outperformed the MISD-M 2. This result confirmed that insertion of signal processing layers between two BLSTM layers is effective. The MISD-M 2 with $N_{sb} = 2$ underperformed the MISD-M 2 with $N_{sb} = 1$. The input feature of the first block in the MISD-M 2 was extracted from the separation result with a randomly initialized parameter. On the other hand, the input feature of the second block of the MISD-M 2 was extracted from the separation result with the refined parameter by the first block. In the training stage, the MISD-M 2 was also trained with the input feature of the first block. Thus, there is the input-mismatch problem between the training stage and the inference stage in the second block of the MISD-M 2 with $N_{sb} = 2$, because the MISD-M 2 was trained with the randomly initialized parameter as the input parameter. When the input parameter of the MISD-M 2 is an refined parameter, the input mismatch results in poor separation performance. We guessed that this is the reason why the separation performance of the MISD-M 2 fluctuated depending on the number of blocks. On contrary, the proposed method with $N_{sb} = 2, 3, 4$ increased separation performance monotonically. Thus, it can be said that the input-mismatch problem was reduced in the proposed method. After $N_{sb} = 5, 6$, the separation result degraded. This result indicates that the second BLSTM layer cannot increase separation performance after $N_{sb} = 5$. However, when we can utilize the third BLSTM layer, there is possibility to achieve more separation performance after $N_{sb} = 5$. It is one of future works.

V. CONCLUSIONS

In this paper, we proposed a supervised speech separation technique which can flexibly change the number of BLSTM layers depending on the available computational resource. The proposed structure is useful for industrial applications, because the upper bound of the computational resource is varying from device to device in industrial applications. The proposed method trains two types of the BLSTM layers sequentially, i.e., a BLSTM layer for initialization of a separation parameter and a BLSTM layer for increasing separation performance. The number of the BLSTM layers is variable by stacking the second BLSTM layer. Experimental results show that speech separation performance of the proposed method can be improved by increasing the number of the BLSTM layers. The proposed architecture will be applied to the other kinds of array signal processing techniques such as dereverberation and speech enhancement.

REFERENCES

[1] S. Makino, *Audio Source Separation*. Springer Publishing Company, Incorporated, 2018.
[2] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*. Springer Publishing Company, Incorporated, 2007.
[3] P. Common, "Independent component analysis, a new concept ?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.

[4] A. Hiroe, "Solution of permutation problem in frequency domain ica using multivariate probability density functions," in *Proceedings ICA*, Mar. 2006, pp. 601–608.
[5] T. Kim, H. 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.
[6] 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.
[7] ——, *Determined Blind Source separation with Independent Low-Rank Matrix Analysis*. Springer Publishing Company, Incorporated, 2018, ch. 6, pp. 125–155.
[8] 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.
[9] 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.
[10] 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.
[11] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, K. Yoshii, and T. Kawahara, "Bayesian multichannel nonnegative matrix factorization for audio source separation and localization," in *ICASSP 2017*, March 2017, pp. 551–555.
[12] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel nonnegative matrix factorization," in *ICASSP 2018*, April 2018, pp. 31–35.
[13] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition," 04 2018, pp. 5734–5738.
[14] 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] A. Nugraha, A. Liutkus, and E. Vincent, "Deep neural network based multichannel audio source separation," in *Audio Source Separation*. Springer, Mar. 2018.
[16] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for multichannel audio source separation," in *EUSIPCO 2018*, Sep. 2018, pp. 1557–1561.
[17] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, in *Proc. of APSIPA ASC*, 2018.
[18] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *ICASSP 2018*, April 2018, pp. 716–720.
[19] H. Kameoka, L. Li, S. Inoue, and S. Makino. (2018) Semi-blind source separation with multichannel variational autoencoder.
[20] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda. (2018) Generalized multichannel variational autoencoder for underdetermined source separation.
[21] 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.
[22] 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.
[23] M. Togami, "Multi-channel Itakura Saito distance minimization with deep neural network," in *ICASSP 2019*, May 2019, pp. 536–540.
[24] 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.
[25] 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.
[26] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016. [Online]. Available: http://arxiv.org/abs/1603.06937
[27] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep Griffin-Lim iteration," *CoRR*, vol. abs/1903.03971, 2019. [Online]. Available: http://arxiv.org/abs/1903.03971
[28] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP 2015*, April 2015, pp. 708–712.
[29] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP 2016*, 2016, pp. 196–200.
[30] 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.
[31] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," *IWAENC 2014*, pp. 313–317, 2014.
[32] 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.
[33] 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.
[34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*, 2015.
[35] 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.