REAL-TIME BLIND SOURCE SEPARATION AND DOA ESTIMATION USING SMALL 3-D MICROPHONE ARRAY

Ryo Mukai *Hiroshi Sawada* *Shoko Araki* *Shoji Makino*

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

ABSTRACT

We present a prototype system for real-time blind source separation (BSS) and directions of arrival (DOA) estimation. Our system uses a small three-dimensional array with 8 microphones 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. The system estimates DOA of the source signals as a by-product of the separation process. In our previous work [1], we presented a batch-type BSS of recorded signals. In contrast, this prototype system performs a real-time separation.

1. INTRODUCTION

Blind source separation (BSS) is a technique for estimating original source signals using only observed mixtures. The BSS of audio signals has a wide range of applications including speech enhancement [2] for speech recognition, hands-free telecommunication systems and high-quality hearing aids. In most realistic applications, the sources are located in three-dimensional space and the locations may change. In this contribution, we present our prototype system for real-time BSS of three-dimensionally located signals (Fig. 1) and describe the techniques used in the system.

2. FREQUENCY DOMAIN BSS

Independent component analysis (ICA) [3] is one of the main statistical methods used for BSS. In a reverberant environment, the signals are mixed in a convolutive manner with reverberations, and the separation system is a matrix of filters. 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 [4, 5, 6]. The other approach is frequency domain BSS, where ICA is applied to multiple instantaneous mixtures in the frequency domain [7, 8, 9]. The computation cost of the frequency domain approach is much less than that of the time domain approach. Our system employs frequency domain ICA using a blockwise batch algorithm [10].

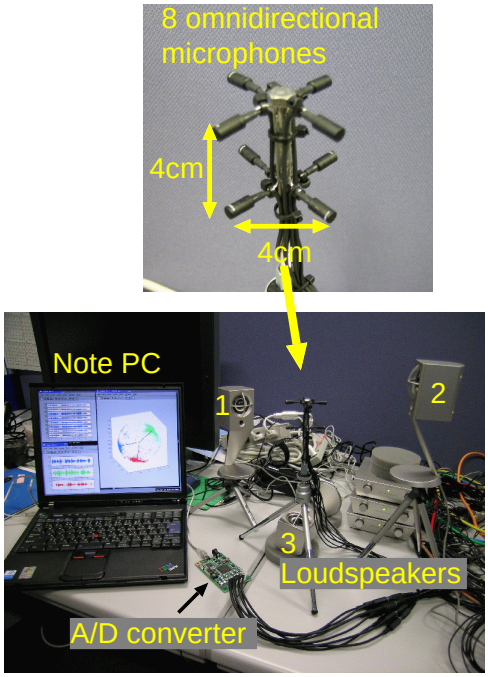


Figure 1: Prototype system for real-time BSS and DOA estimation of 3-D located signals

2.1. ICA in the 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

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

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:

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

Figure 2 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

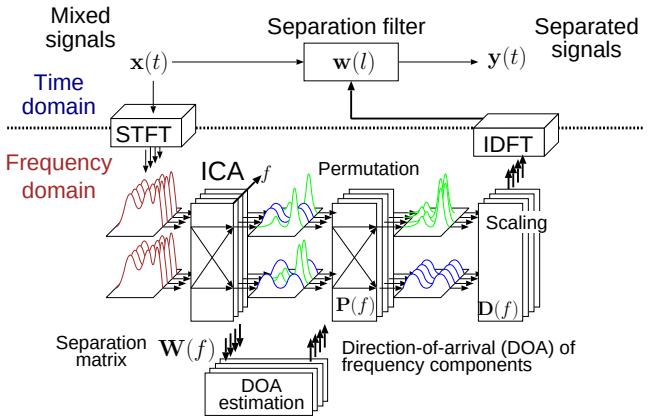


Figure 2: Flow of frequency domain BSS

(DFT), the mixing model is approximated as:

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

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 responses $H_{ji}(f)$ 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) [11] or the projection back method [12], and we can use it. On the other hand, the permutation problem is complicated. 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.

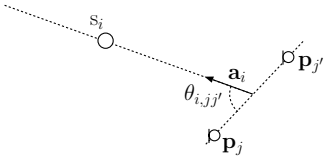
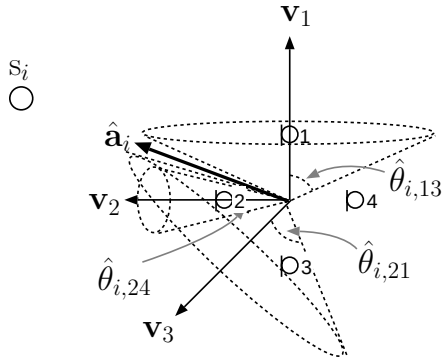


Figure 3: Direction of source i relative to sensor pair j and j'



Index of sensor pairs

$j(1)j'(1) = 13,\ j(2)j'(2) = 24,\ j(3)j'(3) = 21$

Figure 4: Solving ambiguity of estimated DOAs

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. With a far-field model, a frequency response is formulated as:

$$H_{ji}(f) = e^{j2\pi f c^{-1}\mathbf{a}_i^T\mathbf{p}_j}, \tag{7}$$

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

$$H_{ji}(f)/H_{j'i}(f) = e^{j2\pi f c^{-1}\mathbf{a}_i^T(\mathbf{p}_j - \mathbf{p}_{j'})} \tag{8}$$

$$= e^{j2\pi f c^{-1}\|\mathbf{p}_j - \mathbf{p}_{j'}\|\cos\theta_{i,jj'}(f)}, \tag{9}$$

where $\theta_{i,jj'}(f)$ is the direction of source i relative to the sensor pair j and j' (*relative DOA*). Figure 3 shows the relation of the absolute DOA and the relative DOA.

By using the argument of (9) and (6), we can estimate:

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

$\hat{\theta}_{i,jj'}(f)$ 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. 4). 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:

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

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:

$$\mathbf{V}\mathbf{a}_i = \mathbf{c}_i, \quad (12)$$

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:

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

This can be solved approximately by using the Moore-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 5 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 [13], 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 problem

The permutation problem is the most critical issue as regards frequency domain BSS. There are two major approaches for solving this problem: the DOA based method and the correlation based method. The estimated DOA is useful for solving the permutation problem, however the estimation suffers from errors in a reverberant environment and the classification according to the DOA is inconsistent in some frequency bins. Thus we employ the correlation based method for such frequency bins. The combination of these two methods provides a good solution. The procedure is presented in detail in [9]. Recently, we have also proposed another method to solve the permutation problem by using basis vector clustering, which is detailed in [14].

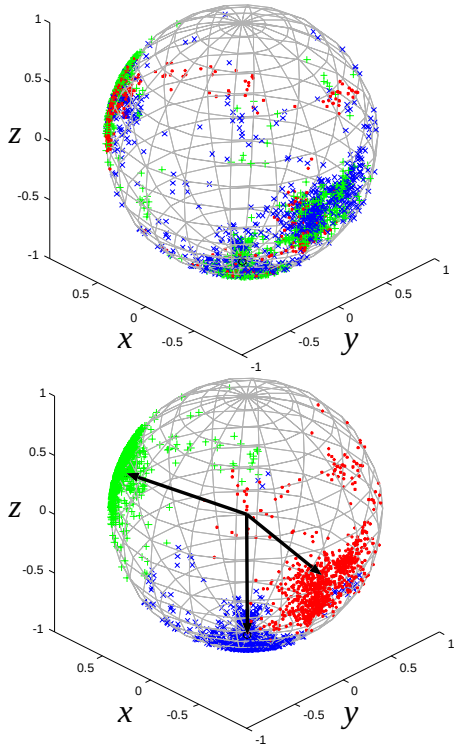


Figure 5: Estimated DOAs of frequency components (above) and clustered result (below)

3. PROTOTYPE SYSTEM

We have developed a prototype system for real-time BSS and DOA estimation. 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. This system is implemented in software (MATLAB + C) and needs no special hardware except for an A/D converter. We adopted a low-delay block-wise batch implementation [10], where ICA is applied for a block of a few seconds but the system input-output delay is kept as small as less than a second. We calculate $\mathbf{W}$ by using a complex-valued version of InfoMax [15] combined with the natural gradient whose nonlinear function is based on the polar coordinate [16].

Table 1: Experimental conditions

Microphone	8 omni-directional microphones
Sampling rate	8 kHz
Frame length	1024 points (128 ms)
Frame shift	256 points (32 ms)
Learning block size	3.2 s
Filter update interval	1.6 s
ICA algorithm	Infomax (complex valued)

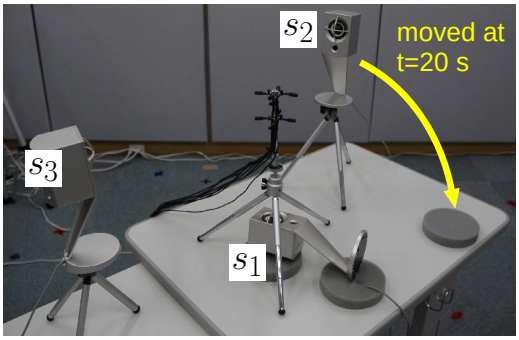


Figure 6: Source locations. s_2 was moved at $t = 20$ s.

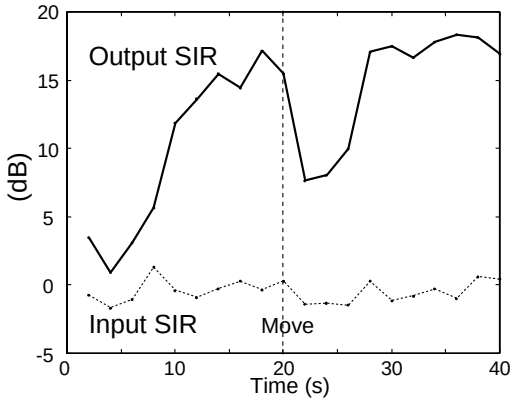


Figure 7: Experimental results

4. EXPERIMENTS

To examine the performance of our system, we carried out experiments using mixture of three speech signals recorded in a room. The layout of the microphone array and loudspeakers is shown in Fig. 6. In this experiment, we assumed that s_1 is a target signal and other two are interference signals. One of the interference signals, s_2, was moved at $t = 20$ s. Other conditions are summarized in Table 1. We calculated time-varying separating filters by using the mixtures of live recorded source signals, and evaluated Signal to Interference Ratio (SIR) by using individually activated source signals. The results are shown in Fig. 7. We can see that the separation performance declines when the interference signal moves and that it recovers after about 6 s.

5. CONCLUSION

We have developed a prototype system for the real-time BSS of speech signals distributed in three-dimensional space. The system estimates DOA of the source signals as a by-product of the separation process. Some sound examples can be found on our web site [17].

6. REFERENCES

[1] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation and DOA estimation using small 3-D microphone array," in *Proc. HSCMA 2005*, 2005, pp. d.9–10.

[2] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, 2005.

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

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Trans. Fundamentals*, vol. E87-A, no. 8, pp. 1941–1948, 2004.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] http://www.kecl.ntt.co.jp/icl/signal/mukai/demo/iwaenc2005/