BLIND SOURCE SEPARATION OF 3-D LOCATED MANY SPEECH SIGNALS

Ryo Mukai *Hiroshi Sawada* *Shoko Araki* *Shoji Makino*
NTT Communication Science Laboratories, NTT Corporation
2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619–0237, Japan
ryo@cslab.kecl.ntt.co.jp
http://www.kecl.ntt.co.jp/icl/signal/mukai/

ABSTRACT

This paper presents a prototype system for Blind Source Separation (BSS) of many speech signals and describes the techniques used in the system. Our system uses 8 microphones located at the vertexes of a 4cm×4cm×4cm cube and has the ability to separate signals distributed in three-dimensional space. The mixed signals observed by the microphone array are processed by Independent Component Analysis (ICA) in the frequency domain and separated into a given number of signals (up to 8). We carried out experiments in an ordinary office and obtained more than 20 dB of SIR improvement.

1. INTRODUCTION

The Blind Source Separation (BSS) [1] of audio signals has a wide range of applications. In most realistic applications, the number of source signals is large, and the signals are mixed in a convolutive manner with reverberations. Independent component analysis (ICA) [2] is one of the main statistical methods used for BSS. It is theoretically possible to solve the BSS problem with a large number of sources by ICA, if we assume that the number of sensors is equal to or greater than the number of source signals. However, there are many practical difficulties. Although many studies have been undertaken on BSS in a reverberant environment [3], most of them have assumed two source signals, and only a few studies have dealt with more than two source signals.

In this paper, we present techniques for the BSS of many speech signals distributed in three-dimensional space and a prototype system that we have developed. In our previous work [4], we described the separation of six source signals consisting of simulated data, i.e. signals made by convolving impulse responses. In contrast, this prototype system performs an on-the-spot BSS of live recorded signals. This paper is an extended version of our previous work [5].

2. FREQUENCY DOMAIN BLIND SOURCE SEPARATION

There are two major approaches to solving the convolutive BSS problem. The first is the time domain approach, where ICA is applied directly to the convolutive mixture model [6, 7, 8]. The time domain approach incurs considerable computational cost, and it is difficult to obtain a solution in a practical time when the number of source signals is large.

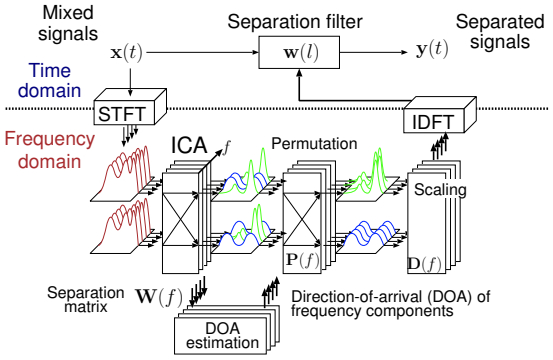

Figure 1: Flow of frequency domain BSS

The other approach is frequency domain BSS, where ICA is applied to multiple instantaneous mixtures in the frequency domain [9, 10, 11]. This approach takes much less computation time than time domain BSS.

2.1. ICA in frequency domain

When N source signals are $\mathrm{s}_1(t), ..., \mathrm{s}_N(t)$ and the signals observed by M sensors are $\mathrm{x}_1(t), ..., \mathrm{x}_M(t)$, the mixing model can be described by

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

where $\mathrm{h}_{ji}(l)$ is the impulse response from source i to sensor j. The separation system typically consists of a set of FIR filters $\mathrm{w}_{kj}(l)$ of length L designed to produce N separated signals $\mathrm{y}_1(t), ..., \mathrm{y}_N(t)$, and it is described as:

$$\mathrm{y}_k(t) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} \mathrm{w}_{kj}(l)\mathrm{x}_j(t-l). \quad (2)$$

Figure 1 shows the flow of BSS in the frequency domain. Each convolutive mixture in the time domain is converted into multiple instantaneous mixtures in the frequency domain. By using a short-time discrete Fourier transform (DFT), the mixing model is approximated as:

$$\mathbf{x}(f,m) = \mathbf{H}(f)\mathbf{s}(f,m), \quad (3)$$

where f denotes the frequency, m is the frame index, $\mathbf{s}(f,m) = [s_1(f,m), ..., s_N(f,m)]^T$ is the vector of the source signals in the frequency bin f, $\mathbf{x}(f,m) = [x_1(f,m), ..., x_M(f,m)]^T$ is the vector of the observed signals, and $\mathbf{H}(f)$ is a matrix consisting of the frequency re-

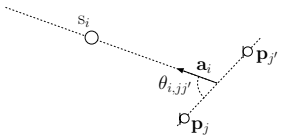

Figure 2: Direction of source i relative to sensor pair j and j'

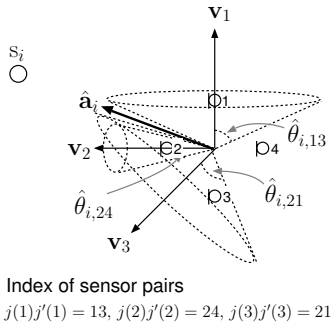

Figure 3: Solving ambiguity of estimated DOAs

sponses $H_{ji}(f)$ from source i to sensor j. The separation process can be formulated in each frequency bin as:

$$\mathbf{y}(f,m) = \mathbf{W}(f)\mathbf{x}(f,m), \tag{4}$$

where $\mathbf{y}(f,m) = [y_1(f,m), ..., y_N(f,m)]^T$ is the vector of the separated signals, and $\mathbf{W}(f)$ represents the separation matrix. Therefore, we can apply an ordinary (instantaneous) ICA algorithm to each frequency bin and calculate the separation matrices. $\mathbf{W}(f)$ is determined so that the elements of $\mathbf{y}(f,m)$ become mutually independent for each f.

The ICA solution suffers from scaling and permutation ambiguities. This is because that if $\mathbf{W}(f)$ is a solution, then $\mathbf{D}(f)\mathbf{P}(f)\mathbf{W}(f)$ is also a solution, where $\mathbf{D}(f)$ is a diagonal complex valued scaling matrix, and $\mathbf{P}(f)$ is an arbitrary permutation matrix. There is a simple and reasonable solution for the scaling problem:

$$\mathbf{D}(f) = \mathrm{diag}\{[\mathbf{P}(f)\mathbf{W}(f)]^{-1}\}, \tag{5}$$

which is obtained by the minimal distortion principle (MDP) [12] or the projection back method [13], and we can use it. On the other hand, the permutation problem is complicated, especially when the number of source signals is large. Before constructing a separation filter in the time domain, we have to align the permutation so that each channel contains frequency components from one source signal. The time domain filters are obtained by the inverse discrete Fourier transform of frequency domain separation matrices.

2.2. DOA estimation using ICA solution

The frequency response matrix $\mathbf{H}(f)$ is closely related to the locations of the sources and sensors. If a separation matrix $\mathbf{W}(f)$ is calculated successfully and it extracts source signals with a scaling ambiguity, there is a diagonal matrix $\mathbf{D}(f)$, and $\mathbf{D}(f)\mathbf{W}(f)\mathbf{H}(f) = \mathbf{I}$ holds. Because of the scaling ambiguity, we cannot obtain $\mathbf{H}(f)$ simply from the ICA solution $\mathbf{W}(f)$. However, the ratio of elements in the same column $H_{ji}(f)/H_{j'i}(f)$ is invariable in relation to $\mathbf{D}(f)$, and is given by

$$\frac{H_{ji}(f)}{H_{j'i}(f)} = \frac{[\mathbf{W}^{-1}(f)\mathbf{D}^{-1}(f)]_{ji}}{[\mathbf{W}^{-1}(f)\mathbf{D}^{-1}(f)]_{j'i}} = \frac{[\mathbf{W}^{-1}(f)]_{ji}}{[\mathbf{W}^{-1}(f)]_{j'i}}, \tag{6}$$

where $[\cdot]_{ji}$ denotes the ji-th element of the matrix.

We can estimate the DOA of a source signal by using this invariant [4]. With a far-field model, a frequency response is formulated as:

$$H_{ji}(f) = e^{\jmath 2\pi f c^{-1}\mathbf{a}_i^T\mathbf{p}_j}, \tag{7}$$

where c is the wave propagation speed, $\mathbf{a}_i$ is a unit vector that points to the direction of source i (*absolute* DOA), and $\mathbf{p}_j$ represents the location of sensor j. According to this model, we have

$$H_{ji}(f)/H_{j'i}(f) = e^{\jmath 2\pi f c^{-1}\mathbf{a}_i^T(\mathbf{p}_j - \mathbf{p}_{j'})} \tag{8}$$

$$= e^{\jmath 2\pi f c^{-1}\|\mathbf{p}_j - \mathbf{p}_{j'}\|\cos\theta_{i,jj'}(f)}, \tag{9}$$

where $\theta_{i,jj'}(f)$ is the direction of source i relative to the sensor pair j and j' (*relative DOA*). Figure 2 shows the relation of the absolute DOA and the relative DOA. By using the argument of (9) and (6), we can estimate:

$$\hat{\theta}_{i,jj'}(f) = \arccos\frac{\arg(H_{ji}/H_{j'i})}{2\pi f c^{-1}\|(\mathbf{p}_j - \mathbf{p}_{j'})\|}$$

$$= \arccos\frac{\arg([\mathbf{W}^{-1}]_{ji}/[\mathbf{W}^{-1}]_{j'i})}{2\pi f c^{-1}\|(\mathbf{p}_j - \mathbf{p}_{j'})\|}. \tag{10}$$

$\hat{\theta}_{i,jj'}(f)$ is estimated for each frequency bin f, but we omit the argument f to simplify the notation in the following description.

The DOA estimation involves certain ambiguities. When we use only one pair of sensors or a linear array, the estimated $\hat{\theta}_{i,jj'}$ determines a cone rather than a direction. This ambiguity can be solved by using multiple sensor pairs (Fig. 3). If we use sensor pairs that have different axis directions, we can estimate cones with various vertex angles for one source direction. If the relative DOA $\hat{\theta}_{i,jj'}$ is estimated without any error, the absolute DOA $\mathbf{a}_i$ satisfies:

$$\frac{(\mathbf{p}_j - \mathbf{p}_{j'})^T\mathbf{a}_i}{\|\mathbf{p}_j - \mathbf{p}_{j'}\|} = \cos\hat{\theta}_{i,jj'}. \tag{11}$$

When we use L sensor pairs whose indexes are $j(l)j'(l)(1 \le l \le L)$, $\mathbf{a}_i$ is given by the solution of the following equation:

$$\mathbf{V}\mathbf{a}_i = \mathbf{c}_i, \tag{12}$$

where $\mathbf{V} \triangleq (\mathbf{v}_1, ..., \mathbf{v}_L)^T$, $\mathbf{v}_l \triangleq \frac{\mathbf{p}_{j(l)} - \mathbf{p}_{j'(l)}}{\|\mathbf{p}_{j(l)} - \mathbf{p}_{j'(l)}\|}$ is a normalized axis, and $\mathbf{c}_i \triangleq [\cos(\hat{\theta}_{i,j(1)j'(1)}), ..., \cos(\hat{\theta}_{i,j(L)j'(L)})]^T$. Sensor pairs should be selected so that $\mathrm{rank}(\mathbf{V}) \ge 3$ if the potential source locations are three-dimensional.

In a practical situation, $\hat{\theta}_{i,j(l)j'(l)}$ has an estimation error, and (12) has no exact solution. Thus we adopt an optimal solution by employing certain criteria such as:

$$\hat{\mathbf{a}}_i = \underset{\mathbf{a}}{\mathrm{argmin}}\ ||\mathbf{V}\mathbf{a} - \mathbf{c}_i||\ \ (\text{subject to}\ \ ||\mathbf{a}|| = 1) \tag{13}$$

This can be solved approximately by using the Moore-

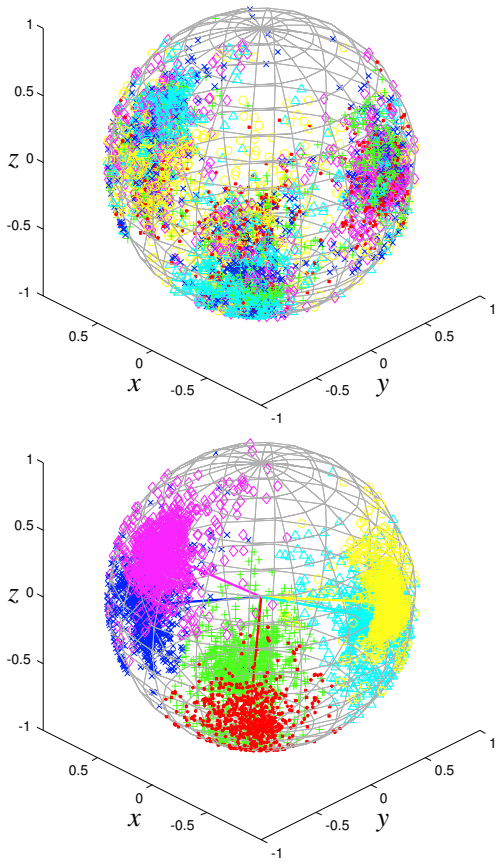

Figure 4: Estimated DOAs of frequency components (above) and clustered result (below)

Penrose pseudo-inverse $\mathbf{V}^+ \triangleq (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T$, and we have:

$$\hat{\mathbf{a}}_i \approx \frac{\mathbf{V}^+\mathbf{c}_i}{||\mathbf{V}^+\mathbf{c}_i||}. \quad (14)$$

Accordingly, we can determine a unit vector $\hat{\mathbf{a}}_i$ pointing to the direction of source s_i.

Figure 4 shows an example of a DOA estimation result. Each point plotted on a unit sphere denotes the estimated DOA of a frequency component in one frequency bin. The points can be clustered by using an ordinary clustering method such as the k-means algorithm [14], then the DOAs of source signals are given as the centroids of the clusters. This information is useful for solving the permutation problem.

2.3. Permutation solver using DOA and correlation

This subsection outlines the procedure for permutation alignment by integrating a DOA based approach and a correlation approach. This procedure has been detailed in [11], and consists of the following steps:

1. Cluster separated frequency components $y_k(f,m)$ for all k and all f by using the estimated DOA, and decide the permutations at certain frequencies where the confidence of DOA estimation is sufficiently high.
2. Decide the permutations to maximize the sum of the inter-frequency correlation of the separated signals. The correlation should be calculated for the amplitude $|y_k(f,m)|$ or (log-scaled) power $|y_k(f,m)|^2$ instead of the raw complex-valued signals $y_k(f,m)$, since the correlation of raw signals would be very low because of the short-time DFT property. The sum of the correlations between $|y_k(f,m)|$ and $|y_k(g,m)|$ within distance δ (i.e. $|f-g| < \delta$) is used as a criterion. The permutations are decided for frequencies where the criterion gives a clear-cut decision.
3. Calculate the correlations between $|y_k(f,m)|$ and its harmonics $|y_k(g,m)|$ ($g = 2f, 3f, 4f, ...$), and decide the permutations that maximize the sum of the correlations. The permutations are decided for frequencies where the correlation among harmonics is sufficiently high.
4. Decide the permutations for the remaining frequencies based on neighboring correlations.

The DOA estimation suffers from errors in a reverberant environment and the classification according to the DOA is inconsistent in some frequency bins. The correlation based method is not robust since a misalignment at one frequency bin causes consecutive misalignments. The main advantage of the integrated method is that it does not cause a large misalignment as long as the permutations fixed by the DOA based approach are correct. Moreover, the correlation part (steps 2, 3 and 4) compensates for the lack of preciseness of the DOA based approach. The correlation part consists of three steps for two reasons. First, the harmonics part (step 3) works well if most of the other permutations are fixed. Second, the method becomes more robust by quitting the step 2 if there is no clear-cut decision. With this structure, we can avoid fixing the permutations for consecutive frequencies without high confidence. This integrated method is effective when the number of source signals is large.

2.4. Spectral smoothing with error minimization

Frequency domain BSS is influenced by the circularity of the discrete frequency representation. This causes a problem when we convert separation matrices in the frequency domain into a separation filter in the time domain. This problem is not apparent when there are two sources, however it is crucial when the number of source signals exceeds two. Our technique for solving this problem involves spectral smoothing of separation filters by using a window that tapers smoothly to zero at each end. The direct application of windowing changes the frequency responses for separation obtained by ICA and causes an error. Therefore, we adjust the frequency responses before windowing so that the error is minimized. The procedure is presented in detail in [15].

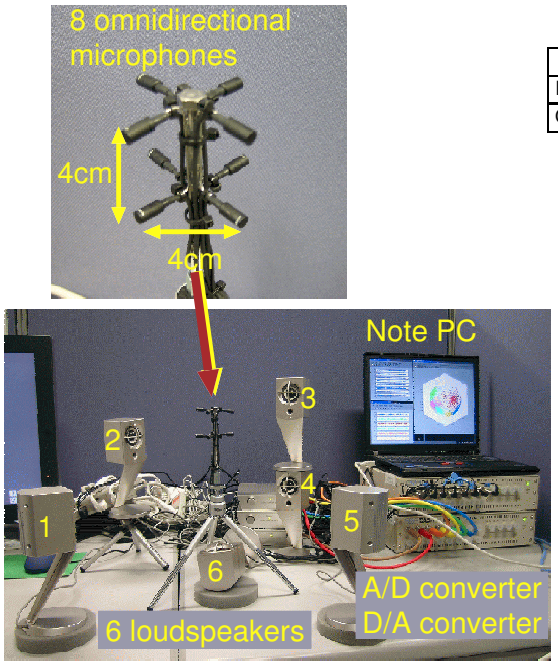

Figure 5: Prototype system and experimental settings

Table 1: Specifications of prototype system

Microphone	8 omni-directional microphones
Sampling rate	8 kHz
Frame length	2048 points (256 ms)
Frame shift	512 points (64 ms)
ICA algorithm	FastICA + Infomax (complex valued)
CPU	Intel Pentium M (2.0 GHz)
Coding	MATLAB + C
Computation time	25 s for 6 sources 8 s data

3. PROTOTYPE SYSTEM AND EXPERIMENTS

We have developed a prototype using the techniques described above. Our system uses 8 microphones located at the vertexes of a 4cm×4cm×4cm cube and has the ability to separate six signals distributed in three-dimensional space (Fig. 5). The system specifications are summarized in Table 1. This system is implemented in software (MATLAB + C) and needs no special hardware except for an A/D converter. We calculated $\mathbf{W}$ by using a complex-valued version of FastICA [16] and improved it further by using InfoMax [17] combined with the natural gradient whose nonlinear function is based on the polar coordinate [18].

We carried out experiments in an ordinary office and evaluated the Signal to Interference Ratio (SIR) performance. The source locations are shown in Fig. 5. We calculated the separation filter by using live recorded mixtures, and evaluated the SIRs by using individually activated source signals. The experimental results are shown in Table 2. We obtained good separation performance in spite of the very low input SIR. The average SIR improvement was more than 20 dB.

Table 2: Experimental results (dB)

	SIR_1	SIR_2	SIR_3	SIR_4	SIR_5	SIR_6	ave.
Input SIR	−11.6	−9.0	−9.0	−6.6	−6.9	−2.5	−7.6
Output SIR	7.6	12.2	16.4	14.4	13.6	13.7	13.0

4. CONCLUSION

We have developed a prototype system for the BSS of many speech signals distributed in three-dimensional space. Our experimental result in an ordinary office showed good separation performance. Some sound examples can be found on our web site [19].

5. REFERENCES

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering*. John Wiley & Sons, 2000.

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

[3] C. G. Puntonet and A. Prieto, Eds., *Independent Component Analysis and Blind Signal Separation (LNCS 3195)*. Springer-Verlag, 2004.

[4] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation for many speech signals," in *Proc. ICA2004 (LNCS 3195)*. Springer-Verlag, 2004, pp. 461–469.

[5] ——, "Blind source separation and DOA estimation using small 3-D microphone array," in *Proc. HSCMA 2005*, 2005, pp. d.9–10.

[6] S. C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Communication*, vol. 39, pp. 65–78, 2003.

[7] K. Matsuoka, Y. Ohba, Y. Toyota, and S. Nakashima, "Blind separation for convolutive mixture of many voices," in *Proc. IWAENC 2003*, 2003, pp. 279–282.

[8] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

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

[10] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 6, pp. 352–362, Sept. 2002.

[11] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[12] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Intl. Workshop on Independent Component Analysis and Blind Signal Separation (ICA'01)*, 2001, pp. 722–727.

[13] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

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

[15] H. Sawada, R. Mukai, S. de la Kethulle, S. Araki, and S. Makino, "Spectral smoothing for frequency-domain blind source separation," in *Proc. IWAENC 2003*, 2003, pp. 311–314.

[16] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *International Journal of Neural Systems*, vol. 10, no. 1, pp. 1–8, Feb. 2000.

[17] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[18] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590–596, Mar. 2003.

[19] http://www.kecl.ntt.co.jp/icl/signal/mukai/demo/waspaa2005/