Spatial Normalization to Reduce Positional Complexity in Direction-aided Supervised Binaural Sound Source Separation

Ryu Takeda*, Kazuhiro Nakadai† and Kazunori Komatani*
* The Institute of Scientific and Industrial Research (SANKEN), Osaka University, Japan
E-mail: {rtakeda,komatani}@sanken.osaka-u.ac.jp Tel/Fax: +81-6-6879-8416/+81-6-6879-8438
† Honda Research Institute Japan, Co, Ltd., Japan
E-mail: nakadai@jp.honda-ri.com

***Abstract*—This paper describes a novel normalization method for mask-based binaural sound source separation using neural networks (NNs). Given a target source direction, the NNs estimate masks that extract target source components in the time-frequency domain. The numerous patterns of sound source numbers and positions make it difficult to train the NNs because some equivalent patterns are treated as different ones. We therefore propose a spatial normalization method of input signals as a pre-processing of the mask estimation. This normalization can reduce the essential positional complexity by converting the transfer functions of input signals into a canonical form using the target direction. This normalization improves the mask estimation and achieves the optimization of spatial pre-filters. Experiments using mixtures of two, three, and four sources showed that the NNs with our spatial normalization improved the signal-to-distortion ratio by 2.1 dB compared with the NNs without the spatial normalization.**

I. Introduction

A. Background and Motivation

Online sound source separation using two microphones (binaural) is attracting interest in terms of computational auditory scene analysis [1], [2] because the minimum number of microphones to utilize spatial information (e.g., direction of sound sources) is two. Additionally, binaural separation is more difficult than multi-microphone separation because of the limited number of microphones. An underdetermined situation where the number of sound sources is more than the number of microphones is a typical example of such cases.

Selective source separation based on a given direction or position is one of the most important functions in binaural sound source separation. Here, "selective" separation means the extraction of a target sound signal from microphone inputs, not the separation of all source signals from the inputs. This is a kind of attentional function using directional information that is usually independent of signal characteristics. For example, a spoken dialogue robot will sometimes listen to a speaker from a certain direction. However, in actual situations, the assumption that a target source is always located at a fixed direction [3] is too restrictive. Moreover, the position and the number of sound sources will change dynamically. We here assume that the position of sound sources can be estimated by sound source localization [4]. The target speaker direction can be also specified by vision and dialogue context in human-robot interaction.

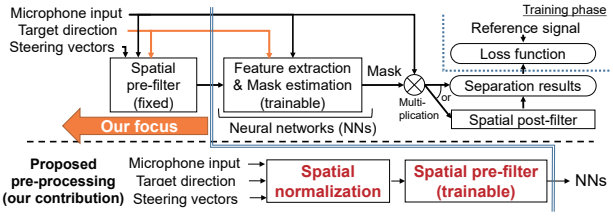

Fig. 1. Overview of supervised source separation using target source direction and our research focus

Recent approaches for such a dynamic situation are based on supervised source separation using neural networks (NNs) [5], but there has been little research in the binaural domain [6], as mentioned in [7]. The approach using a time-frequency (TF) mask in the short-time Fourier transformation (STFT) domain [3] is compatible with beamforming that can utilize the positional information of a target and other sound sources. Figure 1 shows the typical configuration of mask-based direction-aided separation. A fixed spatial pre-filter using steering vectors, such as the delay-and-sum (DS) or another beamformer, is first applied to enhance the target source signal [7], [8], [9], [10]. Here, a steering vector represents a relative transfer function between a sound source and each microphone. Various features for NNs, such as the interaural intensity difference (IID) and interaural phase difference (IPD), are extracted [11], [12], and noise components are masked by using NNs outputs. The spatial post-filter, e.g., time-varying Weiner filter, can be applied [9]. The spatial pre-filters are usually designed manually and have not been optimized.

The fundamental problem of selective source separation lies in the numerous patterns of sound source number and positions. The key points for this problem are the fixed spatial pre-filters and the pre-processing of the input signals. Most previous studies do not explicitly consider this problem, as they have focused mainly on the feature extraction [8], [13],

NN model [14], or loss function [15], [16]. A straightforward idea for the fixed pre-filters, for example, is to optimize them by training. However, such NN training never succeeds because, while optimization usually determines a fixed-parameter set, truly optimal filters will change according to the number of sound sources and their directions. Therefore, normalizing input signals is required to reduce the positional complexity for the pre-filter optimization.

In light of the above, we propose a novel spatial normalization method and an optimization of the spatial pre-filters as shown in Fig. 1. The *spatial normalization* converts the transfer functions of input signals into a canonical form by using a given target direction and the steering vectors. Because this normalization can fix the target source direction virtually and creates a *one-versus-the-rest* situation in terms of directions of the target and the other sound sources, the mask estimation by NNs improves. The optimization of the spatial pre-filters also succeeds because this normalization can reduce the fundamental complexity of the source directions. We demonstrate the effectiveness of our method through experiments using more than two sound sources and also show it does not depend on kinds of sound sources.

B. Relation To Prior Work

Although mask-based separation using NNs has been widely researched for the monaural and multi-channel situations, there are few studies on the binaural situation, as mentioned in Section I-A. Most studies assume *batch* processing and focus on the structure of NNs, the feature engineering, or the loss function, and pay less attention to the pre-processing of input signals and spatial pre-filters.

The most recent monaural separation methods are completely based on NNs and supervised training because spatial information cannot be utilized in the monaural situation. There are two main strategies to separate multiple sound sources: permutation invariant training [15] and clustering [16]. Many studies examined features [13], training targets [17], [18], and deep models [14], [19], [20], [21], [22] within this context.

The multi-channel approaches are usually based on the combination of TF-mask using NNs and the spatial filters. The source selectivity using a target source direction has recently been the subject of research focus [8], [10], and a special structure, such as direction attractor networks [9], is used in the NNs for online mask estimation. A clustering approach is extended to the stereo or multi-microphone situation [23], [24]. Most studies evaluate the situation where the number of dominant sound sources is less than the number of microphones, which does not hold for the actual binaural separation.

II. Preliminaries

This section explains the standard mask estimation using NNs. In this paper, all the variables in models are represented in the STFT domain with a frame index $t \in [1, T]$ and a frequency-bin index $w \in [1, W]$ [25].

A. Observation Model

The arrival process of the sound from M sound sources to two microphones, k_1 and k_2, is modeled as a linear system. The observed spectrum vector $\boldsymbol{x}_{w,t} = [x_{k_1,w,t}, x_{k_2,w,t}]^{\mathsf{T}} \in \mathbb{C}^2$ is represented as

$$\boldsymbol{x}_{w,t} = \sum_{m=1}^{M} \mathbf{h}_w(\boldsymbol{r}_{m,t}) s_{m,w,t} + \boldsymbol{n}_{w,t}, \tag{1}$$

where $s_{m,w,t}$ represents an m-th source sound spectrum and $\boldsymbol{n}_{w,t} = [n_{k_1,w,t}, n_{k_2,w,t}]^{\mathsf{T}} \in \mathbb{C}^2$ is a noise spectrum vector. Here, $\boldsymbol{n}_{w,t}$ may also include late-reverberations. $\boldsymbol{r}_{m,t}$ is a position of the m-th source at frame t, and its simplest representation is an azimuth $\theta_{m,t} \in \mathbb{R}$ from the center of the microphones. $\mathbf{h}_w(\boldsymbol{r}) = [h_{k_1,w}(\boldsymbol{r}), h_{k_2,w}(\boldsymbol{r})]^{\mathsf{T}} \in \mathbb{C}^2$ represents a transfer function from the reference position, $\boldsymbol{r}$, to each microphone.

B. Real-valued Mask-based Separation using NNs

In a binaural sound source separation given a target source position $\boldsymbol{r}_{c,t}$ corresponding to the source index c, the target source spectrum $y_{w,t} = h_{k_1,w}(\boldsymbol{r}_{c,t}) s_{c,w,t}$ is estimated from the observed spectrum $\boldsymbol{x}_{w,t}$. Here, we aim to recover the source signal observed at microphone k_1 without loss of generality.

The real-valued mask approach achieves the separation by using mask function $m_{w,t} \in [0, 1]$ to extract the target source spectrum. The estimated target source spectrum, $\hat{y}_{w,t}$, becomes

$$\hat{y}_{w,t} = m_{w,t} x_{k_1,w,t}. \tag{2}$$

The $m_{w,t}$ is usually modeled by NNs with parameter set $\boldsymbol{\Theta}$, and it is a function of the target source direction $\boldsymbol{r}_{c,t}$ and the observed vector set $\boldsymbol{X}_t = \{\boldsymbol{x}_{w',t'}\}_{w'=1,..,W}^{t'=t-D,...,t+D}$ with a splicing parameter D, that is, $m_{w,t}(\boldsymbol{X}_t, \boldsymbol{r}_{c,t}; \boldsymbol{\Theta})$. **Since this model requires several frames to estimate the masks, it is suitable for online processing**.

The parameter set $\boldsymbol{\Theta}$ of NNs is optimized by using a training data set. In this paper, the loss function G is an L1-norm distance between the reference source spectrum $y_{w,t}$ and the estimated spectrum $\hat{y}_{w,t}$. $y_{w,t}$ can be synthesized by using sound corpora and impulse responses. The gradient for NNs is calculated by differentiating the loss function in the log amplitude domain, as

$$G(\boldsymbol{\Theta}) = \sum_{w,t} |\log|y_{w,t}| - \log|\hat{y}_{w,t}||. \tag{3}$$

The reason we use the logarithm is to relax the different dynamic ranges of the amplitude spectra among each frequency-bin.

C. Fixed Spatial Pre-Filtering and Directional Feature

The actual inputs to NNs are features extracted from the observed vector set $\boldsymbol{X}_t$. One of the major input features is based on a combination of $\boldsymbol{X}_t$ and the output of a DS beamformer to the target source direction $\boldsymbol{r}_{c,t}$ [7], [9]. We consider this feature as a baseline, but other beamformers can be applied such as the one used in [8].

The observed vectors are converted as

$$\boldsymbol{z}_{w,t} = [\boldsymbol{x}_{w,t}^{\mathsf{T}}, \boldsymbol{a}_w(\boldsymbol{r}_{c,t})^{\mathsf{H}} \boldsymbol{x}_{w,t}]^{\mathsf{T}}, \tag{4}$$

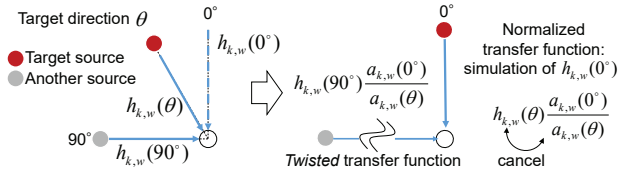

Fig. 2. Image of spatial normalization

where $\boldsymbol{a}_w(\boldsymbol{r}_{c,t}) = [a_{k_1,w}(\boldsymbol{r}_{c,t}), a_{k_2,w}(\boldsymbol{r}_{c,t})]^{\mathsf{T}}$ ($||\boldsymbol{a}_w(\boldsymbol{r}_{c,t})|| = 1$) is a steering vector of the direction $\boldsymbol{r}_{c,t}$, and $\cdot^{\mathsf{H}}$ is a conjugate transformation operator. Note that *the steering vectors are obtained by means of a physical model or simulated/measured impulse responses.* The amplitude spectra of $|\boldsymbol{z}_{w,t}|$ are used as intensity features for NNs. Here, the absolute function for a vector means an element-wise operation.

III. Proposed Method

This section explains the spatial normalization and the optimization of spatial pre-filters. The spatial normalization converts the observed spectrum into a canonical form, which enables us to optimize the spatial pre-filters.

A. Spatial Normalization

The *spatial normalization* reduces the essential complexity of sound source positions by converting the transfer function of the observed vector into a canonical form based on directional information. For example, when a target source stands at the angle θ in azimuth, we transform the observed vector as if the target sound source stands at a fixed *standard position*, e.g. 0° (Fig. 2).

We represent a target source position as $\boldsymbol{r}_{c,t}$ and the fixed *standard position* as $\boldsymbol{r}'$. We assume the steering vectors corresponding to these positions have been already prepared. The spatially normalized vector $\boldsymbol{x}'_{w,t}$ is represented by

$$\boldsymbol{x}'_{w,t} = \boldsymbol{x}_{w,t} \otimes \boldsymbol{a}_w(\boldsymbol{r}') \oslash \boldsymbol{a}_w(\boldsymbol{r}_{c,t}), \tag{5}$$

where $\otimes$ and $\oslash$ represent an element-wise product and division, respectively. Since the steering vector is a relative transfer function, the transfer function of the observed target source is cancelled out by $\boldsymbol{a}_w(\boldsymbol{r}_{c,t})$, and it becomes the transfer function of the standard position by $\boldsymbol{a}_w(\boldsymbol{r}')$.

This normalization can be applied to any existing method because all we do is convert the observed spectrum. Although the transfer function of other sound sources can be *twisted* and the spatial normalization can also include conversion errors, such mismatch or errors can be reduced by training the NNs.

B. Model of Spatial Pre-Filters and Optimization

The key contribution of the spatial normalization is that we can assume a target sound source is always located at the standard point. The spatial pre-filters can be optimized into a single best filter thanks to this property.

The transformed spectrum $\boldsymbol{x}'_{w,t}$ is filtered by complex-valued linear projection, which corresponds to the fixed beamformer. We represent the j-th filter as $\mathbf{w}_{j,w} = [w_{j,k_1,w}, w_{j,k_2,w}]^{\mathsf{T}} \in \mathbb{C}^2 (j = 1, .., J)$, where J is the number of filters and it controls the number of output elements. The new input for NNs is obtained by

$$\boldsymbol{z}_{w,t} = \mathbf{W}_w^{\mathsf{H}} \boldsymbol{x}'_{w,t} + \mathbf{b}_w, \tag{6}$$

where $\mathbf{W}_w = [\mathbf{w}_{1,w}, ..., \mathbf{w}_{J,w}]$ is a filter matrix and $\mathbf{b}_w \in \mathbb{C}^2$ is a bias vector. These parameters are optimized by back-propagation under the constraint of $||\mathbf{w}_{j,w}|| = 1$ $(j = 1, ..., J)$. $|\boldsymbol{z}_{w,t}|$ is used as the input of NNs. The optimum number of J can be determined by experiments.

Intuitively, the target source component $|y_{w,t}|$ could be estimated just by the differential information between the spectral enhancement and cancellation of the target source direction. The NNs are expected to capture such differential information and utilize it for mask estimation. Therefore, we investigate the frequency responses of the trained filters to understand the behavior.

IV. Experiments

Our aim with the experiments is to show the fundamental effectiveness of our spatial normalization for more than two sources including unknown non-speech signals under weak a reverberant environment. Note that highly reverberant speech can be overcome by preparing reverberant training data as monaural separation methods do [26], [27]. Our test sets are designed to be open in terms of speaker, kind of sound source, sound source position, and number of sound sources.

A. Training Set and Four Test Sets

We prepared two kinds of sound sources, human speech and non-speech signals, to check whether our proposed method is less dependent on signal characteristics or not. For the speech signals, we used speech data from the Corpus of Spontaneous Japanese (CSJ) [28]. The training set contains 223 hours of academic lecture presentations (by 799 men and 168 women). We selected speech signals from the official evaluation sets (eval2 and eval3) defined in the CSJ for test sets. This set contains 100 minutes of speech by ten men and ten women, and the length of the test signals ranges from three to ten seconds. For the non-speech signals, we selected ones from the RWCP Sound Scene Database in Real Acoustical Environments[1] [29] as a test set. This corpus includes about 60 kinds of non-speech signal, such as a bell sound.

All binaural data were synthesized by the impulse responses recorded in a real anechoic room to investigate the fundamental behavior of the spatial normalization. Note that reflections from the floor exist in such data. Two-channel impulse responses were recorded at 16 kHz by using microphones horizontally attached to an egg-shaped surface[2]. The resolution of the azimuth was 1° (360 directions), and there were two combinations of distance and height, as shown in Fig. 3. The steering vectors used in Eqs. (4) and (5) were the impulse

[1] http://research.nii.ac.jp/src/en/RWCP-SSD.html
[2] http://www.sifi.co.jp/system/modules/pico/index.php?content_id=39

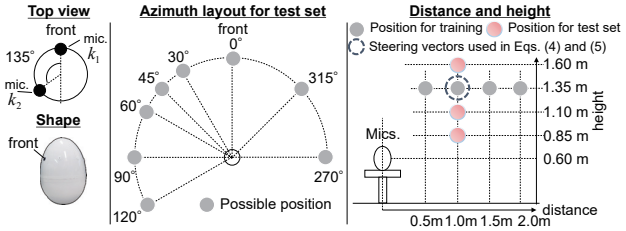

Fig. 3. Layout of microphones and source positions

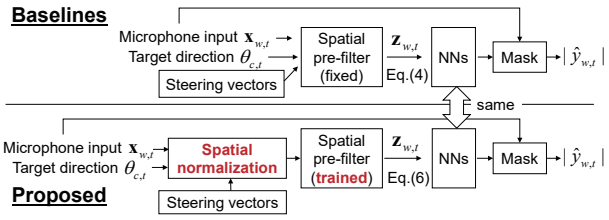

Fig. 4. Block-diagram of comparative methods

responses at the height of 1.35 m and the distance of 1.0 m. The impulse responses at the height of 1.35 m and the distance of 0.5, 1.0, 1.5, and 2.0 m were used for the training set, and those at the height of 0.85, 1.10, and 1.60 m were used for the test sets.

The training data were a mixture of three speakers' speech signals, and most of them consisted of different speakers' utterances. The target source position $\mathbf{r}_{c,t}$ was assumed as the azimuth $\theta_{c,t}$. The $\theta_{c,t}$ was assumed to be time-invariant and uniformly selected from 0° to 359°. The directions of the other two sound sources were randomly selected from $(\theta_{c,t} + 20 + u)^\circ$ and $(\theta_{c,t} + 340 - u)^\circ$, respectively, where $u \in [0, 145]$ was a randomly selected integer value.

The four types of test data were a mixture of two, three, and four sound signals that were not included in the training set. The height and distance of the sound sources in each mixture were same. A **2sp** set was a mixture of two speech signals, and the combinations of source directions were $[0^\circ, 30^\circ]$, $[0^\circ, 45^\circ]$, and $[0^\circ, 60^\circ]$. A **2sp+1n** set was a mixture of two speech and one non-speech signals, and the combinations of source directions were the same as those of **sp2**. A **3sp** set was a mixture of three speech signals, and the combinations of source directions were $[0^\circ, 30^\circ, 60^\circ]$, $[0^\circ, 45^\circ, 90^\circ]$, and $[0^\circ, 60^\circ, 120^\circ]$. A **4sp** set was a mixture of four speech signals, and the combination of source directions was $[0^\circ, 45^\circ, 270^\circ, 315^\circ]$. Finally, real recorded background noise was added to each mixture with randomly selected signal-to-noise ratios from 30 to 15 dB.

B. Comparative Methods and NN Configurations

We tested two kinds of input feature for NNs without spatial normalization (Spatial norm.) to be evaluated with our methods. The first is the two-channel observed vector and DS beamformer described by Eq. (4) in Section II-C. The second is the trainable spatial pre-filters described in Section III-B, and the parameters $\mathbf{W}_w$ were randomly initialized. We applied our spatial normalization to these two kinds of input data as depicted in Fig. 4. Each sound source was separated independently by changing the target direction $\theta_{c,t}$. The *standard direction* was set to 0°, and accordingly the DS beamformer after spatial normalization always focused on the direction 0°. The target direction was set to $(\text{ground truth azimuth} + 3)^\circ$ when separating data in the test sets.

The NN configuration was the same among all methods in this paper. It closely resembles the one described in [30] in that its structure follows a traditional speech feature extraction process and thus is explained only briefly here. The NN consisted of a feature-extraction network and a fully connected network. The former network was for Mel-filterbank feature extraction, and the parameters were optimized by back-propagation. In this experiment, it used the functions in the following order: **10-ms frame shift**, fast Fourier transform (**512 points window**), absolute, linear projection (filterbank, 64 dim.), absolute, power (instead of log), frame concatenation, and linear projection (bottleneck, 256 dim.). The splicing parameter was $D = 32$, which corresponds to a 640-ms duration. The fully connected network had seven layers with a sigmoid function. The output layer with 256 dim. (D.C. was removed) was a sigmoid function to represent the mask $m_{w,t} \in [0, 1]$. We used AdaGrad [31] for the optimization, and the number of epochs was 18.

C. Evaluation Criteria: SDR and CD

We used the signal-to-distortion ratio improvement (SDRi) and cepstral distortion (CD) as evaluation metrics.

The SDRi measures the improvement of the SDR of the separated signals compared to that of the observed signal. A higher SDRi indicates a better performance. The SDR was calculated in the amplitude spectrum domain as $|\hat{y}_{w,t}| = |y_{w,t}| + e_{w,t}$. The SDR is the ratio of the log total power of $|y_{w,t}|$ and $|e_{w,t}|$ over w and t.

The CD is calculated by using the cepstral coefficients computed by applying discrete cosine transformation to the log-amplitude spectrum, where a lower value indicates a better performance. We used the dimensions of the coefficients to range from 1 to 24, and the mean absolute error between reference and separated speech was calculated.

SDRi and CD were averaged over all separated sound sources for each test set. If the input data includes several sound sources, we separate each source from the others independently in accordance with its source direction.

D. Results

Table I shows the SDRi and CD of each method with different mixture signals. The "Ave." column means the averaged performance over four test set. J is the output dimension of the spatial pre-filter described in Section III-B. The original SDRs of the observed signal for 2sp, 2sp+1n, 3sp, and 4sp were 0.63, -5.67, -3.16, and -5.20, respectively. The CDs of the observed signal for 2sp, 2sp+1n, 3sp, and 4sp were 2.45, 3.59, 3.06, and 3.41, respectively.

TABLE I
EXPERIMENTAL RESULTS. AVERAGED SDRI IN DB AND CEPSTRAL DISTORTION (CD).

	Configuration		SDRi					Cepstral distortion				
	Spatial norm.	Spatial pre-filter	2sp	2sp+1n	3sp	4sp	Ave.	2sp	2sp+1n	3sp	4sp	Ave.
Baseline		DS ($J = 3$)	5.38	7.36	7.46	6.80	6.75	1.65	2.42	1.91	2.32	2.07
Proposed	✓	DS ($J = 3$)	6.41	9.74	8.75	8.13	8.26	1.54	2.16	1.74	2.13	1.89
	✓	Train ($J = 3$)	7.10	9.91	9.09	8.22	8.58	1.50	2.17	1.73	2.10	1.87
	✓	Train ($J = 2$)	6.50	9.50	8.84	8.04	8.22	1.55	2.20	1.75	2.14	1.91
	✓	Train ($J = 6$)	**7.52**	**10.25**	**9.50**	8.39	**8.92**	**1.48**	2.14	1.70	**2.08**	1.85
	✓	Train ($J = 7$)	7.24	10.19	9.40	**8.46**	8.82	**1.48**	**2.12**	**1.69**	**2.08**	**1.84**
For comparison		Train ($J = 3$)	0.92	3.41	3.89	5.05	3.32	2.13	2.95	2.42	2.68	2.54

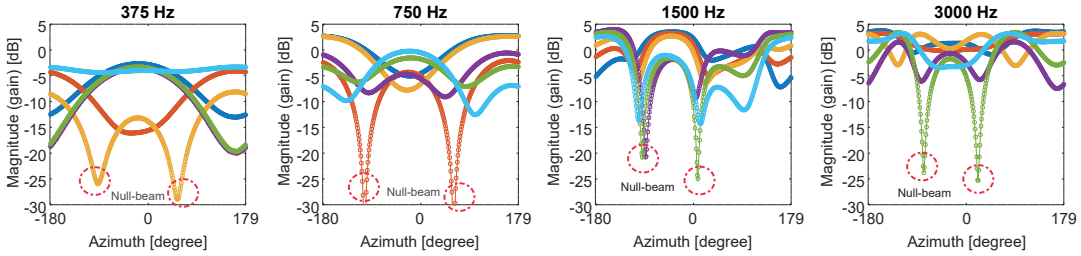

Fig. 5. Frequency response of optimized spatial pre-filters in dB at 375, 750, 1500 and 3000 Hz: each line corresponds to the response of the filter $\mathbf{w}_{j,w}$ ($J = 6$ and $j = 1, ..., 6$).

We found that our normalization and trained pre-filter improved the separation performance from the baseline even in the case of non-speech sound (2sp+1n) and a four-speaker set (4sp). The results under same condition, $J = 3$, showed that the spatial normalization outperformed the baseline by 1.0 to 2.3 dB in SDRi and by 0.1 to 0.26 in CD. The training of the spatial pre-filter with spatial normalization also improved the SDRi and CD. As shown in "For comparison" row, the spatial pre-filter training failed without spatial normalization. The reason why SDRi for 2sp and 4sp was worse than that for 3sp is that the mixture of two and four sound sources was not included in the training set. These results demonstrate that training the NNs with two or four source mixtures will improve the SDRi and CD.

We also found that the SDRi and CD of our method improved as the number of dimensions J increased. When we investigated the performances with $J = 2, 3, ..., 10$, the best averaged SDRi and CD were obtained with $J = 6$ and $J = 7$, respectively. Since the computational cost of spatial pre-filtering is $O(J^2)$, the J should be decided by considering the trade-off between computational cost and performance.

The directivity of the optimized pre-filters had a complementary beam patterns, i.e. the combination of the enhancement and the cancellation (null-beam) patterns. Figure 5 shows the magnitude responses (gain) in dB of each channel of the trained filters with $J = 6$, $\mathbf{w}_{j,w} (j = 1, ..., 6)$. The horizontal axis denotes azimuth (in degree), and the target source was assumed to locate at 0°. The impulse responses at the height of 1.35 m and the distance of 1.0 m were used for this analysis. Note that these amplitude responses were influenced by the spatial normalization term in Eq. (5). Each trained filter had different directive patterns, and some of them had null-directivities (blind spots) at each frequency. This result indicates that the NNs can find a target source not only by the enhancement patterns but also by the cancellation patterns in terms of direction. Therefore, the performance of the separation may not improve even if we add more DS-like filter channels to the spatial pre-filter of the baseline method. The asymmetric beam patterns were caused partly by the microphone arrangement and the shape of the device.

V. CONCLUSION

In this paper, we proposed a novel spatial normalization of observed signals for direction-aided supervised sound source separation. The proposed normalization converts the transfer functions of the observed signals into a canonical form, which improves the mask estimation and achieves the optimization of the spatial pre-filters. Experiments demonstrated the effectiveness of our spatial normalization and its independence from signal characteristics.

Remaining issues and future work include examining the reverberation and narrow angle cases of speaker positions. Data generation and augmentation is also be promising to alleviate these problems because the performance of the NNs improves as the amount of data increases. We will also investigate the separation performance with the optimization of the steering vectors used in the spatial normalization.

REFERENCES

[1] A.S. Bregram, *Auditory scene analysis*, MIT Press, 1990.
[2] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.
[3] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
[6] X. Sun, R. Xia, J. Li, and Y. Yan, "A deep learning based binaural speech enhancement approach with spatial cues preservation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 5766–5770.
[7] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.
[8] R. Gu, Lianwu Chen, S. Zhang, J. Zheng, Yinlong Xu, M. Yu, Dan Su, Y. Zou, and Dong Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. of Interspeech*, 2019, pp. 4290–4209.
[9] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi, "Deep speech extraction with time-varying spatial filtering guided by desired direction attractor," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 671–675.
[10] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. of IEEE Spoken Language Technology Workshop*, 2018, pp. 558–565.
[11] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," in *Proc. of the International Joint Conference on Neural Networks*, 2001, vol. 4, pp. 2861–2866.
[12] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, pp. 175–179.
[13] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.
[14] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.
[15] D. Yu, M. Kolbak, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 241–245.
[16] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 31–35.
[17] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.
[18] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.
[19] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. of IEEE Global Conference on Signal and Information Processing*, 2014, pp. 577–581.
[20] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *Proc. of Interspeech*, 2018, pp. 342–346.
[21] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech and Lanugage Processing*, vol. 26, no. 4, pp. 787–796, 2018.
[22] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 46–50.
[23] Yang Yu, Wenwu Wang, and Peng Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP Journal Audio Speech Music. Process.*, vol. 2016, pp. 7, 2016.
[24] Z. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 1–5.
[25] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 85–88.
[26] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal*, 2020, pp. 46–50.
[27] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 696–700.
[28] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.
[29] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound scene data collection in real acoustical environments," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 225–231, 1999.
[30] R. Takeda, K. Nakadai, and K. Komatani, "Multi-timescale feature-extraction architecture of deep neural networks for acoustic model training from raw speech signal," in *Proc .of IEEE/RSJ International Conference on Intelligent Robots and System*, 2018, pp. 2508–2510.
[31] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.