Underdetermined Sparse Source Separation of Convolutive Mixtures 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
‡ Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo-shi, Hokkaido 060-0814, Japan
{shoko,sawada,ryo,maki}@cslab.kecl.ntt.co.jp

***Abstract*— We propose a new method for solving the underdetermined sparse signal separation problem. Some sparseness based methods have already been proposed. However, most of these methods utilized a linear sensor array (or only two sensors), and therefore they have certain limitations; e.g., they cannot separate symmetrically positioned sources. To allow the use of more than three 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. Our proposed method can handle both underdetermined case and (over-)determined cases. We show practical results for speech separation with non-linear/non-uniform sensor arrangements. We obtained promising experimental results for the cases of 3×4, 4×5 (#sensors $\times$ #sources) in a room (RT$_{60}$= 120 ms).**

I. Introduction

In this paper, we consider the blind source separation (BSS) in a real environment, i.e., the BSS of convolutive mixtures. In particular, we deal with the underdetermined BSS where we have less sensors M than sources N ($M < N$). Recently, independent component analysis (ICA) [1] has been widely studied for the convolutive BSS problem. However, ICA cannot be applied when $M < N$. In contrast, we propose a method that can handle both (over-)determined ($M \geq N$) and underdetermined ($M < N$) cases.

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

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

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

There are several approaches [2–6] 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, this was done by using geometric information (e.g., direction of arrival (DOA) and/or distance) about the sources, which is estimated from the phase and/or level difference between two observations of a linear sensor array. Some authors used the level difference between two observations (e.g., [3]), some employed both the level difference and phase difference of two sensors (e.g., [2]). Moreover, [5] used the direction of arrival (DOA) derived from the phase difference in order to normalize its frequency dependence, which causes a permutation problem in different frequency bins [6].

With previously reported approaches, which use two observations (or a linear sensor array), we can solve the underdetermined BSS problem in some cases. However, a linear array has certain limitations; for example, it limits the separation ability on a 2-dimensional half-plane, and offers no possibility of utilizing a source elevation information. In addition, the previous DOA approach needed exact sensor positions and sensor calibration to estimate the geometric features accurately. If we can use more than two sensors arranged freely, we can overcome such limitations.

In this paper, we propose a new method for separating sparse signals that overcomes the above-mentioned limitations of previous methods. Separation is achieved by clustering the normalized observation vectors. Previously, we have applied these normalization and clustering techniques to the basis vectors produced by ICA [7] 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.

With our method, 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 do not need to know the exact sensor locations, simply the maximum distance between a given sensor and any other sensors. This relaxation makes it easy to use a non-uniform sensor arrangement, and also eliminates the need for sensor calibration. We show the experimental results obtained in a room (reverberation time of 120 ms) with non-linear sensor arrays.

II. Proposed approach

A. *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):

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

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

0-7803-9390-2/06/$20.00 ©2006 IEEE

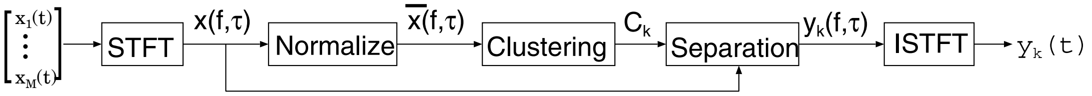

Fig. 1. Flow of proposed method

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:

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

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.

B. Separation procedures

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

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

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.

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 keeps the level differences at all sensors and all the phase differences with respect to the sensor J. In that sense, we can say that our proposed method generalizes the previous method [2, 6] for a multiple sensor case. As shown in the Appendix, $\bar{\mathbf{x}}(f,\tau)$ becomes dependent only on the positions of the sources and sensors. That is, the observation vectors are clustered based on the source geometry.

2) Clustering: The next step is to find clusters $C_1,\ldots,C_N$ formed by all 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].

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):

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

III. Experiments

A. 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 $64(=L/8)$.

B. 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 $\text{OutputSIR}_i - \text{InputSIR}_i$, where

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

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

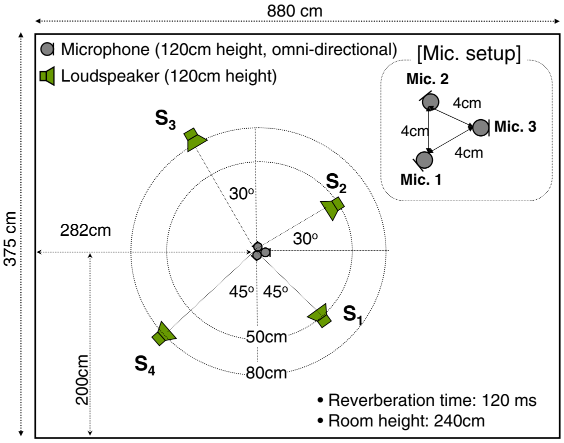

Fig. 2. Experimental setup with a non-linear array

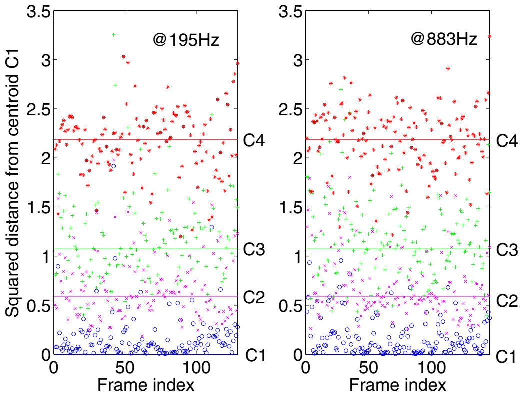

Fig. 3. Example clustering result ($M = 3, N = 4$). o, x, +, * show the cluster members C_1, C_2, C_3 and C_4, respectively.

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:

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

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.

C. 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 I shows the separation result. From Table I, we can see that our proposed method achieved good separation even if we utilized a non-linear sensor arrangement. Table I 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 [9].

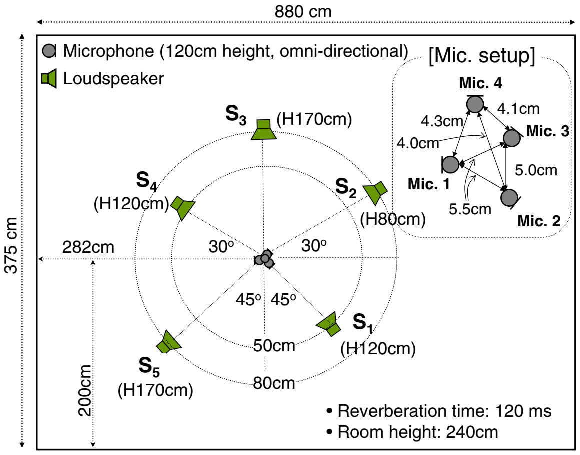

Fig. 4. Experimental setup with a 3-D array

TABLE I
EXPERIMENTAL RESULTS FOR $M = 3, N = 4$,

		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

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 II shows the separation results. We can see from Table II 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 (9). The results of subjective tests can be found in [10]. Some sound examples can be found at [11].

D. Discussion

In this section, we discuss the advantages of our method compared with some previous methods [2–6].

The first advantage of the proposed method is that we can utilize a 2- or 3-dimensionally arranged sensor arrangement. A previously adopted linear sensor array in [2–5] 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. Let us show an example. Figure 5 is an example histogram of the estimated DOA when we use Mic. 1 and Mic. 2 in Fig. 2 for S_1, S_2 and

TABLE II

EXPERIMENTAL RESULTS FOR $M = 4, N = 5$,

		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

S_4. Although there were three sources, we can see only two peaks in the histogram. This is because S_1 and S_4 came from $\pm 45°$, and therefore, they could not be distinguished by Mic. 1 and Mic. 2. On the other hand, because our proposed method makes it easy to employ a two-dimensional non-linear sensor arrangement, we can cope with source arrangements such as that shown in Fig. 2.

The second advantage is that the proposed normalization of the observation vector allows us to cluster all the frequency components together. Previously, [6] has utilized the method for more than two sensors case, however, they still worked in individual frequency bins. Therefore the (inner-)permutation problem remains and a permutation error decreases the separation performance. By contrast, our frequency normalized observation vector does not have this problem inherently.

The third advantage is that, unlike the previous DOA approach [5], we do not need to know the exact sensor locations. We simply need the maximum distance $d_{\max}$ between a given sensor and any other sensor. If 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.

IV. CONCLUSION

We proposed a new method for underdetermined BSS by clustering the normalized observation vectors. Our proposed technique makes it easy to use a non-linear/non-uniform sensor arrangement, and makes it possible to exploit information obtained from all the sensors for separation. In this paper we provided the results solely for underdetermined cases, however, our proposed method can also be applied to (over-)determined cases [10].

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 (near-field) model (Fig. 6)

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

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

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

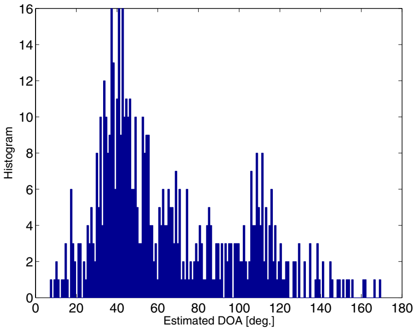

Fig. 5. Example histogram of DOAs for S_1, S_2 and S_4 with sensors 1 and 2 in Fig.2 condition.

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 $\max_{j,k} |d_{jk} - d_{Jk}| \le d_{\max}$, an inequality

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

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.

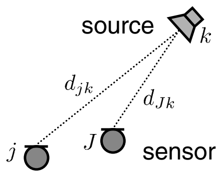

Fig. 6. Direct-path (near-field) model

REFERENCES

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

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

[3] F. Theis, E. Lang, and C. Puntonet, "A geometric algorithm for overcomplete linear ICA," *Neurocomputing*, vol. 56, pp. 381–398, 2004.

[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. ICASSP 2004*, vol. III, May 2004, pp. 881–884.

[6] 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.

[7] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of a dominant source from mixtures of many sources using ICA and time-frequency masking,"," in *Proc. ISCAS 2005*, May 2005, pp. 5882–5885.

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

[9] 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.

[10] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. IWAENC 2005*, Sept. 2005, pp. 117–120.

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