ACOUSTICAL LETTER

Interactive speech source separation based on independent low-rank matrix analysis

Fuga Oshima[1], Masaki Nakano[2] and Daichi Kitamura[1,*]

[1]*National Institute of Technology, Kagawa College, Chokushi 355, Takamatsu, 761–8058 Japan*
[2]*University of Tsukuba, 1–1–1 Tennodai, Tsukuba, 305–8577 Japan*

(*Received 25 February 2021, Accepted for publication 16 March 2021*)

Keywords: Blind source separation, Independent component analysis, Nonnegative matrix factorization, User annotation

1. Introduction

Blind source separation (BSS) [1] is a technique to estimate source signals in an observed mixture signal without knowing any prior information. When the number of sources is equal to or greater than the number of sensors (microphones in audio cases), which is called the (over-)determined situation, independent component analysis (ICA) [2] and its extensions have been studied. Since audio sources are convolutively mixed with room reverberation, frequency-domain ICA (FDICA) [3] is a basic algorithm for audio BSS.

In FDICA, scales and permutation of frequency-wise estimated signals cannot be determined, and post-processes are required to fix these ambiguities. The scale ambiguity can easily be recovered by the back projection [4], whereas the permutation alignment of the estimated signals for all the frequencies (so-called the permutation problem [5]) has been tackled for many years. In particular, independent vector analysis (IVA) [6,7] and independent low-rank matrix analysis (ILRMA) [8,9] are the most successful approaches for solving the permutation problem. These methods assume specific time-frequency structures of sources (source models) to avoid encountering the permutation problem during the optimization.

Although ILRMA accurately separates mixtures of music sources, the performance of IVA and ILRMA often degrades for speech mixtures. The main reason of the degradation in IVA and ILRMA is a block permutation problem [10,11], namely, the permutation misalignment of the estimated signals occurs with subband blocks (see Fig. 1). This is caused by a mismatch between the source model assumed in IVA or ILRMA and actual time-frequency structure of speech sources.

In this letter, we propose to solve the block permutation problem based on user annotations and develop a new interactive speech source separation system. In the proposed system, ILRMA-based source separation is performed while the user annotation is utilized for escaping from bad local minima in ILRMA optimization and searching for more accurate separation results.

2. BSS based on ILRMA

Let N and M be the numbers of sources and microphones, respectively. The time-frequency components of source, mixture, and estimated signals are respectively defined as

$$\boldsymbol{s}_{i,j} = [s_{i,j,1}, \cdots, s_{i,j,n}, \cdots, s_{i,j,N}]^{\mathrm{T}} \in \mathbb{C}^N, \tag{1}$$

$$\boldsymbol{x}_{i,j} = [x_{i,j,1}, \cdots, x_{i,j,m}, \cdots, x_{i,j,M}]^{\mathrm{T}} \in \mathbb{C}^M, \tag{2}$$

$$\boldsymbol{y}_{i,j} = [y_{i,j,1}, \cdots, y_{i,j,n}, \cdots, y_{i,j,N}]^{\mathrm{T}} \in \mathbb{C}^N, \tag{3}$$

where $i = 1, 2, \cdots, I$, $j = 1, 2, \cdots, J$, $n = 1, 2, \cdots, N$, and $m = 1, 2, \cdots, M$ are the indices of frequency, time, source, and microphone, respectively, and $\cdot^{\mathrm{T}}$ denotes the transpose. We assume that the observed signal is represented by

$$\boldsymbol{x}_{i,j} = \boldsymbol{A}_i \boldsymbol{s}_{i,j}, \tag{4}$$

where $\boldsymbol{A}_i \in \mathbb{C}^{M \times N}$ is a frequency-wise mixing matrix. Hereafter, we consider the determined situation ($M = N$). If $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1}$ exists for all the frequencies, the estimated signal can be obtained by

$$\boldsymbol{y}_{i,j} = \boldsymbol{W}_i \boldsymbol{x}_{i,j}, \tag{5}$$

where $\boldsymbol{W}_i = [\boldsymbol{w}_{i,1}\ \boldsymbol{w}_{i,2} \cdots \boldsymbol{w}_{i,N}]^{\mathrm{H}}$ is called the demixing matrix and $\cdot^{\mathrm{H}}$ denotes the Hermitian transpose.

ILRMA is an extension of IVA: the low-rank source model based on nonnegative matrix factorization (NMF) [12] is introduced. ILRMA optimizes both the demixing matrix $\boldsymbol{W}_i$ and the NMF source model $\boldsymbol{T}_n \boldsymbol{V}_n$, where $\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I \times L}$ and $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{L \times J}$ are the basis and activation matrices in NMF.

In ILRMA, the following cost function is minimized:

$$\mathcal{J} = -2J \sum_i \log |\det \boldsymbol{W}_i| + \sum_{i,j,n} \left[\frac{|\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{i,j}|^2}{\sum_l t_{i,l,n} v_{l,j,n}} + \log \sum_l t_{i,l,n} v_{l,j,n} \right], \tag{6}$$

where $t_{i,l,n}$ and $v_{l,j,n}$ are the elements of $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$, respectively, and $l = 1, 2, \cdots, L$ is the index of NMF bases. The minimization of (6) w.r.t. $\boldsymbol{W}_i$ is performed via iterative projection [7] as

$$\boldsymbol{U}_{i,n} = \frac{1}{J} \sum_j \frac{1}{\sum_l t_{i,l,n} v_{l,j,n}} \boldsymbol{x}_{i,j} \boldsymbol{x}_{i,j}^{\mathrm{H}}, \tag{7}$$

$$\boldsymbol{w}_{i,n} \leftarrow (\boldsymbol{W}_i \boldsymbol{U}_{i,n})^{-1} \boldsymbol{e}_n, \tag{8}$$

$$\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n} (\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{U}_{i,n} \boldsymbol{w}_{i,n})^{-\frac{1}{2}}, \tag{9}$$

where $\boldsymbol{e}_n \in \mathbb{R}_{\{0,1\}}^N$ is a unit vector whose nth element is unity. Also, the minimization of (6) w.r.t. $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$ is performed by iterating the following update rules:

[*]e-mail: kitamura-d@t.kagawa-nct.ac.jp

[doi:10.1250/ast.42.222]

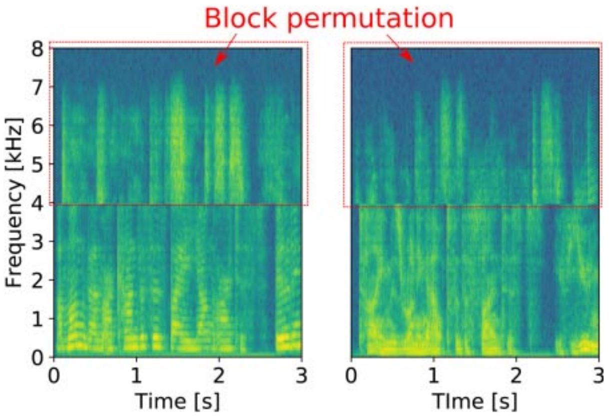

Fig. 1 Example of estimated signals with block permutation problem. Components over 4 kHz are swapped as subband block because of permutation misalignment.

$$t_{i,l,n} \leftarrow t_{i,l,n}\sqrt{\frac{\sum_j |\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{i,j}|^2 \left(\sum_{l'} t_{i,l',n} v_{l',j,n}\right)^{-2} v_{l,j,n}}{\sum_j v_{l,j,n}\left(\sum_{l'} t_{i,l',n} v_{l',j,n}\right)^{-1}}}, \tag{10}$$

$$v_{l,j,n} \leftarrow v_{l,j,n}\sqrt{\frac{\sum_i |\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{i,j}|^2 \left(\sum_{l'} t_{i,l',n} v_{l',j,n}\right)^{-2} t_{i,l,n}}{\sum_i t_{i,l,n}\left(\sum_{l'} t_{i,l',n} v_{l',j,n}\right)^{-1}}}. \tag{11}$$

Thus, all the variables can be optimized by iterating (7)–(11) with appropriate initialization, and the estimated signal is obtained by (5). However, since the cost function (6) is nonconvex, there exist many local minima in this optimization: some of them provide poor separation performance because of the block permutation problem.

3. Interactive speech source separation

3.1. Motivation and system overview

BSS of IVA and ILRMA often encounters the block permutation problem [10,11]. This is because the parameter optimization gets stuck in a bad local minimum, which confuses the permutation-misaligned estimated signals with the source signals. However, as shown in Fig. 1, the boundary frequency of the block permutation problem is often visually recognizable by users. On the basis of this idea, we propose to solve the block permutation problem using user annotations and develop a new interactive speech source separation system.

Figure 2 shows the user interface of the proposed system, where the system is implemented to work on web browsers. The system returns spectrograms of the estimated signals obtained by ILRMA to the user as a temporary result. The user can listen these estimated signals by clicking play buttons. When the user noticed that the estimated signals contain a permutation-misaligned subband or the separation performance is not satisfactory, a region-based annotation (see the red rectangle in Fig. 2) can be fed back to the system. The detailed explanations of this annotation and processing after the feedback are presented in the following subsections. Then, ILRMA re-estimates the demixing matrix using the user annotation to output permutation-fixed and more accurate estimated signals. This interaction between ILRMA and the user can be iterated until the satisfactory separation is achieved.

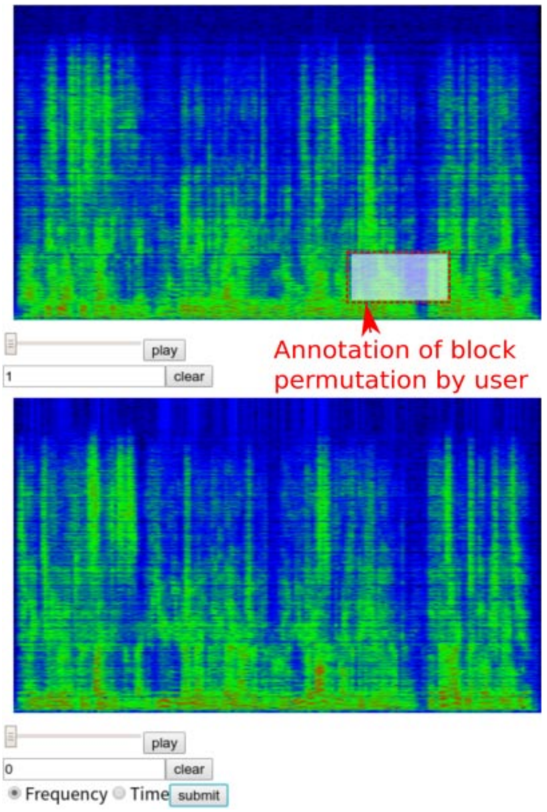

Fig. 2 User interface of proposed interactive speech source separation system in two-source case. Spectrograms are temporarily estimated signals obtained by ILRMA, where block permutation problem is occurred in annotated frequency subband.

3.2. Frequency annotation with block permutation problem

When the user noticed that the estimated signals contain a permutation-misaligned subband, frequency annotation should be fed back to the system. Let $i = i_{\mathrm{s}}$ and $i = i_{\mathrm{e}}$ $(1 \leq i_{\mathrm{s}} < i_{\mathrm{e}} \leq I)$ be the frequency indices of the lowest and the highest frequencies in the permutation-misaligned subband, respectively. Also, let $n = n_{\mathrm{s}}$ and $n = n_{\mathrm{t}}$ be the source and target indices of the signals, respectively, i.e., $y_{i_{\mathrm{s}},j,n_{\mathrm{s}}}, y_{i_{\mathrm{s}}+1,j,n_{\mathrm{s}}}, \cdots, y_{i_{\mathrm{e}},j,n_{\mathrm{s}}}$ and $y_{i_{\mathrm{s}},j,n_{\mathrm{t}}}, y_{i_{\mathrm{s}}+1,j,n_{\mathrm{t}}}, \cdots, y_{i_{\mathrm{e}},j,n_{\mathrm{t}}}$ are mistakenly swapped. These indices are obtained from the frequency annotation, which is a frequency range of the selected region, by the user. In this case, since we need to swap the demixing filters $\boldsymbol{w}_{i,n}$ and the basis components $t_{i,k,n}$, the following process is performed:

$$\begin{aligned}&\boldsymbol{w}_{i_{\mathrm{s}},n_{\mathrm{s}}}, \boldsymbol{w}_{i_{\mathrm{s}}+1,n_{\mathrm{s}}}, \cdots, \boldsymbol{w}_{i_{\mathrm{e}},n_{\mathrm{s}}}\\ &\qquad \Leftrightarrow \boldsymbol{w}_{i_{\mathrm{s}},n_{\mathrm{t}}}, \boldsymbol{w}_{i_{\mathrm{s}}+1,n_{\mathrm{t}}}, \cdots, \boldsymbol{w}_{i_{\mathrm{e}},n_{\mathrm{t}}},\end{aligned} \tag{12}$$

$$\begin{aligned}&t_{i_{\mathrm{s}},k,n_{\mathrm{s}}}, t_{i_{\mathrm{s}}+1,k,n_{\mathrm{s}}}, \cdots, t_{i_{\mathrm{e}},k,n_{\mathrm{s}}}\\ &\qquad \Leftrightarrow t_{i_{\mathrm{s}},k,n_{\mathrm{t}}}, t_{i_{\mathrm{s}}+1,k,n_{\mathrm{t}}}, \cdots, t_{i_{\mathrm{e}},k,n_{\mathrm{t}}} \ \forall k,\end{aligned} \tag{13}$$

where $\Leftrightarrow$ denotes the swapping process between the each component of left-hand and right-hand sides. In addition, the activation matrices of the corresponding sources are reset as

$$v_{k,j,n_\mathrm{s}}, v_{k,j,n_\mathrm{t}} \leftarrow \rho, \rho \ \forall k, j, \tag{14}$$

where $\leftarrow$ denotes the substitution of each component and ρ is a random value that obeys the uniform distribution in the range $(0, 1)$. The other parameters are inherited from the previously performed ILRMA. Thus, we expect that the optimization parameters can escape from the bad local minimum with block permutation problem by continuing the ILRMA algorithm after applying (12)–(14).

3.3. Time annotation with silent segment of the other sources (I)

When the user does not satisfied with the quality of separation, a time annotation is useful to improve the performance. In this case, the user annotates a silent time segment of the other sources, namely, the silent segment must contain only one source components. This annotation is reasonable because most of time of a typical speech mixture (conversation) contains only one speaker, which can easily be annotated by listening the observed or the temporarily estimated signal. Similar approach was proposed with IVA [13], but, to the best of our knowledge, ILRMA-based BSS using the time annotation has not been proposed yet.

Let $j = j_\mathrm{s}$ and $j = j_\mathrm{e}$ $(1 \leq j_\mathrm{s} < j_\mathrm{e} \leq J)$ be the time indices of the beginning and the ending time frames in the silent segment, respectively. Also, let $n = n_\mathrm{t}$ be the active source in the silent segment, i.e., the sources $n \neq n_\mathrm{t}$ are inactive. These indices are obtained from the time annotation, which is a time range of the selected region, by the user. In this case, since we need to suppress the time-frequency variance $\boldsymbol{T}_n \boldsymbol{V}_n$ of the silent sources, the following process is performed:

$$\begin{aligned} &v_{k,j_\mathrm{s},n}, v_{k,j_\mathrm{s}+1,n}, \cdots, v_{k,j_\mathrm{e},n} \\ &\qquad \leftarrow \varepsilon, \varepsilon, \cdots, \varepsilon \ \forall k, n \neq n_\mathrm{t}, \end{aligned} \tag{15}$$

$$\boldsymbol{w}_{i,n} \leftarrow [\rho, \rho, \cdots, \rho]^\mathrm{T} \ \forall i, n, \tag{16}$$

where ε is a machine epsilon. Similarly to Sect. 3.2, the other parameters are inherited from the previously performed ILRMA.

3.4. Time annotation with silent segment of the other sources (II)

The process (15) in Sect. 3.3 substitutes ε to the corresponding elements in $\boldsymbol{V}_n$ $(n \neq n_\mathrm{t})$. In addition to (15), we also propose to reset the other components in $\boldsymbol{V}_n$ as follows:

$$\begin{aligned} &v_{k,1,n}, v_{k,2,n}, \cdots, v_{k,j_\mathrm{s}-1,n} \\ &\qquad \leftarrow \alpha, \alpha, \cdots, \alpha \ \forall k, n \neq n_\mathrm{t}, \end{aligned} \tag{17}$$

$$\begin{aligned} &v_{k,j_\mathrm{e}+1,n}, v_{k,j_\mathrm{e}+2,n}, \cdots, v_{k,J,n} \\ &\qquad \leftarrow \alpha, \alpha, \cdots, \alpha \ \forall k, n \neq n_\mathrm{t}, \end{aligned} \tag{18}$$

$$v_{k,j,n_\mathrm{t}} \leftarrow \alpha \ \forall k, j, \tag{19}$$

where α is a random value that obeys the uniform distribution in the range $[1.0 \times 10^5, 1.1 \times 10^5]$, which is sufficiently larger than ε. Compared with the processing (15) and (16) (hereafter referred to as `time annotation (I)`), the processing (15)–(19) (hereafter referred to as `time annotation (II)`) provides more drastic change to the ILRMA optimization, resulting in more effective induction to escape from the bad local minimum.

Table 1 Sources obtained from SiSEC2011 dataset [14].

Mixture	Source signals
No. 1	dev1_female3_synthconv_130ms_5cm_sim_1
	dev1_female3_synthconv_130ms_5cm_sim_2
No. 2	dev1_male3_synthconv_130ms_5cm_sim_1
	dev1_male3_synthconv_130ms_5cm_sim_2
No. 3	dev1_male3_synthconv_130ms_5cm_sim_1
	dev1_female3_synthconv_130ms_5cm_sim_2

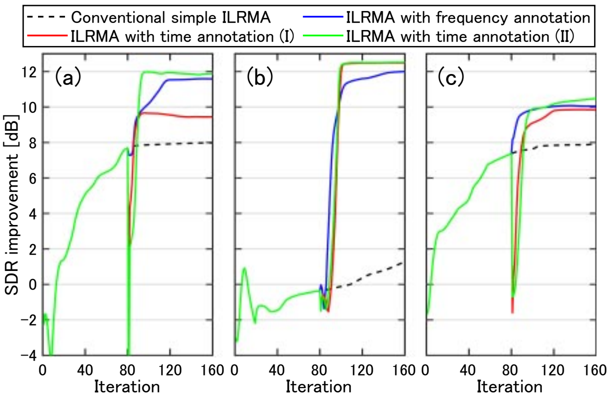

Fig. 3 SDR improvements of the proposed system with frequency or time annotation for mixtures (a) no. 1, (b) no. 2, and (c) no. 3.

4. Experiments

4.1. Conditions

We evaluated the performance of conventional simple ILRMA and the proposed interactive speech source separation system. In this experiment, we produced three mixture signals (nos. 1–3) as listed in Table 1, where the speech sources were obtained from SiSEC2011 dataset (see [14] for the detailed conditions). In short-time Fourier transform, we used 128-ms-long hamming window with half-overlap shifting. The number of bases in ILRMA was set to $L = 3$, and the improvement of source-to-distortion ratio (SDR) [15] was calculated.

In this experiment, the conventional ILRMA estimates $\boldsymbol{y}_{i,j}$ with 160 iterations (parameter updates). The proposed system returns temporarily estimated results to the user after 80 iterations of ILRMA and requires the frequency or time annotation at that time. After the annotation feedback and the processing described in Sect. 3, the proposed system continues ILRMA optimization with 80 iterations. The annotation is subjectively provided.

4.2. Results

Figure 3 shows behaviors of SDR improvement for the mixtures nos. 1–3 when the frequency or time annotation is fed back by the user at the 80th iteration. For all the cases, we can confirm that the proposed system outperforms the conventional ILRMA. In particular, the conventional ILRMA for the mixture no. 2 (Fig. 3(b)) cannot separate the sources because of encountering the block permutation problem

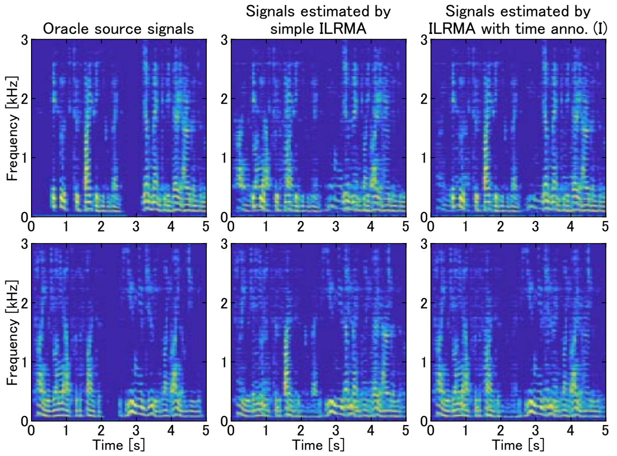

Fig. 4 Spectrograms of first (top) and second (bottom) speech sources for mixture no. 2.

during the parameter optimization. However, in the proposed system, the user annotation assists to escape from the bad local minimum and provides significant improvement for the correct estimation of the demixing matrix $\boldsymbol{W}_i$.

Regarding the type of annotations, we can see that both the frequency and time annotations effectively improve the separation accuracy in the proposed system. Since the demixing filters $\boldsymbol{w}_{i,n}$ are completely reset using random values ρ in the time annotations (I) and (II), the value of SDR improvement drops after the annotation feedback. In other words, this performance drop is necessary for escaping from the bad local minimum or solving the block permutation problem. Thus, it is important to sufficiently iterate ILRMA after the time annotation is fed back to the system (at least 40 iterations).

Figure 4 shows the spectrograms of oracle sources and estimated signals obtained by simple ILRMA and ILRMA with time annotation (I) for the mixture no. 2. The estimated signals obtained by simple ILRMA (the center column in Fig. 4) include the block permutation problem: the components between 0.4–1.7 kHz are swapped, whereas the proposed ILRMA (the right column in Fig. 4) correctly solves this misalignment.

This experiment only compares the performance at the 160th iteration in conventional ILRMA and the proposed system. Note that the interaction between ILRMA and the user can be iterated until the satisfactory separation is achieved in an actual application. The judgment can be subjectively taken by listening the estimated signals $\boldsymbol{y}_{i,j}$.

5. Conclusion

In this letter, we proposed a new ILRMA-based interactive speech source separation system, where frequency and time annotations were utilized. The experimental results show that these annotations significantly improve the separation performance by escaping from the bad local minima in ILRMA optimization.

Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Numbers 19K20306 and 19H01116.

References

[1] H. Sawada, N. Ono, H. Kameoka, D. Kitamura and H. Saruwatari, “A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF,” *APSIPA Trans. Signal Inf. Process.*, **8**(e12), 1–14 (2019).

[2] P. Comon, “Independent component analysis, a new concept?,” *Signal Process.*, **36**, 287–314 (1994).

[3] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, **22**, 21–34 (1998).

[4] 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).

[5] 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 Process.*, **12**, 530–538 (2004).

[6] T. Kim, H. T. Attias, S.-Y. Lee and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. Audio Speech Lang. Process.*, **15**, 70–79 (2007).

[7] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 189–192 (2011).

[8] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, **24**, 1626–1641 (2016).

[9] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, “Determined blind source separation with independent low-rank matrix analysis,” in *Audio Source Separation*, S. Makino, Ed. (Springer, Cham, 2018), Chap. 6, pp. 125–155.

[10] Y. Liang, S. M. Naqvi and J. A. Chambers, “Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm,” *Electron. Lett.*, **48**, 460–462 (2012).

[11] Y. Mitsui, D. Kitamura, N. Takamune, H. Saruwatari, Y. Takahashi and K. Kondo, “Independent low-rank matrix analysis based on parametric majorizaion-equalization algorithm,” *Proc. Comput. Adv. Multi-Sensor Adapt. Process.*, pp. 98–102 (2017).

[12] C. Févotte and J. Idier, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural Comput.*, **21**, 793–830 (2011).

[13] T. Ono, N. Ono and S. Sagayama, “User-guided independent vector analysis with source activity tuning,” *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2417–2420 (2012).

[14] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe and A. Benichoux, “The 2011 signal separation evaluation campaign (SiSEC2011): Audio source separation,” *Proc. Latent Variable Anal. Signal Separation*, pp. 414–422 (2012).

[15] E. Vincent, R. Gribonval and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio Speech Lang. Process.*, **14**, 1462–1469 (2006).