NORMALIZED OBSERVATION VECTOR CLUSTERING APPROACH FOR SPARSE SOURCE SEPARATION

Shoko Araki[†‡] *Hiroshi Sawada*[†] *Ryo Mukai*[†] *Shoji Makino*[†‡]

† NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
‡ Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo-shi, Hokkaido 060-0814, Japan
Email: shoko@cslab.kecl.ntt.co.jp

ABSTRACT

This paper presents a new method for the blind separation of sparse sources whose number N can exceed the number of sensors M. Recently, sparseness based blind separation has been actively studied. However, most methods utilize a linear sensor array (or only two sensors), and therefore have certain limitations; e.g., they cannot be applied to symmetrically positioned sources. To allow the use of more than two sensors that can be arranged in a non-linear/non-uniform way, we propose a new method that includes the normalization and clustering of the observation vectors. We report promising results for the speech separation of 3-dimensionally distributed five sources with a non-linear/non-uniform array of four sensors in a room (RT_{60}= 120 ms).

1. INTRODUCTION

Blind source separation (BSS) is an approach for estimating source signals that uses only the mixed signal information observed at each sensor. The BSS technique of speech focused on in this paper, has many applications including hands-free teleconference systems and automatic conference minute generators. Such applications usually have to deal with underdetermined situations where the N sources outnumber the M sensors ($N > M$).

In this paper, we propose a novel BSS method that can handle the underdetermined convolutive BSS problem. Two approaches have been widely studied and employed to solve such a BSS problem; one is based on independent component analysis (ICA) (e.g.,[1]) and the other relies on the sparseness of source signals (e.g., [2]). Recently, many ICA methods have been proposed for the convolutive BSS problem (e.g., [1, 3–5]), however, ICA cannot be applied when $N > M$. Some [6, 7] handle cases where $N > M$ with ICA, however, they only separate a few dominant sources and still cannot separate all of the sources.

On the other hand, a sparseness based method can separate all N sources even when $N > M$. There are several approaches [2, 8–13] that rely on the sparseness of the source signals. If the signals are sufficiently sparse, we can assume that the sources rarely exist simultaneously. Therefore, we can estimate each source by collecting observation samples that appear to belong to one of the sources. Previously, such observation samples were evaluated by using geometric information, which is estimated from *only two* sensor observations. Some authors used the level difference [8–10] or phase difference [11] between two observations, and some employed both the level difference and phase difference between two sensor observations [2, 12].

However, it is difficult to extend these methods to $M \geq 3$, especially for a non-linearly arranged sensor array. A two sensor system (or a linear sensor array) limits the separation ability on a 2-dimensional half-plane, e.g., the previous methods cannot separate sources placed at a mirror image point. To allow the free location of sources, we need more than two 2- or 3-dimensionally arranged sensors.

Although the method in [13] is generalized for more than two sensors case, they work in *each* frequency bin to handle convolutive mixtures. Therefore, they have to solve so called the permutation problem after separating the signals at each frequency. This can cause errors and degrade the separation performance. Furthermore, when the observation data length is short (e.g., with an on-line implementation and moving source tracking), working at each frequency is inefficient because fewer sample data are available at each frequency. To avoid such problems, a frequency normalization should be employed to make it possible to handle all the frequency components together.

In this paper, we propose a new source separation method that can be applied to multiple sensors arranged non-linearly. First, we normalize all the observations with regard to one of the observations. Our normalization also includes frequency normalization. The normalized observation vectors maintain the level and phase difference information of all the sensor pairs. Then, we cluster the normalized observation vectors. This clustering is executed for whole frequency components. Finally, we design binary masks using the clustering result and estimate the separated signals with the masks.

With our proposed separation method, we do not need the exact sensor locations, but simply the maximum distance between a given sensor and any other sensor. This allows us to use a non-linear/non-uniform arrangement of multiple sensors. Therefore, we can separate signals that are distributed 2- or 3-dimensionally. In addition, by employing frequency normalization, we can handle all the frequency components together, and therefore obtain a promising result even when only short observations are available. We show experimental results obtained in a real room (reverberation time of 120 ms) with non-linear sensor arrays in underdetermined scenarios (#sources $\times$ #sensors $= 3 \times 4$ or 4×5).

2. PROPOSED APPROACH

2.1 Problem description

Suppose that sources $\mathrm{s}_1, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at M sensors

$$\mathrm{x}_j(t) = \sum_{i=1}^{N} \sum_l \mathrm{h}_{ji}(l)\,\mathrm{s}_i(t-l),\ j=1,\ldots,M, \quad (1)$$

where $\mathrm{h}_{ji}(l)$ represents the impulse response from source i to sensor j. In this paper, we focus particularly on a situation where the number of sources N can exceed the number of sensors M ($N > M$). We assume that N and M are known. The goal is to obtain separated signals $\mathrm{y}_k(t)$ that are estimations of s_i solely from M observations.

2.2 Frequency domain operation

Figure 1 shows the flow of our method. First, time-domain signals $\mathrm{x}_j(t)$ sampled at frequency f_s are converted into frequency-domain time-series signals $x_j(f,\tau)$ with an L-point short-time Fourier transform (STFT):

$$x_j(f,\tau) \leftarrow \sum_{r=-L/2}^{L/2-1} \mathrm{x}_j(\tau+r)\,\mathrm{win}(r)\,e^{-\jmath 2\pi f r}, \quad (2)$$

where $f \in \{0, \frac{1}{L}f_s, \ldots, \frac{L-1}{L}f_s\}$ is a frequency, $\mathrm{win}(r)$ is a window that tapers smoothly to zero at each end, such as a Hanning window $\frac{1}{2}(1+\cos\frac{2\pi r}{L})$, and τ is a time index.

The remaining operations are performed in the frequency domain. There are two advantages to this. First, convolutive mixtures (1) can be approximated as instantaneous mixtures at each frequency:

$$x_j(f,\tau) \approx \sum_{i=1}^{N} h_{ji}(f)s_i(f,\tau), \quad (3)$$

or in vector notation,

$$\mathbf{x}(f,\tau) \approx \sum_{i=1}^{N} \mathbf{h}_i(f)s_i(f,\tau), \quad (4)$$

where $h_{ji}(f)$ is the frequency response from source i to sensor j, and $s_i(f,\tau)$ is a frequency-domain time-series signal of $\mathrm{s}_i(t)$ obtained by the same operation as (2), $\mathbf{x} = [x_1,\ldots,x_M]^T$ is *an observation vector* and $\mathbf{h}_i = [h_{1i},\ldots,h_{Mi}]^T$ is a mixing vector that consists of the frequency responses from source s_i to all sensors.

The second advantage is that the sparseness of a source signal becomes prominent in the time-frequency domain [2, 8, 10, 12], if the source is colored and non-stationary such as speech. The possibility of $s_i(f,\tau)$ being close to zero is much higher than that of $\mathrm{s}_i(t)$. When the signals are sufficiently sparse in the time-frequency domain, we can assume that the sources rarely overlap, and (3) and (4) can be approximated respectively as

$$x_j(f,\tau) \approx h_{jk}(f)s_k(f,\tau), \quad k \in \{1,\cdots,N\}, \quad (5)$$
$$\mathbf{x}(f,\tau) \approx \mathbf{h}_k(f)s_k(f,\tau), \quad k \in \{1,\cdots,N\}, \quad (6)$$

where $s_k(f,\tau)$ is a dominant source at the time-frequency point (f,τ).

2.3 Proposed separation procedures

2.3.1 Normalization

The new method involves normalizing all observation vector components $x_j(f,\tau)$ ($j = 1,\ldots,M$) for all frequency bins ($f = 0, \frac{1}{L}f_s, \ldots, \frac{L-1}{L}f_s$) such that they form clusters, each of which corresponds to an individual source. Our normalization procedure includes phase-normalization, frequency-normalization and unit-norm normalization.

First, with phase-normalization, we eliminate the phase inconstancy due to the scalar $s_k(f,\tau)$ in the observation samples (5). This can be normalized by taking the ratio of two observation components x_j and one arbitrarily selected observation x_J: $x_j(f,\tau)/x_J(f,\tau) \approx h_{jk}(f,\tau)/h_{Jk}(f,\tau)$. Then, we perform frequency-normalization because the phase of $h_{jk}(f,\tau)/h_{Jk}(f,\tau)$ is proportional to the frequency if we can assume that $h_{jk}(f,\tau)$ has linear phase characteristics. This normalization allows us to handle all the frequency components together. The frequency is normalized by dividing the phase of $x_j(f,\tau)/x_J(f,\tau)$ by f.

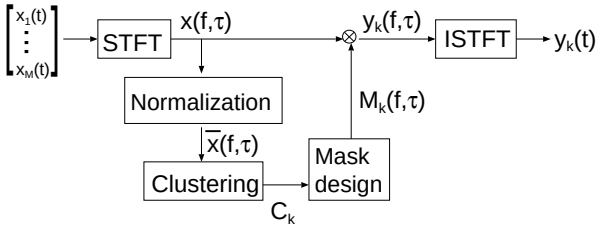

Figure 1: Flow of proposed method.

Our normalization, which includes the above normalizations, is performed for the all components of $\mathbf{x}(f,\tau)$ by selecting one of the sensors J:

$$\bar{x}_j(f,\tau) \leftarrow |x_j(f,\tau)| \exp\left[-\jmath \frac{\arg[x_j(f,\tau)/x_J(f,\tau)]}{4fc^{-1}d_{\max}}\right] \quad (7)$$

where c is the propagation velocity and $d_{\max}$ is the maximum distance between sensor J and sensor $^{\forall}j \in \{1,\ldots,M\}$. This normalization makes the phase zero at sensor J. The rationale for the frequency normalization $4fc^{-1}d_{\max}$ can be found in Section 2.4. Here, we maintain the amplitudes at all sensors to utilize the level difference information.

We also apply unit-norm normalization to facilitate clustering,

$$\bar{\mathbf{x}}(f,\tau) \leftarrow \bar{\mathbf{x}}(f,\tau) \,/\, ||\bar{\mathbf{x}}(f,\tau)|| \quad (8)$$

for $\bar{\mathbf{x}}(f,\tau) = [\bar{x}_1(f,\tau),\ldots,\bar{x}_M(f,\tau)]^T$.

We describe the properties of the normalized observation vector in Section 2.4.

2.3.2 Clustering

The next step is to find clusters $C_1,\ldots,C_N$ formed by all normalized vectors $\bar{\mathbf{x}}(f,\tau)$. Note that the normalized vectors $\bar{\mathbf{x}}(f,\tau)$ are complex M-dimensional vectors, and therefore the clustering is carried out in an M-dimensional space.

After setting appropriate initial centroids $\mathbf{c}_k$ ($k = 1,\cdots,N$), clustering is realized by the following iterative updates:

$$C_k = \{\bar{\mathbf{x}}(f,\tau) \mid k = \mathrm{argmin}_i ||\bar{\mathbf{x}}(f,\tau) - \mathbf{c}_i||^2\} \quad (9)$$
$$\mathbf{c}_k \leftarrow E[\bar{\mathbf{x}}(f,\tau)]_{\bar{\mathbf{x}}\in C_k},\ \mathbf{c}_k \leftarrow \mathbf{c}_k/||\mathbf{c}_k||, \quad (10)$$

where $E[\cdot]_{\bar{\mathbf{x}}\in C_k}$ is a mean operator for the members of a cluster C_k. That is the cluster members are determined by (9) and their centroid is calculated by (10). This minimization can be performed efficiently with the k-means clustering algorithm [14] with a given source number N.

2.3.3 Mask design and separated signal reconstruction

Finally, we design a time-frequency binary mask that extracts the time-frequency points in one of the clusters

$$M_k(f,\tau) = \begin{cases} 1 & \bar{\mathbf{x}}(f,\tau) \in C_k \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

and obtain the separated signals $y_k(f,\tau)$ by

$$y_k(f,\tau) = M_k(f,\tau)x_{J'}(f,\tau)$$

where $J' \in \{1,\cdots,M\}$ is a selected sensor index.

At the end of the flow, we obtain outputs $\mathrm{y}_k(t)$ by an inverse STFT (ISTFT):

$$\mathrm{y}_k(\tau+r) \leftarrow \frac{1}{L\cdot \mathrm{win}(r)} \sum_{f\in\{0,\frac{1}{L}f_s,\ldots,\frac{L-1}{L}f_s\}} y_k(f,\tau)\, e^{\jmath 2\pi f r}. \tag{12}$$

2.4 Properties of normalized observation vector

We explain why normalized observation vectors $\bar{\mathbf{x}}(f,\tau)$ form a cluster that corresponds to an individual source.

Let us approximate the frequency response $h_{ji}(f)$ by using a direct-path (near-field) model

$$h_{jk}(f) \approx \frac{q(f)}{d_{jk}} \exp\left[-\jmath 2\pi f c^{-1}(d_{jk}-d_{Jk})\right], \tag{13}$$

where $d_{jk} > 0$ is the distance between source k and sensor j. We assume that the phase $2\pi f c^{-1}(d_{jk}-d_{Jk})$ depends on the distance difference to the reference sensor J. We also assume that the attenuation $q(f)/d_{jk}$ depends on both the distance and a frequency-dependent constant $q(f) > 0$.

Substituting (13) and (5) into (7) and (8) yields

$$\bar{x}_j(f,\tau) \approx \frac{1}{d_{jk}D}\exp\left[-\jmath\frac{\pi}{2}\frac{(d_{jk}-d_{Jk})}{d_{\max}}\right],\ D=\sqrt{\textstyle\sum_{j=1}^{M}\frac{1}{{d_{jk}}^2}}$$

We can see that the normalized observation vector $\bar{\mathbf{x}}(f,\tau)$ is independent of frequency, and dependent only on the positions of the sources and sensors due to the term $4fc^{-1}d_{\max}$ in (7). Therefore the observation vectors are clustered based on the source geometry. We can also see that the observation level information $\frac{1}{d_{jk}}$ remains, because our normalization (7 maintains the amplitude of each observation.

We also show the rationale behind the frequency normalization $4fc^{-1}d_{\max}$ in (7). The frequency normalization $4fc^{-1}d_{\max}$ provides us with an optimal argument property for the clustering. From the fact that $\max_{j,k}|d_{jk}-d_{Jk}| \le d_{\max}$, an inequality

$$-\pi/2 \le \arg[\bar{x}_j(f,\tau)] \le \pi/2 \tag{14}$$

holds. This property is important for two reasons. The first is that $|\bar{x}-\bar{x}'|$ increases monotonically as $|\arg(\bar{x})-\arg(\bar{x}')|$ increases. This is important for the distance measure in (9). The second reason is that the arguments of normalized observation vectors are the most scattered. This is a preferable property for small sensor array systems (e.g., see Section 3). If we use such systems, phase differences between sensors are more reliable than level differences for clustering. The frequency normalization with $4fc^{-1}d_{\max}$ allows us to make full use of the phase (argument) information.

3. EXPERIMENTS

3.1 Experimental conditions

We performed experiments to verify that our method can separate signals mixed in a reverberant condition. We measured impulse responses $\mathrm{h}_{jk}(l)$ under the conditions shown in Figs. 2 and 4. Mixtures were made by convolving the impulse responses and 5-second English speeches. The reverberation time of the room was RT_{60} =120 ms. The sampling rate was 8 kHz. The frame size L for STFT was 512, and we changed the frame shift from $256(=L/2)$ to $64(=L/8)$.

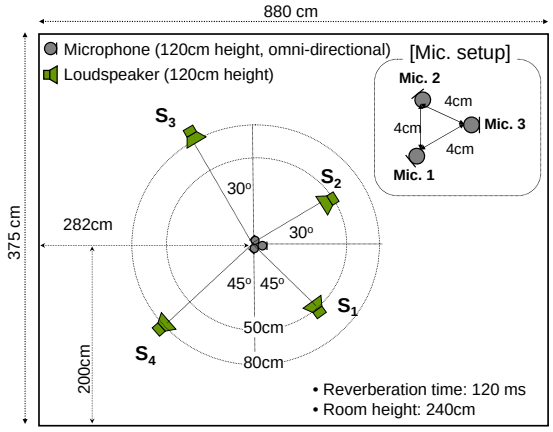

Figure 2: Experimental setup with a non-linear array

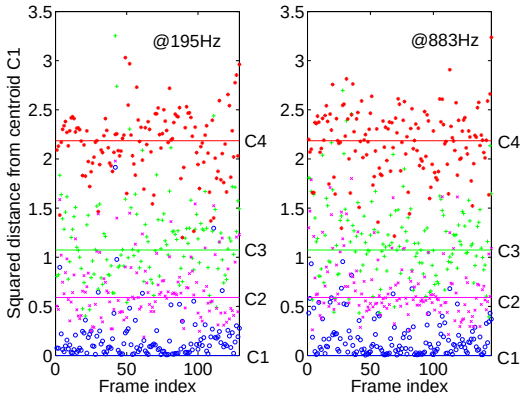

Figure 3: Example clustering result ($N=4, M=3$). o, x, +, * show the cluster members C_1, C_2, C_3 and C_4, respectively.

3.2 Performance measures

The separation performance was evaluated in terms of the improvement in the signal-to-interference ratio (SIR) for each output i. This improvement was calculated by $\mathrm{OutputSIR}_i - \mathrm{InputSIR}_i$, where

$$\mathrm{InputSIR}_i = 10\log_{10}\frac{\langle|\mathrm{x}_{J'i}(t)|^2\rangle_t}{\langle|\sum_{k\neq i}\mathrm{x}_{J'k}(t)|^2\rangle_t}\ \text{(dB)}, \tag{15}$$

$$\mathrm{OutputSIR}_i = 10\log_{10}\frac{\langle|\mathrm{y}_{ii}(t)|^2\rangle_t}{\langle|\sum_{k\neq i}\mathrm{y}_{ik}(t)|^2\rangle_t}\ \text{(dB)}, \tag{16}$$

where $\mathrm{x}_{J'k}(t) = \sum_l \mathrm{h}_{J'k}(l)\,\mathrm{s}_k(t-l)$ and $\mathrm{y}_{ik}(t)$ is the component of s_k that appears at output $\mathrm{y}_i(t)$: $\mathrm{y}_i(t) = \sum_{k=1}^{N}\mathrm{y}_{ik}(t)$. Moreover, we used the signal to distortion ratio (SDR) as a measure of sound quality:

$$\mathrm{SDR}_i = 10\log_{10}\frac{\langle|\mathrm{x}_{J'i}(t)|^2\rangle_t}{\langle|\mathrm{x}_{J'i}(t)-\alpha\mathrm{y}_{ii}(t-D)|^2\rangle_t}\ \text{(dB)}, \tag{17}$$

where α and D are parameters used to compensate for the amplitude and phase difference between $\mathrm{x}_{J'i}$ and y_{ii}. We investigated four combinations of speakers and averaged the results.

Table 1: Experimental results for $N = 4, M = 3$,

		y_1	y_2	y_3	y_4
InputSIR$_i$		-6.3	-6.4	-4.8	-2.3
Shift $L/2$	SIR$_i$	15.5	10.8	14.0	13.8
	SDR$_i$	5.0	4.7	6.1	7.3
Shift $L/4$	SIR$_i$	16.5	12.1	15.2	14.5
	SDR$_i$	5.6	5.5	6.9	8.0
Shift $L/8$	SIR$_i$	17.0	12.2	15.8	14.8
	SDR$_i$	5.8	5.6	7.1	8.3

3.3 Results

First, we show the result we obtained for four sources with three sensors that were arranged non-linearly (Fig. 2). Figure 3 shows an example clustering result for normalized observation vectors at two frequencies. Each point shows the squared distance $||\bar{\mathbf{x}} - \mathbf{c}_1||^2$ between normalized vectors $\bar{\mathbf{x}}$ and one of the centroids $\mathbf{c}_1$. We can see that the clustering was accomplished successfully using our clustering method. Moreover, it can be seen that the clustering is independent of frequency. Therefore, we can cluster all the frequency components together.

Table 1 shows the separation result. From Table 1, we can see that our proposed method achieved good separation even if we utilized a non-linear sensor arrangement. Table 1 also shows the SIR and SDR values when we changed the frame shift from $256(= L/2)$ to $64(= L/8)$. By using a fine-shift ($L/4$ and $L/8$), the SDR values increase without any reduction in the SIR values. This is because the fine-shift and the overlap-add realize a gradual change in the spectrogram of the separated signal [15].

We also applied our method to a non-uniform 3-dimensional sensor arrangement for a five sources and four sensors case (Fig. 4). Here, the system knew just the maximum distance ($d_{\max}$ =5.5 cm) between the reference microphone (Mic. 1) and the others. Table 2 shows the separation results. We can see from Table 2 that our proposed method can be applied to such a non-uniform 3-dimensional microphone array system.

We have also considered the musical noise problem, which usually occurs when we use a time-frequency binary mask like (11). The results of subjective tests can be found in [16]. Some sound examples can be found at [17].

4. DISCUSSIONS

In this section, we discuss the advantages of our method compared with some previous methods [2, 8, 11, 13].

4.1 Arbitrary source and sensor arrangements

The first advantage of the proposed method is that we do not need the exact sensor locations, but simply the maximum distance $d_{\max}$ between a given sensor and any other sensor. Even when we do not have the maximum distance, we can still use an arbitrary (slightly large) figure as $d_{\max}$, and employ our proposed normalization method. Therefore, we can utilize 2- or 3-dimensionally arranged sensors, that can be arranged in a non-linear/non-uniform way.

A previously adopted linear sensor array [2, 8, 11] limits the separation ability on a 2-dimensional half-plane: the previous methods cannot separate sources placed at symmetrical positions with respect to the sensor axis. On the other hand, because our proposed method makes it easy to employ a 2 or 3-dimensional non-linear sensor arrangement, we can cope with arbitrary source arrangements such as those shown in Figs. 2 and 4.

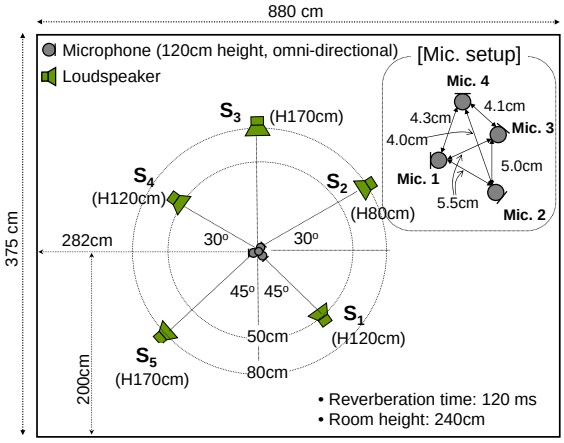

Figure 4: Experimental setup with a 3-D array

4.2 Avoidance of (inner-)permutation problem

The second advantage is that the proposed normalization of the observation vector allows us to cluster all the frequency components together. Previously, [13] has shown a separation result for more than two sensors case, however, they still worked in individual frequency bins. Therefore the (inner-)permutation problem remains and permutation error degrades the separation performance. By contrast, our frequency normalized observation vector clustering inherently avoids this problem.

4.3 Applicability to short observations

Moreover, frequency normalization allows us to utilize enough data samples and obtain good performance even if we use short observations. This applicability to short data is important e.g., for on-line implementations and moving source separation. On the other hand, if we do not apply frequency normalization, the clustering should be performed at each frequency, and therefore, fewer sample data are available for clustering than with frequency normalization. This can degrade the separation performance.

To confirm the effectiveness of frequency normalization to short observations, we separated some short speech signals with and without frequency normalization. The setup was the same as that shown in Fig. 2 and the frame size and frame shift were $L = 512$ and $256(= L/2)$, respectively. We utilized the same 5-second data as Section 3, however, we divided these data into some short (1 second or 0.5 seconds) blocks and separated them block by block. We evaluated the SIR improvements and SDR for whole 5-second outputs for four outputs $y_1, \cdots, y_4$ and averaged the results (Table 3).

With our proposed frequency normalization technique (labeled with "yes (Eq. 7)" on Table 3), we performed clustering by using all the frequency data samples. By contrast, for conditions without frequency normalization (labeled with "no (Eq. 18)" on Table 3), the observation vector was only

Table 2: Experimental results for $N = 5, M = 4$,

		y_1	y_2	y_3	y_4	y_5
InputSIR_i		-11.1	-3.0	-4.5	-10.6	-4.7
Shift $L/2$	SIR_i	18.4	13.7	6.5	15.5	16.0
	SDR_i	2.8	4.6	3.5	3.3	6.3
Shift $L/4$	SIR_i	20.1	15.0	6.9	16.6	17.7
	SDR_i	3.1	5.1	3.8	3.9	7.0
Shift $L/8$	SIR_i	20.7	15.5	6.9	17.2	18.2
	SDR_i	3.3	5.2	3.9	4.1	7.2

phase normalized as in [13]

$$\bar{x}_j(f,\tau) \leftarrow \arg[x_j(f,\tau)/x_J(f,\tau)], \qquad (18)$$

and clustering was executed frequency by frequency. We utilized the k-means algorithm for the clustering. If the clustering did not converge for a frequency, we set the binary mask (11) 0 for all k at that frequency. The average number of failed frequency bins per block is also shown in Table 3 (see rows labeled with "NG"). The permutation problem in the condition "no (Eq. 18)" was solved by clustering the (frequency normalized) centroids of each frequency and each block.

Table 3 summarizes the results. With frequency normalization ("yes (Eq. 7)"), even when we partitioned the observations into blocks of 0.5 seconds, we observed no degradation. On the other hand, without frequency normalization ("no (Eq. 18)"), the performance worsened as the block length decreased. The large degradation in SDR for the short block length reveals that clustering failed in many frequency bins in each block (see "NG" figures in the Table 3). The failure means that the k-means algorithm could not make N clusters. This is because there were fewer than N sources in each frequency due to source sparseness. The SIR degradation shows that the reliability of the clustering decreased because of the inadequate sample data.

From these results, we can say that our proposed method with frequency normalization is promising as regards the online implementation of underdetermined BSS.

5. CONCLUSION

We proposed a novel source separation method for cases where $N > M$ that assumes source sparseness. The method is based on the normalization and clustering of the observation vectors. Our proposed normalization technique makes it easy to employ multiple sensors arranged in a non-linear/non-uniform way. We obtained promising experimental results under reverberant underdetermined conditions even when we utilized observations of less than one second.

If we know the sensor locations, we can also estimate the directions of arrival of sources by using the cluster centroids obtained with our proposed method. The method and results can be found in [18].

Table 3: Effect of frequency normalization ($N = 4, M = 3$). BL: block length, NG: average number of failed frequency bins per block,

BL	averaged	yes (Eq. 7)	no (Eq. 18)
5 sec.	SIR_i [dB]	13.5	13.8
	SDR_i [dB]	5.8	6.1
	NG [bins]	0	0
1 sec.	SIR_i [dB]	13.6	12.4
	SDR_i [dB]	5.8	4.8
	NG [bins]	0	5.6
0.5 sec.	SIR_i [dB]	13.4	11.4
	SDR_i [dB]	5.3	2.5
	NG [bins]	0	64.7

REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[2] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on SP*, vol. 52, no. 7, pp. 1830–1847, 2004.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[4] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer, Mar. 2005, pp. 299–327.

[6] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of a dominant source signal from mixtures of many sources," in *Proc. ICASSP2005*, vol. III, Mar. 2005, pp. 61–64.

[7] H. Sawada, R. M. S. Araki, and S. Makino, "Real-time blind extraction of dominant target sources from many background interferences," in *Proc. IWAENC2005*, Sept. 2005, pp. 73–76.

[8] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," in *Proc. ICA2000*, 2000, pp. 87–92.

[9] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing*, vol. 55, pp. 627–641, 2003.

[10] A. Blin, S. Araki, and S. Makino, "Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1693–1700, 2005.

[11] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," in *Proc. ICASSP 2004*, vol. III, May 2004, pp. 881–884.

[12] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

[13] J. M. Peterson and S. Kadambe, "A probabilistic approach for blind source separation of underdetermined convolutive mixtures," in *Proc. ICASSP 2003*, vol. VI, 2003, pp. 581–584.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley Interscience, 2000.

[15] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc. ICASSP2005*, vol. III, Mar. 2005, pp. 81–84.

[16] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC 2005)*, Sept. 2005, pp. 117–120.

[17] http://www.kecl.ntt.co.jp/icl/signal/araki/xcluster_fine.html

[18] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. ICASSP2006*, May 2006, (accepted).