On Calculating the Inverse of Separation Matrix in Frequency-Domain Blind Source Separation

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

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

Abstract. For blind source separation (BSS) of convolutive mixtures, the frequency-domain approach is efficient and practical, because the convolutive mixtures are modeled with instantaneous mixtures at each frequency bin and simple instantaneous independent component analysis (ICA) can be employed to separate the mixtures. However, the permutation and scaling ambiguities of ICA solutions need to be aligned to obtain proper time-domain separated signals. This paper discusses the idea that calculating the inverses of separation matrices obtained by ICA is very important as regards aligning these ambiguities. This paper also shows the relationship between the ICA-based method and the time-frequency masking method for BSS, which becomes clear by calculating the inverses.

1 Introduction

With acoustical applications of blind source separation (BSS), such as solving a cocktail party problem, signals are generally mixed in a convolutive manner with reverberations. Let $\mathrm{s}_1, \ldots, \mathrm{s}_N$ be N source signals and $\mathrm{x}_1, \ldots, \mathrm{x}_M$ be M sensor observations. Then, the convolutive mixture model is formulated as

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

where t represents time and $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j. If we consider sounds mixed in a practical room situation, impulse responses $\mathrm{h}_{jk}(l)$ tend to have hundreds or thousands of taps even with an 8 kHz sampling rate. This makes the convolutive BSS problem very difficult compared with the BSS of simple instantaneous mixtures.

A practical approach for such convolutive mixtures is frequency-domain BSS [1-10], where we apply a short-time Fourier transform (STFT) to the sensor observations $\mathrm{x}_j(t)$. Then, the convolutive model (1) can be approximated as an instantaneous mixture model at each frequency:

$$x_j(f,t) = \sum_{k=1}^{N} h_{jk}(f) s_k(f,t), \quad j=1, \ldots, M, \tag{2}$$

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 691–699, 2006.
© Springer-Verlag Berlin Heidelberg 2006

where f represents frequency, t is now down-sampled with the distance of the frame shift, $h_{jk}(f)$ is the frequency response from source k to sensor j, and $s_k(f,t)$ is a frequency-domain time-series signal of $\mathrm{s}_k(t)$ obtained with an STFT. The vector notation of (2) is

$$\mathbf{x}(f,t) = \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,t), \tag{3}$$

where $\mathbf{x} = [x_1, \ldots, x_M]^T$ and $\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$.

Once we assume instantaneous mixtures at each frequency, and also if the number of sources N does not exceed the number of sensors M, we can apply standard instantaneous independent component analysis (ICA) [11] to the mixtures $\mathbf{x}(f,t)$ to obtain separated frequency components:

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

where $\mathbf{y} = [y_1, \ldots, y_N]^T$ is the vector of separated frequency components and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]^H$ is an $N \times M$ separation matrix [1-6]. However, the ICA solution has the permutation and scaling ambiguities. We need to align these ambiguities to obtain proper time-domain separated signals.

Various studies have tried to solve these permutation and scaling problems because they constitute a critical issue. Some of these studies have attempted to solve the problems by using information obtained from (4), i.e the separation matrix $\mathbf{W}$ and/or the separated signals $\mathbf{y}$, or by imposing some constraints on $\mathbf{W}$. By contrast, we believe that the inverses of separation matrices $\mathbf{W}$ provide useful information for solving these problems. The main topic of this paper is to discuss the procedures for solving these problems by calculating the inverses.

There is also a frequency-domain BSS method that is based on time-frequency (T-F) masking [7-10]. It does not employ a standard ICA to separate the mixtures, and can be applied even if the number of sources N exceeds the number of sensors M. The method relies on the sparseness of source signals. It classifies the mixtures $\mathbf{x}(f,t)$ based on spatial information extracted from them. As the second topic of this paper, we show a link between the ICA-based method and the T-F masking method. The link becomes clear once we have the decomposition (6) of mixtures by calculating the inverse of $\mathbf{W}$. Based on the link, we see that some of the techniques used in solving the permutation problem can also be used for classifying the mixtures in the T-F masking method, and vice versa.

2 Calculating the Inverses of Separation Matrices

Figure 1 shows the flow of ICA-based frequency-domain BSS that we consider in this paper. The inverse of separation matrix $\mathbf{W}$ is represented as

$$[\mathbf{a}_1, \cdots, \mathbf{a}_N] = \mathbf{W}^{-1},\ \mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T, \tag{5}$$

which we call basis vectors obtained by ICA, because the mixture $\mathbf{x}(f,t)$ is represented by their linear combination by multiplying $\mathbf{W}^{-1}$ and (4):

$$\mathbf{x}(f,t) = \sum_{i=1}^{N} \mathbf{a}_i(f) y_i(f,t). \tag{6}$$

The basis vectors provide the key information with which to solve the permutation and scaling problems as shown in the following sections. If $\mathbf{W}$ is not square, we use the Moore-Penrose pseudoinverse instead of the inverse. It is not difficult to make $\mathbf{W}$ invertible by using an appropriate ICA procedure, such as whitening followed by unitary transformation (e.g. FastICA [11]).

3 Solving the Permutation Problem

Various methods have been proposed for solving the permutation problem:

1. making the separation matrices $\mathbf{W}(f)$ smooth along frequencies f [1, 2],
2. maximizing the correlation of separated signal envelopes $|y_i|$ [3],
3. analyzing the directivity patterns calculated from $\mathbf{W}(f)$ [4],
4. manipulating basis vectors $\mathbf{a}_i(f)$ [5, 6].

The third and fourth methods utilize the same information because $\mathbf{W}$ and $\mathbf{a}_i$ are related to each other through the inversion (5). However, the fourth method is easier to apply when there are more than two sources [5, 6], because basis vectors $\mathbf{a}_i$ directly represent estimated mixing system information (6). This section describes how to utilize this information for solving the permutation problem.

3.1 Assumption and Basic Idea

If ICA works well, we obtain separated signals $y_i(f,t)$ that should be close to source signals $s_k(f,t)$ up to the permutation and scaling ambiguities. If we compare (3) and (6), we see that the basis vectors $\mathbf{a}_i(f)$, which are obtained by ICA and the subsequent inversion of $\mathbf{W}$, should be close to the vectors $\mathbf{h}_k(f)$, again, up to the permutation and scaling ambiguities. The use of different subscripts, i and k, indicates the permutation ambiguity.

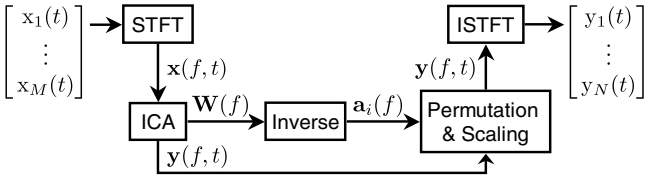

Fig. 1. Flow of ICA-based frequency-domain BSS

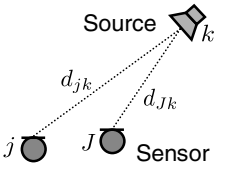

Fig. 2. Direct-path model

The method presented here assumes a direct-path model (Fig. 2) for the vector $\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$, even though in reality signals are mixed in a multi-path model (1). This simplified model is expressed in the frequency domain:

$$h_{jk}(f) \approx \lambda_{jk} \cdot e^{-\jmath 2\pi f \tau_{jk}}, \tag{7}$$

where τ_{jk} and $\lambda_{jk} \geq 0$ are the time delay and attenuation from source k to sensor j, respectively. Since we cannot distinguish the phase (or amplitude) of $s_k(f,t)$ and $h_{jk}(f)$, these two parameters can be considered to be relative (this fact causes the scaling ambiguity). Thus, without loss of generality, we normalize them and align the scaling ambiguity by

$$\tau_{jk} = (d_{jk} - d_{Jk})/c, \tag{8}$$

$$\textstyle\sum_{j=1}^{M} \lambda_{jk}^2 = 1, \tag{9}$$

where d_{jk} is the distance from source k to sensor j (Fig. 2), and c is the propagation velocity of the signal. Normalization (8) makes $\tau_{Jk} = 0$, i.e. the relative time delay is zero at a reference sensor J.

By following the normalizations (8) and (9), the scaling ambiguity of basis vectors $\mathbf{a}_i$ is aligned by the operation

$$\mathbf{a}_i \leftarrow \frac{\mathbf{a}_i}{||\mathbf{a}_i||} e^{-\jmath \arg(a_{Ji})} \tag{10}$$

which makes $\arg(a_{Ji}) = 0$ and $||\mathbf{a}_i|| = 1$. Now, the task as regards the permutation problem is to determine a permutation Π_f that relates the subscript i and k with $i = \Pi_f(k)$, and to estimate parameters τ_{jk}, λ_{jk} that make the model (7) match the $a_{ji}(f)$ element of a basis vector. This can be formulated so as to find Π_f, τ_{jk} and λ_{jk} that minimize the cost function:

$$\mathcal{J} = \sum_{f \in \mathcal{F}} \sum_{k=1}^{N} \sum_{j=1}^{M} |a_{ji}(f) - \lambda_{jk} \cdot e^{-\jmath 2\pi f \tau_{jk}}|^2, \quad i = \Pi_f(k), \tag{11}$$

where $\mathcal{F}$ is the set of frequencies that we have to consider.

3.2 Clustering Frequency-Normalized Basis Vectors

If we consider the frequency range where spatial aliasing does not occur:

$$\mathcal{F} = \{f : -\pi < 2\pi f \tau_{jk} < \pi, {}^{\forall} j, k\} \tag{12}$$

we can introduce the frequency normalization technique [6] to minimize the cost function (11). Let $d_{\max}$ be the maximum distance between the reference sensor J and any other sensor. Then the relative time delay is bounded by

$$\max_{jk} |\tau_{jk}| \leq d_{\max}/c \tag{13}$$

and therefore the frequency range $\mathcal{F}$ can be expressed with

$$\mathcal{F} = \{f : 0 < f < \frac{c}{2d_{\max}}\}. \tag{14}$$

The frequency normalization technique [6] removes frequency dependence from the elements of scale-normalized basis vectors (10):

$$\bar{a}_{ji}(f) \leftarrow |a_{ji}(f)| \exp\left[\jmath \frac{\arg[a_{ji}(f)]}{4fc^{-1}d_{\max}}\right]. \tag{15}$$

The rationale of dividing the argument by $4fc^{-1}d_{\max}$ is discussed in [6]. With this operation, the cost function (11) is converted to

$$\bar{\mathcal{J}} = \sum_{f\in\mathcal{F}} \sum_{k=1}^{N} \sum_{j=1}^{M} |\bar{a}_{ji}(f) - \bar{h}_{jk}|^2, \;\; i = \mathit{\Pi}_f(k) \tag{16}$$

where

$$\bar{h}_{jk} = \lambda_{jk} \cdot \exp[-\jmath \frac{\pi}{2} \cdot \frac{c \cdot \tau_{jk}}{d_{\max}}] \tag{17}$$

is a frequency-normalized model. In a vector notation

$$\bar{\mathcal{J}} = \sum_{f\in\mathcal{F}} \sum_{k=1}^{N} ||\bar{\mathbf{a}}_i(f) - \bar{\mathbf{h}}_k||^2, \;\; i = \mathit{\Pi}_f(k), \tag{18}$$

where $\bar{\mathbf{a}}_i = [\bar{a}_{1i}, \ldots, \bar{a}_{Mi}]^T$ and $\bar{\mathbf{h}}_k = [\bar{h}_{1k}, \ldots, \bar{h}_{Mk}]^T$. Because the model $\bar{\mathbf{h}}_k$ do not depend on frequency, $\bar{\mathcal{J}}$ can be minimized efficiently by a clustering algorithm that iterates the following two updates until convergence:

$$\mathit{\Pi}_f \leftarrow \mathrm{argmin}_{\mathit{\Pi}} \sum_{k=1}^{N} ||\bar{\mathbf{a}}_{\mathit{\Pi}(k)}(f) - \bar{\mathbf{h}}_k||^2, \;\; \text{for each } f \in \mathcal{F}, \tag{19}$$

$$\bar{\mathbf{h}}_k \leftarrow \sum_{f\in\mathcal{F}} \bar{\mathbf{a}}_{\mathit{\Pi}_f(k)}(f), \;\; \bar{\mathbf{h}}_k \leftarrow \bar{\mathbf{h}}_k / ||\bar{\mathbf{h}}_k||, \;\; \text{for each } k = 1, \ldots, N. \tag{20}$$

The first update (19) optimizes the permutation $\mathit{\Pi}_f$ for each frequency f with the current model $\bar{\mathbf{h}}_k$. The second update (20) calculates the most probable model $\bar{\mathbf{h}}_k$ with the current permutations. This set of updates is very similar to that of the k-means algorithm [12].

After the algorithm has converged, the permutation ambiguity in each frequency bin is aligned by

$$\mathbf{a}_k(f) \leftarrow \mathbf{a}_{\mathit{\Pi}_f(k)}(f), \;\; y_k(f,t) \leftarrow y_{\mathit{\Pi}_f(k)}(f,t), \;\; k = 1, \ldots, N. \tag{21}$$

In addition to aligning the permutations, the method estimates the model parameters by

$$\tau_{jk} = -\frac{2}{\pi} \cdot \frac{d_{\max}}{c} \arg(\bar{h}_{jk}), \quad \lambda_{jk} = |\bar{h}_{jk}|. \tag{22}$$

From these parameters and sensor array geometry, we can perform source localization, such as direction-of-arrival (DOA) estimation.

4 Solving the Scaling Problem

The ultimate goal as regards the scaling problem is to recover each source $\mathrm{s}_k(t)$, i.e. multichannel blind deconvolution. However, this is very difficult with colored source signals, such as speech. A feasible goal [13, 14] is simply to recover the observation of each source k at a reference sensor J

$$\sum_l \mathrm{h}_{Jk}(l)\,\mathrm{s}_k(t-l)\,. \tag{23}$$

If we consider the frequency-domain counterpart of the above discussion, there is no practical way to recover the amplitude and phase of $s_k(f,t)$ blindly, but there is a feasible way to recover those of

$$h_{Jk}(f)s_k(f,t) \tag{24}$$

instead [3, 13]. We use this criterion for the scaling problem.

Calculating the inverse (5) and obtaining the linear combination form (6) of $\mathbf{x}(f,t)$ provides an instant solution to the scaling problem. If ICA works well and the permutation ambiguity is solved, we obtain separated signals $y_k(f,t)$ that should be close to source signals $s_k(f,t)$, now only up to the scaling ambiguity. If we compare (3) and (6), we see that $\mathbf{a}_k(f)y_k(f,t)$ should be close to $\mathbf{h}_k(f)s_k(f,t)$ and therefore $a_{Jk}(f)y_k(f,t)$ should be close to $h_{Jk}(f)s_k(f,t)$. Thus the scaling alignment can be performed simply by

$$y_k(f,t) \leftarrow a_{Jk}(f)y_k(f,t)\,. \tag{25}$$

In other words, there is no scaling ambiguity to be considered in (6) if we do not discriminate between $\mathbf{a}_i(f)$ and $y_i(f,t)$.

5 A Link to the Time-Frequency Masking Method

This section reveals a link between the ICA-based method and the time-frequency (T-F) masking method. The link becomes clear by the linear combination form (6) of $\mathbf{x}(f,t)$ obtained by the inverse (5) of the ICA separation matrix $\mathbf{W}(f)$.

Let us explain the T-F masking method, in which we assume the sparseness of source signals, i.e. at most only one source is active for each time-frequency slot (f,t). Based on this assumption, the mixture model (3) can be simplified as

$$\mathbf{x}(f,t) = \mathbf{h}_k(f)s_k(f,t), \quad k \in \{1,\ldots,N\}. \tag{26}$$

where k depends on each time-frequency slot (f,t). Then, the method classifies the observation vectors $\mathbf{x}(f,t), {}^{\forall}f,t$ into N clusters $C_1,\ldots,C_N$ so that the k-th cluster contains observation vectors in which the k-th source is the only active source. After the classification, time domain separated signals $\mathrm{y}_k(t)$ are obtained by applying an inverse STFT (ISTFT) to the following classified frequency components

$$y_k(f,t) = \begin{cases} x_J(f,t) & \mathbf{x}(f,t) \in C_k, \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

In the classification, the spatial information expressed in $\mathbf{x}(f,t)$ is extracted and used. Typically, the phase difference normalized with frequency and/or the amplitude difference between two sensors:

$$\frac{\arg[x_2(f,t)/x_1(f,t)]}{2\pi f} \quad \text{and/or} \quad \left|\frac{x_2(f,t)}{x_1(f,t)}\right| \tag{28}$$

are calculated for the classification [7-9]. However, these papers presented only cases with two sensors. Recently, we proposed a new technique for using all the information of more than two sensors [10]. The technique used there is similar to that presented in Sec. 3. Thus, we consider that various techniques for classifying observation vectors $\mathbf{x}(f,t)$ in the T-F masking method can be used to classify basis vectors $\mathbf{a}_i(f,t)$ for solving the permutation problem in the ICA-based method, and vice versa.

Let us discuss this relationship in the following. If the sparseness assumption is satisfied, the linear combination form (6) obtained by ICA is reduced to

$$\mathbf{x}(f,t) = \mathbf{a}_i(f) y_i(f,t), \quad i \in \{1,\ldots,N\}. \tag{29}$$

where i depends on each time-frequency slot (f,t). If we compare (26) and (29), we see that $\mathbf{h}_k(f)s_k(f,t)$ should be close to $\mathbf{a}_i(f)y_i(f,t)$ for each time-frequency slot (f,t). Thus, the spatial information expressed in observation vectors $\mathbf{x}(f,t)$ with the sparseness assumption (26) is the same as that of basis vectors $\mathbf{a}_i(f,t)$ up to the scaling ambiguity. Therefore, we can use the same techniques for extracting spatial information from observation vectors $\mathbf{x}(f,t)$ and basis vectors $\mathbf{a}_i(f,t)$.

The normalization formulas (10) and (15) and the clustering procedure of (19) and (20) can be used not only for the ICA-based method but also the T-F masking method. Of course, we need to replace $\mathbf{a}_i(f)$ with $\mathbf{x}(f,t)$ and modify (19) for a standard clustering algorithm such as the k-means algorithm [12]. Figure 3 shows the flows of both methods in accordance with this idea.

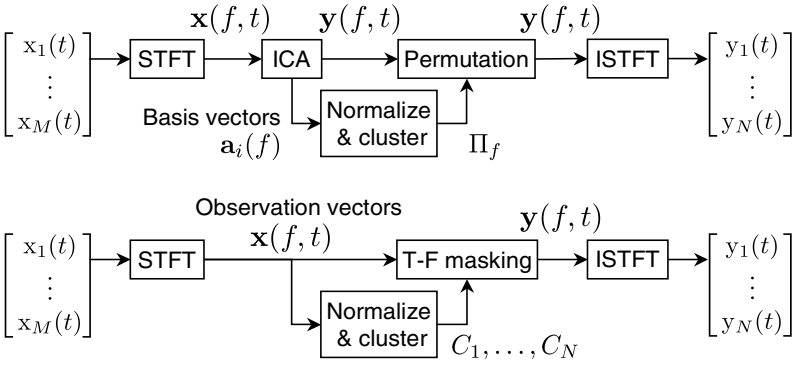

Fig. 3. Flows of ICA-based method (above) and time-frequency (T-F) masking method (below)

6 Experimental Results

We have performed experiments with the conditions shown in Fig. 4. We used 16 combinations of three speeches to evaluate the separation performance. The system did not have to know the sensor geometry for solving the permutation problem, but just the maximum distance $d_{\max} = 4$ cm between the reference sensor and any other sensor. The computational time was less than 3 seconds for 3-second speech mixtures, meaning that real-time processing was possible.

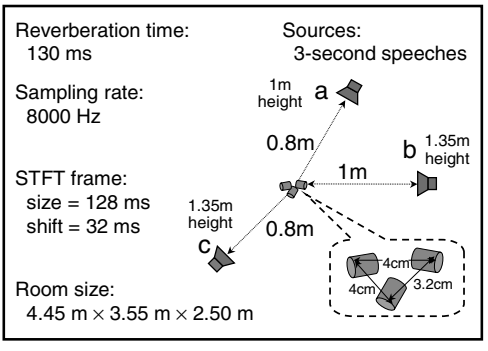

Fig. 4. Experimental conditions

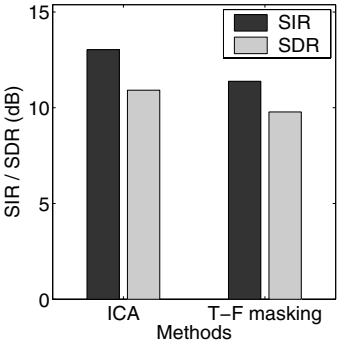

Fig. 5. Separation performance

Figure 5 shows the average signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR), whose detailed definitions can be found in [10]. Basically, the SIR indicates how well the mixtures are separated into the sources, and the SDR indicates how close each separated signal is to the observation of the corresponding source at the reference sensor. Since the number of sensors was sufficient for the number of sources in this case, ICA-based method worked better than T-F masking as shown in Fig. 5. We have also already obtained results with another setup where the number of sensors was insufficient ($N = 4, M = 3$) and the T-F masking method still worked [10].

7 Conclusion

In the ICA-based frequency-domain BSS, the permutation and scaling ambiguity of the ICA solution should be aligned. Once we have the form (6) by calculating the inverses of ICA separation matrices $\mathbf{W}$, the scaling ambiguity does not have to be considered. To align the permutation ambiguity, we can exploit the mixing system information represented in basis vectors $\mathbf{a}_i$, as Sec. 3 presents an efficient method. The form (6) clarifies the relationship between the ICA-based method and the T-F masking method. The same technique as that presented in Sec. 3 can be used in the T-F masking method for clustering observations $\mathbf{x}(f, t)$ and extracting the mixing system information.

References

1. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. Neurocomputing **22** (1998) 21–34
2. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. IEEE Trans. Speech Audio Processing **8** (2000) 320–327
3. Murata, N., Ikeda, S., Ziehe, A.: An approach to blind source separation based on temporal structure of speech signals. Neurocomputing **41** (2001) 1–24
4. Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., Shikano, K.: Blind source separation combining independent component analysis and beamforming. EURASIP Journal on Applied Signal Processing **2003** (2003) 1135–1146
5. Mukai, R., Sawada, H., Araki, S., Makino, S.: Frequency domain blind source separation for many speech signals. In: Proc. ICA 2004 (LNCS 3195). Springer-Verlag (2004) 461–469
6. Sawada, H., Araki, S., Mukai, R., Makino, S.: Blind extraction of a dominant source signal from mixtures of many sources. In: Proc. ICASSP 2005. Volume III. (2005) 61–64
7. Aoki, M., Okamoto, M., Aoki, S., Matsui, H., Sakurai, T., Kaneda, Y.: Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. Acoustical Science and Technology **22** (2001) 149–157
8. Rickard, S., Balan, R., Rosca, J.: Real-time time-frequency based blind source separation. In: Proc. ICA2001. (2001) 651–656
9. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Processing **52** (2004) 1830–1847
10. Araki, S., Sawada, H., Mukai, R., Makino, S.: A novel blind source separation method with observation vector clustering. In: Proc. 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC 2005). (2005)
11. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons (2001)
12. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2nd edn. Wiley Interscience (2000)
13. Matsuoka, K., Nakashima, S.: Minimal distortion principle for blind source separation. In: Proc. ICA 2001. (2001) 722–727
14. Takatani, T., Nishikawa, T., Saruwatari, H., Shikano, K.: Blind separation of binaural sound mixtures using SIMO-model-based independent component analysis. In: Proc. ICASSP 2004. Volume IV. (2004) 113–116