A NOVEL BLIND SOURCE SEPARATION METHOD WITH OBSERVATION VECTOR CLUSTERING

Shoko Araki, Hiroshi Sawada, Ryo Mukai, and Shoji Makino

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{shoko,sawada,ryo,maki}@cslab.kecl.ntt.co.jp

ABSTRACT

We propose a new method for separating sparse signals from their mixtures. Separation is achieved by clustering the normalized observation vectors and extracting each cluster as each separated signal. We show the practical result of speech separation with non-linear sensor arrangements in both a determined and an underdetermined scenarios. We also consider the musical noise problem and show the listening test results.

1. INTRODUCTION

In this paper, we consider the blind source separation (BSS) of speech signals observed in a real environment, i.e., the BSS of convolutive mixtures of speech. Recently, independent component analysis (ICA) [1] has been widely studied for such a BSS problem. However, ICA cannot be applied when $N > M$. In contrast, we propose a method that can handle both a (over-)determined ($N \leq M$) and an underdetermined ($N > M$) cases.

Let us formulate the task. Suppose that sources $\mathrm{s}_1, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at M sensors

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

where $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j. Here we assume that sources s_k are sparse signals, i.e., they have super-Gaussian distributions. For instance this is true for speech signals. The goal is to obtain the separated signals $\mathrm{y}_k(t)$ that are estimations of s_k only from the M observations.

There are several approaches [2, 3, 4] 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 gathering the observation samples that appear to belong to one of the sources. Previously, this was done by using geometric information (e.g., direction of arrival (DOA) and/or distance) about the sources, which is estimated by the phase and/or level difference between two observations of a linear sensor array [2, 5]. However, it is difficult to expand those methods for $M \geq 3$. That is, even if we have $M \geq 3$ microphones, the previous methods cannot utilize all the observation information effectively and straightforwardly. In addition, the previous methods needed an exact sensor arrangement and sensor calibration to estimate the geometric features of the sources accurately.

In this paper, we propose a new method for separating sparse signals from their mixtures. First we normalize all the observations and cluster the normalized observation vectors (see Eq. (5)). Then, we design time-frequency binary masks using the clustering result and estimate the separated signals with the masks. With this approach, we can exploit the information obtained from all the sensors. Moreover, we do not need to know the exact sensor locations, simply the maximum distance between a given sensor and any other sensor. This relaxation makes it easy to use a non-uniform arrangement of sensors, and also eliminates the need for sensor calibration. We show the experimental results obtained in a room (reverberation time of 130 ms) with non-linear sensor arrays in both a determined and an underdetermined scenarios.

Previously, we have applied the normalization and clustering techniques to the basis vectors produced by ICA [6] in order to overcome the permutation problem that we face in frequency domain ICA. In contrast, in this paper, we normalize and cluster the observation vectors themselves and separate the signals directly.

We also consider the musical noise problem, which usually occurs when we use a time-frequency binary mask. We confirm that our reported fine-shift and overlap-add method [7] is also applicable to our observation vector clustering method to reduce musical noise.

2. PROPOSED APPROACH

2.1. 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 new index representing time.

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:

$$x_j(f,\tau) \approx \sum_{k=1}^{N} h_{jk}(f) s_k(f,\tau), \quad (3)$$

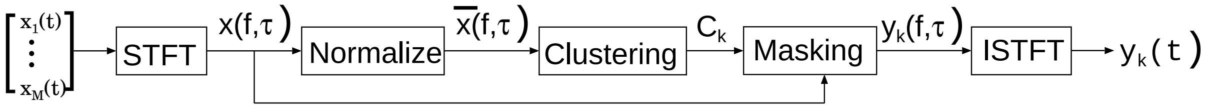

Figure 1: Flow of proposed method

where $h_{jk}(f)$ is the frequency response from source k to sensor j, and $s_k(f,\tau)$ is a frequency-domain time-series signal of $\mathrm{s}_k(t)$ obtained by the same operation as (2). The second advantage is that the sparseness of a source signal becomes prominent in the time-frequency domain if the source is colored and non-stationary such as speech. The possibility of $s_k(f,\tau)$ being close to zero is much higher than that of $\mathrm{s}_k(t)$. When the signals are sufficiently sparse in the time-frequency domain, we can assume that the sources rarely overlap and (3) can be approximated as

$$x_j(f,\tau) \approx h_{jk}(f)s_k(f,\tau), \quad k \in \{1,\cdots,N\}, \tag{4}$$

where $s_k(f,\tau)$ is a dominant source at the time-frequency point (f,τ). We estimate which source is dominant at each time-frequency point (f,τ) by using the procedures described in the following subsection.

2.2. Separation procedures

Let us have a vector notation of the mixing model (3):

$$\mathbf{x}(f,\tau) \approx \textstyle\sum_{k=1}^{N} \mathbf{h}_k(f)s_k(f,\tau), \tag{5}$$

where $\mathbf{x} = [x_1,\ldots,x_M]^T$ is an observation vector and $\mathbf{h}_k = [h_{1k},\ldots,h_{Mk}]^T$ is the vector of the frequency responses from source s_k to all sensors.

2.2.1. Normalization

The new method involves normalizing all observation vectors $\mathbf{x}(f,\tau)$, $j = 1,\ldots,M$, 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. The normalization is performed by selecting a reference sensor J and calculating

$$\bar{x}_j(f,\tau) \leftarrow |x_j(f,\tau)| \exp\left[\jmath \frac{\arg[x_j(f,\tau)/x_J(f,\tau)]}{4fc^{-1}d_{\max}}\right] \tag{6}$$

where c is the propagation velocity and $d_{\max}$ is the maximum distance between the reference sensor J and a sensor ${}^{\forall}j \in \{1,\ldots,M\}$. Then, we apply unit-norm normalization

$$\bar{\mathbf{x}}(f,\tau) \leftarrow \bar{\mathbf{x}}(f,\tau)\,/\,||\bar{\mathbf{x}}(f,\tau)|| \tag{7}$$

for $\bar{\mathbf{x}}(f,\tau) = [\bar{x}_1(f,\tau),\ldots,\bar{x}_M(f,\tau)]^T$. By this normalization, $\bar{\mathbf{x}}(f,\tau)$ becomes independent of frequency, and dependent only on the positions of the sources and sensors. That is, the observation vectors are clustered based on the source geometry. The rationale for this is explained in the Appendix.

2.2.2. Clustering

The next step is to find clusters $C_1,\ldots,C_M$ formed by normalized vectors $\bar{\mathbf{x}}(f,\tau)$. The centroid $\mathbf{c}_k$ of a cluster C_k is calculated by

$$\mathbf{c}_k \leftarrow \textstyle\sum_{\bar{\mathbf{x}}\in C_k} \bar{\mathbf{x}}/|C_k|, \;\; \mathbf{c}_k \leftarrow \mathbf{c}_k/||\mathbf{c}_k||,$$

where $|C_k|$ is the number of vectors in C_k. Each cluster corresponds to an individual source. The clustering criterion is to minimize the total sum $\mathcal{J}$ of the squared distances between cluster members and their centroid

$$\mathcal{J} = \textstyle\sum_{k=1}^{M} \mathcal{J}_k, \; \mathcal{J}_k = \sum_{\bar{\mathbf{x}}\in C_k} ||\bar{\mathbf{x}} - \mathbf{c}_k||^2. \tag{8}$$

This minimization can be performed efficiently with the k-means clustering algorithm [8].

2.2.3. Reconstruction of each separated signal

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

$$M_k(f,\tau) = \begin{cases} 1 & \bar{\mathbf{x}}(f,\tau) \in C_k \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

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 have outputs $\mathrm{y}_k(t)$ by an inverse STFT (ISTFT):

$$\mathrm{y}_k(\tau + r) \leftarrow \frac{1}{L\cdot\mathrm{win}(r)} \sum_{f\in\{0,\,\frac{1}{L}f_s,\,\ldots,\,\frac{L-1}{L}f_s\}} y_k(f,\tau)\,e^{\jmath 2\pi f r}. \tag{10}$$

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$_{60}$ =130 ms. The sampling rate was 8 kHz. The frame size L for STFT was 512, and we changed the frame shift from $64(= L/8)$ to $256(= L/2)$ to observe its effect on musical noise.

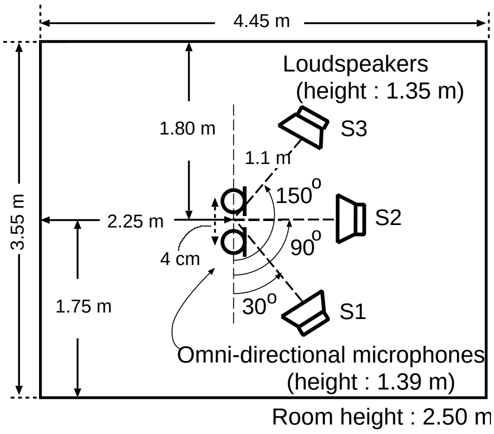

Figure 2: Room setup for primary experiment

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 $\mathsf{OutputSIR}_i - \mathsf{InputSIR}_i$, where

$$\mathsf{InputSIR}_i = 10\log_{10}\frac{\langle|\mathrm{x}_{J'i}(t)|^2\rangle_t}{\langle|\sum_{k\neq i}\mathrm{x}_{J'k}(t)|^2\rangle_t} \quad \text{(dB)}, \quad (11)$$

$$\mathsf{OutputSIR}_i = 10\log_{10}\frac{\langle|\mathrm{y}_{ii}(t)|^2\rangle_t}{\langle|\sum_{k\neq i}\mathrm{y}_{ik}(t)|^2\rangle_t} \quad \text{(dB)}, \quad (12)$$

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:

$$\mathsf{SDR}_i = 10\log_{10}\frac{\langle|\mathrm{x}_{J'i}(t)|^2\rangle_t}{\langle|\mathrm{x}_{J'i}(t) - \alpha\mathrm{y}_{ii}(t-D)|^2\rangle_t} \quad \text{(dB)}, \quad (13)$$

where α and D are parameters used to compensate for the amplitude and phase difference between $\mathrm{x}_{J'i}$ and y_{ii}.
To evaluate the musical noise, we also conducted a subjective test and obtained the mean opinion score (MOS). The listening tests were undertaken by 10 listeners. Each listener awarded a score from one (musical noise is clearly audible) to five (not audible) for each output signal.

3.3. Results

Figure 3 shows an example clustering result for normalized observation vectors when $N = 3$ and $M = 2$ (Fig. 2) 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. As expected, the SIR improvement was almost the same (around 12 dB on average) with both our proposed method and the previous DOAs-based method, because we have only two elements and all sources were at equal distances from the microphones.

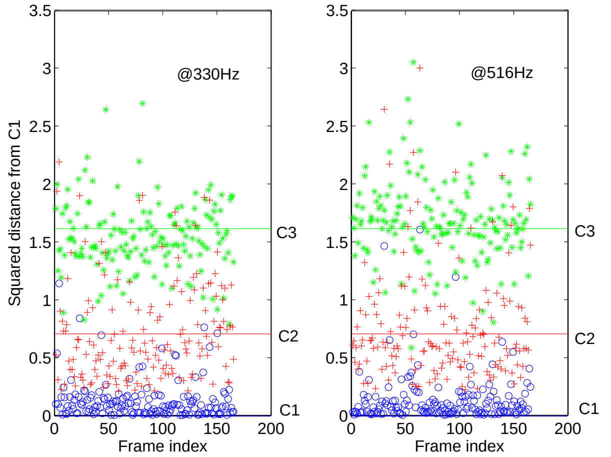

Figure 3: Example clustering result ($N = 3, M = 2$). o, +, * show the cluster members C_1, C_2 and C_3, respectively.

Table 1 shows the separation result for four sources with three sensors ($N = 4, M = 3$, underdetermined), that were arranged non-linearly (Fig. 4 [Setup 1]). We investigated four combinations of speakers and averaged the results. From Table 1, we can see that our proposed method achieved good separation performance even if we utilized the non-linear sensor arrangement. In such a non-linear case, the conventional DOA-based method cannot be applied straightforwardly. The applicability to a non-linear sensor array is one of the advantages of our method.

Table 1 also shows the SIR, SDR and MOS values when we changed the frame shift from $256 (= L/2)$ to $64 (= L/8)$. By using the fine-shift ($L/4$ and $L/8$), the SDR and MOS values increase without any reduction in the SIR values. The MOS was significantly different for each shift rate (the significant level was .01). Our subjective test confirmed that the fine-shift reduced the audible musical noise. This is because the fine-shift in (2) and the overlap-add in (10) realize a gradual change in the spectrogram of the separated signal [7]. We can say that the fine-shift effectively reduces the signal distortion when we employ it for our new method.

We also applied our method to a 3-dimensional sensor arrangement (Fig. 4 [Setup 2]). Table 2 shows the example separation result for four sources (two male and two female speakers) with four sensors ($N = M = 4$). Here, the system knew just the maximum distance (5.0 cm) between the reference microphone (Mic. 3) and the others. We can see from Table 2 that our proposed method can be applied to such a 3-dimensional microphone array system. Here, although there was no significant difference between the MOS results for shift=$L/4$ and shift=$L/8$, the fine-shift (shift=$L/4$ or $L/8$) reduced the musical noise.

Some sound examples can be found at [9].

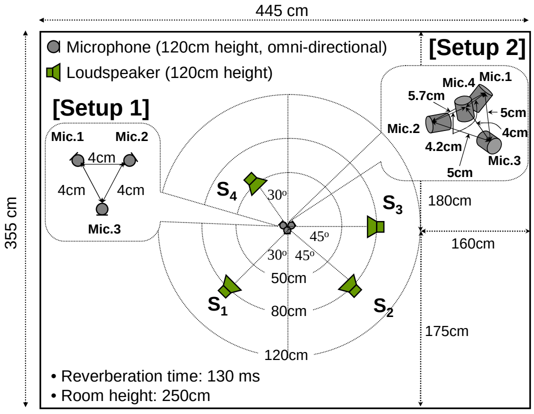

Figure 4: Experimental setups with non-linear arrays

4. CONCLUSION

We proposed a new method for separating sparse signals by clustering the normalized observation vectors. Our proposed technique 1) provides a novel BSS method that can be applied to both (over-)determined and underdetermined cases, 2) provides a new feature vector for clustering the observation vectors, 3) makes it easy to use a non-linear/non-uniform sensor arrangement, and 4) makes it possible to exploit the information obtained from all the sensors for separation.

Appendix

This appendix explains why normalized observation vectors $\bar{\mathbf{x}}(f,\tau)$ form a cluster for a source. Let us approximate the multi-path mixing model (1) by using a direct-path (nearfield) model (Fig. 5)

$$h_{jk}(f) \approx \frac{q(f)}{d_{jk}} \exp\left[\jmath\, 2\pi f c^{-1}(d_{jk} - d_{Jk})\right], \qquad (14)$$

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 normalized with the distance 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 (14) and (4) into (6) and (7) yields

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

which is independent of frequency, and dependent only on the positions of the sources and sensors. That is, the observation vectors are clustered based on the source geometry. From the fact that $\max_{j,k} |d_{jk} - d_{Jk}| \leq d_{\max}$, an inequality

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

holds. This property is important for the distance measure (8), since $|\bar{x} - \bar{x}'|$ increases monotonically as $|\arg(\bar{x}) - \arg(\bar{x}')|$ increases.

Table 1: Average SIR improvement [dB], average SDR [dB] and average MOS ($N = 4, M = 3$),

		y_1	y_2	y_3	y_4	MOS
InputSIR$_i$		-7.4	-6.2	-6.0	-1.0	-
Shift $L/2$	SIR$_i$	16.4	10.5	14.2	11.1	1.9
	SDR$_i$	4.7	3.9	4.8	7.9	
Shift $L/4$	SIR$_i$	17.4	11.6	15.5	12.0	2.6
	SDR$_i$	5.5	4.7	5.6	8.7	
Shift $L/8$	SIR$_i$	17.9	11.8	15.9	12.3	2.9
	SDR$_i$	5.6	4.8	5.8	8.8	

Table 2: Example SIR improvement [dB], SDR [dB] and average MOS ($N = 4, M = 4$),

		y_1	y_2	y_3	y_4	MOS
InputSIR$_i$		-8.1	-5.3	-6.6	-0.6	-
Shift $L/2$	SIR$_i$	17.8	15.6	9.6	15.7	1.9
	SDR$_i$	4.2	5.7	3.6	11.5	
Shift $L/4$	SIR$_i$	19.0	16.5	9.9	17.0	2.7
	SDR$_i$	4.7	6.1	4.1	12.1	
Shift $L/8$	SIR$_i$	19.3	17.3	10.1	17.4	2.7
	SDR$_i$	4.9	6.6	4.2	12.2	

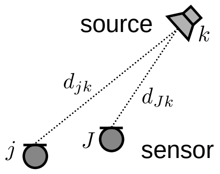

Figure 5: Direct-path (nearfield) model

5. 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] F. J. Theis, C. G. Puntonet, and E. W. Lang, "A histogram-based overcomplete ICA algorithm," in *Proc. ICA2003*, 2003, pp. 1071–1076.

[4] 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.

[5] 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. ICASSP2004*, May 2004, vol. III, pp. 881–884.

[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*, Mar. 2005, vol. III, pp. 61–64.

[7] 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*, Mar. 2005, vol. III, pp. 81–84.

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

[9] http://www.kecl.ntt.co.jp/icl/signal/araki/xcluster.html