ICA-Based Clustering for Resolving Permutation Ambiguity in Frequency-Domain Convolutive Source Separation

Minje Kim and Seungjin Choi
Department of Computer Science
Pohang University of Science and Technology, Korea
{minjekim,seungjin}@postech.ac.kr

Abstract

Permutation ambiguity is an inherent limitation in independent component analysis, which is a bottleneck in frequency-domain methods of convolutive source separation. In this paper we present a method for resolving this permutation ambiguity, where we group vectors of estimated frequency responses into clusters in such a way that each cluster contains frequency responses associated with the same source. The clustering is carried out, applying independent component analysis to estimated frequency responses. In contrast to existing methods, the proposed method does not require any prior information such as the geometric configuration of microphone arrays or distances between sources and microphones. Experimental results confirm the validity of our method.

1. Introduction

Blind source separation (BSS) is a problem of restoring independent sources from their mixtures, without resorting to any information on sources and mixing characteristics [3, 1]. Two inherent indeterminacies in BSS or independent component analysis (ICA) are scaling and permutation ambiguities. In convolutive source separation, mixtures (corresponding microphone signals in the domain of speech processing) are assumed to be generated by convolving sources with a multivariate FIR filter. Frequency-domain methods of convolutive source separation, in general, transform time-domain multivariate signals into frequency-domain using short-time Fourier transform (STFT), so that the problem is converted into a task of demixing instantaneous mixture at each frequency bin. However, a bottleneck in frequency-domain BSS methods, lies in the frequency permutation ambiguity which leads to different ordering of sources at each frequency bin.

Various methods have been developed, in order to tackle this permutation ambiguity. Imposing constraints on demixing filters such as smoothing, might be a good solution, but it can not be used when mixing filter length is too long [10, 8]. Correlations between envelopes of band-passed signals are useful, but are not robust since misalignment at a frequency is propagated through consecutive frequencies [7]. Direction of arrival (DOA) estimation methods were used in [4], where the wavelength should be longer than the half of distance between sensors. This is not adequate for the case where the sampling rate of observed signals is high. Processing signals in the time-domain, is free of frequency permutation ambiguity [2], however, we lose the benefit of frequency-domain methods (fast processing due to FFT) [6, 5].

Recently, Sawada *et al.* proposed a method based on clustering basis vectors of estimated mixing filter (corresponding to estimated frequency responses), where only prior information on the maximal distance between sensors is required and a sophisticated normalization method was introduced [9]. Motivated by this work, we present a method of ICA-based clustering where any prior information or a normalization procedure is not required. The basic idea of our method is to exploit basis vectors of representing estimated frequency responses, in order to group them into clusters, each of which contain frequency responses associated with the same source. We use ICA for frequency-domain convolutive source separation as well as for resolving frequency permutation ambiguity, where the former takes observed signals as input data and the latter takes estimated frequency responses as input data.

2. Frequency-Domain Convolutive Source Separation

Convolutive source separation aims at restoring sources $s_1(t), \ldots, s_n(t)$ from m ($m \geq n$) sensor signals, x_i, $i =$

$1, \ldots, m$, that are modelled as

$$x_i(t) = \sum_{j=1}^{n} \sum_{l} h_{ij}(l) s_j(t-l),\ i = 1, \ldots, m, \qquad (1)$$

where $h_{ij}(l)$ represents the channel impulse response from source j to sensor i. We apply L-point short-time Fourier transform (STFT) to sensor signals, i.e.,

$$x_i(f, \tau) = \sum_{r=-L/2}^{L/2-1} x_i(\tau + r)\rho(r) e^{-\jmath 2\pi f r}, \qquad (2)$$

where $f \in \{0, \frac{1}{L} f_s, ..., \frac{L-1}{L} f_s\}$, f_s represents the sampling frequency, and $\rho(r)$ is a window function (for example, Hanning window).

A linear convolution can be approximated by a circular convolution if the frame size of FFT is much larger than the channel length. The convolutive mixture model in (1) is approximated by

$$x_i(f, \tau) \approx \sum_{j=1}^{n} h_{ij}(f) s_j(f, \tau), \qquad (3)$$

where $h_{ij}(f)$ is the frequency response from source j to sensor i, and $s_j(f, \tau)$ is STFT of $s_j(t)$ as in (2). In a compact form, we can write (3) as

$$\boldsymbol{x}(f, \tau) \approx \boldsymbol{H}(f) \boldsymbol{s}(f, \tau), \qquad (4)$$

where $\boldsymbol{x}(f, \tau) = [x_1(f, \tau), \ldots, x_m(f, \tau)]^\top$, $\boldsymbol{s}(f, \tau) = [s_1(f, \tau), \ldots, s_m(f, \tau)]^\top$, and $\boldsymbol{H}(f) = [h_{ij}(f)] \in \mathbb{C}^{m \times n}$.

Frequency-domain source separation consists in estimating a demixing matrix $\boldsymbol{W}(f)$ such that $\boldsymbol{y}(f, \tau) = \boldsymbol{W}(f)\boldsymbol{x}(f, \tau)$ corresponds to scaled and permuted version of $\boldsymbol{s}(f, \tau)$. Any currently available ICA methods [3, 1] can be applied to estimate the frequency-domain demixing matrix $\boldsymbol{W}(f)$. Estimated mixing matrix $\boldsymbol{A}(f) = \boldsymbol{W}^{-1}(f) = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n(f)]$ is associated with the inverse of $\boldsymbol{W}(f)$. At different frequency bins, $f \in \{0, \frac{1}{L} f_s, ..., \frac{L-1}{L} f_s\}$, different permutation happens, due to the inherence indeterminacy in ICA. This causes a serious problem in transforming frequency-domain signals $\boldsymbol{y}(f, \tau)$ back to time-domain signals $\boldsymbol{y}(t)$ in an appropriate way.

3. ICA-Based Permutation Re-Ordering

We start with a quick review of the normalization method in [9], (since our work was inspired by it), where the following near-field model for channel impulse responses was used:

$$h_{ij}(f) \approx \frac{q(f)}{d_{ij}} \exp\left\{\jmath 2\pi f c^{-1}(d_{ij} - d_{Rj})\right\}, \qquad (5)$$

where c is the propagation velocity and $d_{ij} > 0$ is the distance between source j and sensor i with the subscript R representing the reference sensor. Sawada *et al.* [9] introduced the following normalization

$$\bar{a}_{ik}(f) \leftarrow |a_{ik}(f)| \exp\left\{\jmath \frac{\arg(a_{ik}(f)/a_{Rk}(f))}{4 f c^{-1} d_{\max}}\right\}, \qquad (6)$$

where $d_{\max}$ is the maximum distance between the preselected reference sensor R and a sensor $i \in \{1, ..., m\}$, followed by a unit-norm normalization, for basis vectors $\bar{\boldsymbol{a}}_i(f)$.

Taking the frequency permutation into account, i.e., $\boldsymbol{a}_i(f) \approx \boldsymbol{h}_k(f)$ *(i and j can be different)* and incorporating the normalization (6) into the near-field model (5), leads to

$$\bar{a}_{ij}(f) \approx \frac{1}{d_{ik} D} \exp\left\{\jmath \frac{\pi}{2} \frac{(d_{ik} - d_{Rk})}{d_{\max}}\right\}, \qquad (7)$$

where $D = \sqrt{\sum_{i=1}^{m} 1/d_{ik}^2}$. Note that Eq. (7) is free of frequency-dependent factors and depends only on positions of sensors and sources. For example, see Fig. 1 where we plot real parts of 2-dimensional basis vectors $\boldsymbol{a}_i(f)$, $i = 1, 2$ (i.e., 2 sources and 2 sensors) with respect to frequencies, and associated normalized basis vectors.

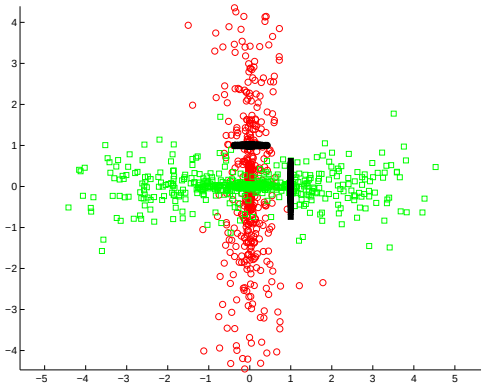


Figure 1. Squares and circles represent scatter plots of real parts of 2-dimensional basis vectors, $\boldsymbol{a}_1(f)$ and $\boldsymbol{a}_2(f)$ over $f \in \{0, \frac{1}{L} f_s, ..., \frac{L-1}{L} f_s\}$. Associated normalized basis vectors, $\bar{\boldsymbol{a}}_1(f)$ and $\bar{\boldsymbol{a}}_2(f)$, form two clusters represented by a horizonal bar and a vertical bar, each of which is the contribution of each source.

Although the sophisticated normalization (7) eliminates frequency-dependent factors in basis vectors, which allows us to easily group basis vectors into clusters, each of which contains basis vectors associated with the same source. However, we found out that the performance depends on

what to choose as a reference sensor. Fig. 1 inspires a conjecture that the $\boldsymbol{a}(f)$ associated with the same source, lies in the same direction, regardless of frequency bins. This direction is referred to as an *intrinsic direction*. Our method is to find these intrinsic directions which are expected to correspond to ICA basis vectors when taking $\boldsymbol{a}_i(f)$ as input data (see Fig. 2). In contrast to the normalization method, our approach does not require any prior knowledge on sensor locations, as well as the sophisticated normalization step.

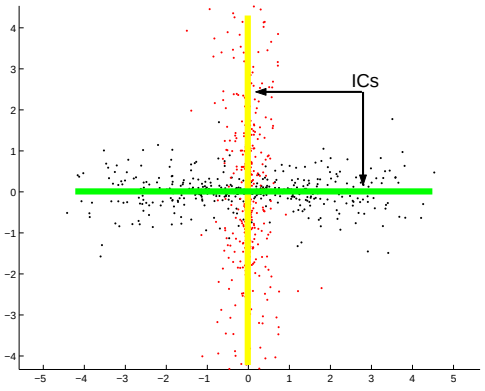


Figure 2. Two basis vectors $\boldsymbol{a}_1(f)$ and $\boldsymbol{a}_2(f)$ form two intrinsic directions that can be determined by basis vectors computed by ICA as taking $\boldsymbol{A}(f)$ as input.

Suppose that frequency-domain convolutive source separation methods already estimated $\boldsymbol{A}(f) = \boldsymbol{W}^{-1}(f)$,

$$\boldsymbol{A}(f) = [\boldsymbol{a}_1(f), \boldsymbol{a}_2(f), ..., \boldsymbol{a}_n(f)].$$

We construct a data matrix $\widetilde{\boldsymbol{X}} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L]$, where $\boldsymbol{A}_k = \boldsymbol{A}\big(\frac{(k-1)f_s}{L}\big)$. Then we apply ICA to $\widetilde{\boldsymbol{X}}$ to find the following decomposition

$$\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{S}},$$

where $\widetilde{\boldsymbol{A}} \in \mathbb{C}^{m\times n}$ and $\widetilde{\boldsymbol{S}} \in \mathbb{C}^{n\times nL}$ denote the ICA basis matrix and encoding variable matrix associated with $\widetilde{\boldsymbol{X}}$, not the sensor signal. Each column vector of $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{a}}_1, \ldots, \widetilde{\boldsymbol{a}}_n]$ is normalized to have unit Euclidean norm.

Clustering is done by considering absolute values of encoding variables that represent the contribution of basis vectors. For the case of $n = 2$, as an example, consider a data point $\widetilde{\boldsymbol{x}}_l$. In such a case, we have two basis vectors $\widetilde{\boldsymbol{a}}_1$ and $\widetilde{\boldsymbol{a}}_2$ and associated encoding variables $\widetilde{\boldsymbol{s}}_l = [\widetilde{s}_{1,l}, \widetilde{s}_{2,l}]^\top$. $\widetilde{\boldsymbol{x}}_l$ is assigned to cluster 1, if $|\widetilde{s}_{1,l}| > |\widetilde{s}_{2,l}|$, and is assigned to cluster 2 otherwise. For $n = 2$, in fact, we have to consider two consecutive data points, $\widetilde{\boldsymbol{x}}_l$ and $\widetilde{\boldsymbol{x}}_{l+1}$, that are associated with two column vectors of $\boldsymbol{A}_l$. If $\widetilde{\boldsymbol{x}}_l$ is assigned to cluster 1, then $\widetilde{\boldsymbol{x}}_{l+1}$ should be assigned to cluster 2, since both vectors should not belong to the same cluster. However, there might be a case where $|\widetilde{s}_{1,l}| > |\widetilde{s}_{2,l}|$ and $|\widetilde{s}_{1,l+1}| > |\widetilde{s}_{2,l+1}|$. In such a case, we take the ratio of encoding variables into account and assign $\widetilde{\boldsymbol{x}}_l$ and $\widetilde{\boldsymbol{x}}_{l+1}$ to cluster 1 and 2, respectively, if $|\widetilde{s}_{1,l}|/|\widetilde{s}_{2,l}| > |\widetilde{s}_{1,l+1}|/|\widetilde{s}_{2,l+1}|$. This idea can be easily generalized to the case of $n > 2$.

Our ICA-based clustering method has some advantages over the Sawada's method [9], in two aspects:

- Our method does not require prior information such as distances between sensors.
- Our method does not require the sophisticated normalization (7), the performance of which depends on the selection of a reference sensor.

4. Numerical Experiments

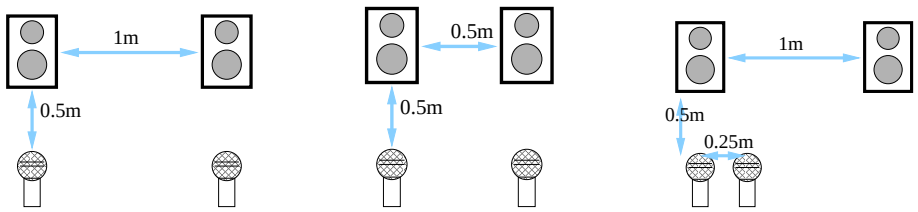
We present three simulation results, with comparison to the normalization method in [9]. Two numerical examples are to emphasize the clustering performance of our ICA-based method, compared to the normalization method.

4.1. Clustering Performance

We measured impulse responses using two speakers and two microphones with three different geometric configurations shown in Fig. 3:

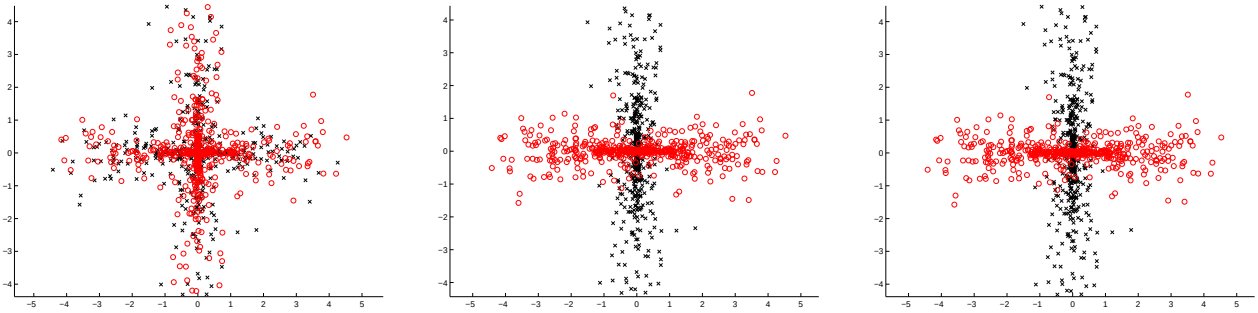
- **Case 1:** The geometric configuration of microphones and speakers, is shown in Fig 3 (a), where the distance between speakers is 1m and each microphone is separated by 0.5m from its associated speaker.
- **Case 2:** The distance between speakers as well as the distance between microphones, are reduced to 0.5m. (see Fig. 3 (b))
- **Case 3:** Both microphones separated by 0.25m, are close to speaker 1. (see Fig. 3 (c))

We transform these measure impulse responses into frequency responses, then artificially make permutation to generate $\boldsymbol{a}_1(f)$ and $\boldsymbol{a}_2(f)$ ($L = 512$ and $n = 2$). ICA-based clustering results (where we applied the natural gradient ICA algorithm with a complex nonlinear function), in terms of accuracy (clustering error rate based on known class labels), are summarized in Table 1. For the case where each microphone is close to a corresponding speaker, basis vectors $\boldsymbol{a}_1(f)$ and $\boldsymbol{a}_2(f)$ are well clustered, since intrinsic directions are well separated. However, if both microphones are close to a single speaker, then clustering performance becomes worse in our method as well as the normalization method. The performance of the normalization



(a) (b) (c)

Figure 3. Three different geometric configurations of speakers and microphones, are shown in (a), (b), (c), corresponding to case 1, case 2, case 3, respectively.



(a) (b) (c)

Figure 4. Results of re-ordering artificially permuted frequency responses, are shown: (a) before re-ordering; (b) re-ordering using the normalization method [9]; (c) re-ordering using our method. Circle points are associated with the first column vectors and cross points are associated with the second ones. Circle and cross points are appropriately lined up in (b) and (c), which implies that the permutation problem is resolved.

method is slightly different, depending on the selection of a reference sensor. Fig 4 shows scatter plots of basis vectors for case 1.

Table 1. Comparison of our ICA-based clustering to the normalization method [9], in terms of accuracy. Norm-1 and Norm-2 represent the normalization method, with sensor 1 or 2 being a reference sensor, respectively.

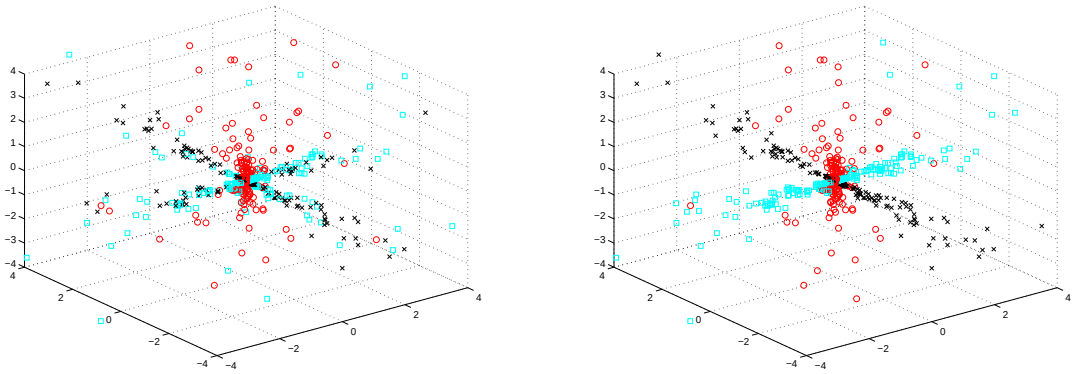
	ICA	Norm-1	Norm-2
Case 1	0.20	0.59	0.59
Case 2	3.71	4.88	5.08
Case 3	27.34	31.05	29.10

4.2. Separation Results

In order to show the validity of our method, we applied our ICA-based clustering to the task of resolving permutation ambiguity in a frequency-domain source separation method. We used the natural gradient ICA algorithm with a complex nonlinear function (for example, [10]) for both frequency-domain convolutive source separation and ICA-based clustering. We recorded two microphone signals using a male voice and a female voice, with the configuration in Fig. 3 (a). The sampling rate was 16kHz and 1024-point FFT was used. The performance in terms of signal-to-interference ratio (SIR) improvement, was investigated using the method in [8], with comparison to the normalization method (see Table 2).

4.3. Clustering for $n = 3$

Fig.5 shows scatter plots of basis vectors $\boldsymbol{a}_i(f)$, $i = 1, 2, 3$, for the case of three sources and three sensors. Our



(a) (b)

Figure 5. Re-ordering results using our method for the case of $n = 3$, is shown: (a) before re-ordering; (b) after re-ordering using our ICA-based clstering method.

Table 2. SIR improvements (dB).

Input SIR	ICA	Norm-1	Norm-2
6.6163	8.2823	7.7444	8.1493

method can be applied to the case of $n > 2$, although the computational complexity slightly increases.

5. Conclusions

We have presented a method of resolving frequency permutation ambiguity in frequency-domain source separation methods. The key idea was to exploit geometric information of basis vectors associated with independent components of frequency responses estimated at frequency bins. These basis vectors of representing frequency responses, led us to group them into clusters containing frequency responses associated with the same source. Experimental comparison have shown that the ICA-based clustering method was superior to the sophisticated normalization method [9] in resolving frequency permutation. Moreover, in contrast to [9], our method did not require any prior knowledge on distances between sensors.

Acknowledgments: This work was supported by ITEP Brain Neuroinformatics Program and Korea MIC under ITRC support program supervised by the IITA (IITA-2005-C1090-0501-0018).

References

[1] S. Choi, A. Cichocki, H. M. Park, and S. Y. Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing - Letters and Review*, 6(1):1–57, 2005.

[2] S. Choi, H. Hong, H. Glotin, and F. Berthommier. Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network. *Neurocomputing*, 49(1-4):299–314, 2002.

[3] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, Inc., 2002.

[4] M. Ikram and D. Morgan. A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 881–884, Salk Lake City, Utah, 2001.

[5] M. Joho and P. Schniter. Frequency domain realization of a multichannel blind deconvolution algorithm based on the natural gradient. In *Proc. Int'l Conf. Independent Component Analysis and Blind Signal Separation*, pages 543–548, Nara, Japan, 2003.

[6] R. Lambert and A. Bell. Blind separation of multiple speakers in a multipath environment. In *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 423–426, Munich, Germany, 1997.

[7] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41:1–24, 2001.

[8] L. C. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. Speech and Audio Processing*, pages 320–327, May 2000.

[9] H. Sawada, S. Araki, R. Mukai, and S. Makino. Blind extraction of a dominant sourcee signal from mixtures of many sources using ICA and time-frequency masking. In *Proc. IEEE Int'l Symp. Circuits and Systems*, pages 5882–5885, Kobe, Japan, 2005.

[10] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34, 1998.