Permutation Alignment Based on MUSIC Spectrum Discrepancy for Blind Source Separation

Yuuki Tachioka
Denso IT Laboratory
Tokyo, Japan
ytachioka@d-itlab.co.jp

***Abstract*—Conventional time-frequency-domain blind source separation (BSS) requires permutation alignment of the sound sources. Permutation alignment methods can be classified into two types: those that use the direction of arrival (DOA) constraints and those that model the sound source characteristics instead of DOA constraints. Multi-channel non-negative matrix factorization (MNMF), which is based on the second type, is one of the most effective BSS methods. However, our experiments revealed that its permutation alignment sometimes fails due to the lack of a DOA constraint. We present a permutation alignment method based on the DOAs directly obtained from a spatial correlation matrix by using multiple signal classification (MUSIC) and that solves the permutation problems by minimizing the discrepancy of the MUSIC spectra, which belong to the same source, in the middle of the BSS algorithm. Our proposed method boosts the second type with a help of the DOA constraint and can be applied in a blind manner to both the mixing system approach, e.g., MNMF, and the demixing system approach, e.g., independent low-rank matrix analysis. Experiments showed that the proposed method is effective for both approaches.**

***Index Terms*—blind source separation, permutation alignment, direction of arrival estimation, multiple signal classification**

I. Introduction

Sound source separation extracts target sources from mixed signals. In particular, the blind source separation (BSS) approach does not need prior knowledge about the positions of sources or microphones and it is especially convenient and is robust to measurement errors. It is more effective to conduct time-frequency-domain BSS than time-domain BSS because the convolutive mixture in the time domain can be dealt with as an instantaneous mixture in the time-frequency domain [1].

The most widely used time-frequency-domain BSS has been independent component analysis [2] but it requires permutation alignment of sound sources after separation because the ambiguity of source index causes source permutations among frequency bins. The permutation alignment methods are classified into two types: ones using the direction of arrival (DOA) and ones that model the sound source characteristics. An example of the first type is DOA clustering [3], [4], which makes it easy to understand the reasons for separation failures.

There are numerous examples of the second type. Some place certain assumptions on the properties of the sources, for instance, on the correlation of the amplitude envelopes among neighboring frequency bins [1]. Others model the spectrum envelopes [5], [6]. In particular, a method called independent vector analysis (IVA) [7]–[9] assumes that the target sources across different frequency bins activate simultaneously.

Advanced forms of IVA, multi-channel non-negative matrix factorization (MNMF) [10] and its rank-1 relaxation, independent low-rank matrix analysis (ILRMA) [11], have been developed. MNMF models the spectrum envelope by using low-rank bases and activations without any explicit DOA constraints. It can perform accurate separation, but sometimes fails. We believe that the permutation alignments based on DOAs are necessary for both MNMF and ILRMA because BSS performance depends heavily on the estimation accuracy of the spatial correlation matrices [12], [13]. The experiments described in Section IV reveal that one of the causes of separation failure in MNMF is insufficient permutation alignment.

This paper proposes a permutation alignment method based on DOA estimated by multiple signal classification (MUSIC) [14]. Conventional DOA clustering [3], [4] relies time-frequency-bin-wise DOA for solving the permutation problems, but bin-wise DOA is poor at estimating accurate DOAs [15]. In fact, in the field of DOA estimation, instead of bin-wise DOA, generalized cross correlation with phase transform [16] or MUSIC-based approaches are the most widely used. Mitianoudis [17] proposed a permutation alignment method by applying the MUSIC algorithm to the demixed signals. In contrast, our method directly applies the MUSIC algorithm to the estimated spatial correlation matrix instead of the demixed signals. To detect the source permutations, we use different metrics to evaluate the discrepancies between the MUSIC spectra. Minimizing the discrepancy leads to the permutation alignments. Our algorithm does not require any prior knowledge about the source or microphone setups and can be applied in a blind manner to both the mixing and demixing system approach. Furthermore, it can be used in underdetermined situations where the number of microphones is less than that of the sound sources due to the sparsity in the time-frequency domain. It aligns the permutation not at the end of the BSS algorithm, as is done in [3], [4], [17], [18], but in the middle of the BSS algorithm, as in permutation-free clustering [15]. The permutations can be aligned along with the BSS algorithms. It improves the performance of BSS with a help of DOA constraints, which is a combination of the above-mentioned two types of permutation alignment methods.

The remainder of this paper is organized as follows. Section II describes the BSS methods, MNMF and ILRMA. Section

978-9-0827-9703-9/19/$31.00

III describes the permutation alignment algorithm based on metrics for evaluating MUSIC spectrum discrepancies. BSS experiments using four different musical pieces examining the effectiveness of the algorithm are described. Section IV compares the BSS performance of MNMF and ILRMA with and without the proposed permutation alignment.

II. BSS METHODS

A. MNMF (mixing system approach)

An observation vector at frequency bin i ($1 \leq i \leq I$) and time frame j ($1 \leq j \leq J$), $\boldsymbol{x}_{i,j}$, is expressed as $[x_1, \ldots, x_m, \ldots, x_M]^\top_{i,j}$, where $^\top$ denotes the transpose and x_m is a complex short-time Fourier transform spectrum observed at the $m(1 \leq m \leq M)$-th microphone. Thus, the i,j element of the tensor containing signals' statistics $\boldsymbol{X} \in (\mathbb{C}^{M\times M})^{I\times J}$ is represented as $\boldsymbol{X}_{i,j} = \boldsymbol{x}_{i,j}\boldsymbol{x}^{\mathrm{H}}_{i,j}$ where $^{\mathrm{H}}$ is the Hermitian transpose. $\boldsymbol{X}$ is a hierarchical matrix whose elements $\boldsymbol{X}_{i,j} \in \mathbb{C}^{M\times M}$ are semi-positive-definite Hermitian matrices. $\boldsymbol{X}$ can be reconstructed as $\hat{\boldsymbol{X}}$ by factorized four matrices:

$$\boldsymbol{X} \cong \hat{\boldsymbol{X}} = [(\boldsymbol{H}\boldsymbol{Z}) \circ \boldsymbol{T}]\,\boldsymbol{V}, \tag{1}$$

where $\circ$ is the Hadamard product. $\boldsymbol{H} \in (\mathbb{C}^{M\times M})^{I\times L}$ is a spatial correlation matrix that indicates the spatial correlation of L sources. The bases matrix $\boldsymbol{T} \in \mathbb{R}^{I\times K}$ is composed of K bases and $\boldsymbol{V} \in \mathbb{R}^{K\times J}$ is the activation of each basis. $\boldsymbol{Z} \in \mathbb{R}^{L\times K}$ is a matrix relating the spatial correlation to each basis. The right-hand side of Eq. (1) is $\hat{\boldsymbol{X}}_{i,j} = \sum_{k,l} \boldsymbol{H}_{i,l} z_{l,k} t_{i,k} v_{k,j}$. The four matrices above are updated to minimize the multi-channel Itakura-Saito divergence between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$:

$$\underset{\boldsymbol{H},\boldsymbol{Z},\boldsymbol{T},\boldsymbol{V}}{\arg\min} \sum_{i,j} \left[\mathrm{tr}(\boldsymbol{X}_{i,j}\hat{\boldsymbol{X}}^{-1}_{i,j}) - \log\det \boldsymbol{X}_{i,j}\hat{\boldsymbol{X}}^{-1}_{i,j} - M\right], \tag{2}$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix.

B. ILRMA (demixing system approach)

ILRMA estimates a demixing matrix directly instead of a spatial correlation matrix. Assuming a rank-1 approximation of the mixed system, the observation $\boldsymbol{x}_{i,j}$ and source image $\boldsymbol{s}_{i,j} \in \mathbb{C}^L$ are related through the demixing matrix $\boldsymbol{W}_i \in \mathbb{C}^{L\times M}$ as follows: $\boldsymbol{s}_{i,j} = \boldsymbol{W}_i\boldsymbol{x}_{i,j} = [\boldsymbol{w}_{i,1}, ..., \boldsymbol{w}_{i,L}]^\top \boldsymbol{x}_{i,j}$, where $\boldsymbol{w}_{i,l} (\in \mathbb{C}^M)$ is the l-th row vector of $\boldsymbol{W}_i$. The mixing matrix $\boldsymbol{A}_i \in \mathbb{C}^{M\times L}$ is the inverse of the demixing matrix $\boldsymbol{W}_i^{-1}$ (in the case of $M \neq L$, the Moore-Penrose pseudo inverse). Here, $\boldsymbol{x}_{i,j} = \boldsymbol{A}_i\boldsymbol{s}_{i,j} = [\boldsymbol{a}_{i,1}, ..., \boldsymbol{a}_{i,L}]\boldsymbol{s}_{i,j}$ ($\boldsymbol{a}_{i,l} \in \mathbb{C}^M$) holds for mixing matrices. In a rank-1 system, the spatial correlation matrix for the l-th source can be represented by the correlation of the l-th column vector of $\boldsymbol{A}_i$: $\boldsymbol{H}_{i,l} = \boldsymbol{a}_{i,l}\boldsymbol{a}^{\mathrm{H}}_{i,l}$.

III. PERMUTATION ALIGNMENT BY USING MUSIC SPECTRA

A. MUSIC spectrum

The MUSIC spectrum of the l-th source at the i-th frequency bin can be obtained from the spatial correlation matrix, $\boldsymbol{H}_{i,l} \in \mathbb{C}^{M\times M}$ by performing an eigenvalue decomposition:

$$\boldsymbol{H}_{i,l} = \boldsymbol{B}_{i,l}\boldsymbol{G}_{i,l}\boldsymbol{B}^{-1}_{i,l} = \left[\boldsymbol{B}^s_{i,l}\boldsymbol{B}^n_{i,l}\right]\boldsymbol{G}_{i,l}\left[\boldsymbol{B}^s_{i,l}\boldsymbol{B}^n_{i,l}\right]^{-1}. \tag{3}$$

Here, $\boldsymbol{G}_{i,l} \in \mathbb{R}^{M\times M}$ is a diagonal matrix whose elements are real-valued eigenvalues (sorted in descending order) because $\boldsymbol{H}_{i,l}$ is an Hermitian matrix. $\boldsymbol{B}_{i,l} \in \mathbb{C}^{M\times M}$ is composed of M-dimensional column eigenvectors, which are orthogonal to each other. $\boldsymbol{B}_{i,l}$ is composed of two sub-matrices $\boldsymbol{B}^s_{i,l} \in \mathbb{C}^{M\times 1}$ and $\boldsymbol{B}^n_{i,l} \in \mathbb{C}^{M\times(M-1)}$, which span the signal and noise subspaces, respectively. Eigenvectors that correspond to eigenvalues other than the maximum eigenvalue span noise subspaces because $\boldsymbol{H}_{i,l}$ corresponds to the single l-th separated source. We assume that, when $\boldsymbol{H}$ is properly obtained, the target source has the largest energy among multiple sources. In the first iterations of BSS algorithm, $\boldsymbol{H}$ cannot be properly obtained, thus we start permutation alignments after some iterations.

The DOA range is discretized in S steps. Under the plane wave assumption for a microphone array with a microphone spacing δ[1], the steering vector $\boldsymbol{a}^P_i(\theta_s) = [a^P_1, ..., a^P_m, ..., a^P_M]^\top$ corresponding to each direction $\theta_s (s \in \{1, ..., S\})$ is

$$a^P_m = \exp\left[\jmath\left(m - \frac{M+1}{2}\right)\frac{2\pi\delta}{c}(i-1)\frac{f_s}{I}\sin\theta_s\right], \tag{4}$$

where $\jmath$ is the imaginary unit, f_s is a sampling frequency [Hz], and c is a sound velocity. Here, θ_s is a candidate of the source directions, which we arbitrary assume.

The MUSIC spectrum with a weight of the maximum eigenvalue $\boldsymbol{G}_{i,l}(1,1)$ is expressed as

$$S_{i,l}(\theta_s) = \frac{\sqrt{\boldsymbol{G}_{i,l}(1,1)}}{(\boldsymbol{a}^P_i(\theta_s))^{\mathrm{H}}\boldsymbol{B}^n_{i,l}(\boldsymbol{B}^n_{i,l})^{\mathrm{H}}\boldsymbol{a}^P_i(\theta_s)}. \tag{5}$$

For MNMF, which can obtain the power spectrum of each source, the power-normalized MUSIC spectrum, which is the MUSIC spectrum divided by the power of each source, can also be used:

$$S'_{i,l}(\theta_s) = \frac{S_{i,l}(\theta_s)}{\sum_j \sum_k z_{l,k}t_{i,k}v_{k,j}}. \tag{6}$$

To obtain the source DOA, the MUSIC spectra are summed over reliable frequency bins for the l-th source:

$$S^{all}_l(\theta_s) = \sum_{i\in[\lfloor f_{min}/f_sI\rfloor,\lceil f_{max}/f_sI\rceil]} S_{i,l}(\theta_s), \tag{7}$$

where the frequency band in $[f_{min}, f_{max}]$[Hz] are considered reliable.

B. Permutation evaluation based on MUSIC spectrum discrepancy

The extent of source permutations can be evaluated by comparing MUSIC spectra from L sound sources after normalizing them to set their sums to unity.

$$S_{i,l}(\theta_s) \leftarrow \frac{S_{i,l}(\theta_s)}{\sum_s S_{i,l}(\theta_s)},\ S^{all}_l(\theta_s) \leftarrow \frac{S^{all}_l(\theta_s)}{\sum_s S^{all}_l(\theta_s)}. \tag{8}$$

[1] δ can be unknown because when the actual microphone spacing is $\delta' (\neq \delta)$, a^P_m is just powered by δ'/δ, $(a^P_m)^{\delta'/\delta}$. The envelope of the MUSIC spectrum is constant, despite that the MUSIC spectrum expands or shrinks. For non-linear array, estimated source direction is unreliable but a distinction between different sources can be done. Thus, this assumption can be used in the BSS framework.

TABLE I
METRICS FOR EVALUATING DISCREPANCY BETWEEN TWO MUSIC SPECTRA $\boldsymbol{S}_1$ AND $\boldsymbol{S}_2$. FOR DPD, γ IS SET TO 0.2.

Metrics	$d_\cdot(\boldsymbol{S}_1, \boldsymbol{S}_2)$
d_{PK}	$\lvert\arg\max_\theta S_1(\theta) - \arg\max_\theta S_2(\theta)\rvert$
d_{CS}	$-\frac{\sum_\theta S_1(\theta)S_2(\theta)}{\sqrt{\sum_\theta S_1^2(\theta)}\sqrt{\sum_\theta S_2^2(\theta)}}$
d_{SE}	$\sum_\theta \lvert S_1(\theta) - S_2(\theta)\rvert^2$
d_{OR}	$-\sum_\theta \min\left(S_1(\theta), S_2(\theta)\right)$
d_{KLD}	$\sum_\theta S_1(\theta)\log\left(\frac{S_1(\theta)}{S_2(\theta)}\right)$
d_{DPD}	$\sum_\theta \Big[\frac{1}{\gamma}\left(S_1(\theta)^\gamma - S_2(\theta)^\gamma\right)$ $-\frac{1}{1+\gamma}\left(S_1(\theta)^{1+\gamma} - S_2(\theta)^{1+\gamma}\right)\Big]$

MUSIC spectra $\boldsymbol{S}$ denotes a vector of $S(\theta_s)$ for all θ_ss, i.e., $\boldsymbol{S} = [S(\theta_1), ..., S(\theta_S)]^\top$. Table I lists metrics that evaluate the discrepancy between two MUSIC spectra. Once the metric has been selected, the total discrepancy is calculated by

$$D_\cdot^1 = \frac{1}{2}\sum_i \sum_l \sum_{l' \neq l} d_\cdot\left(\boldsymbol{S}_{i,l}, \boldsymbol{S}_{i,l'}\right), \qquad D_\cdot^2 = \frac{1}{2}\sum_l \sum_{l' \neq l} d_\cdot\left(\boldsymbol{S}_l^{all}, \boldsymbol{S}_{l'}^{all}\right), \tag{9}$$

where dot "$\cdot$" of $D_\cdot^{\{1,2\}}$ and $d_\cdot$ means an arbitrary metric.

The six metrics have the following properties. Peak, d_{PK} evaluates the discrepancy on the basis of the difference between the peak indexes of MUSIC spectra. Cosine similarity, d_{CS}, is a (minus) cosine similarity of them. Square error, d_{SE}, calculates the square errors between them. Overlapping region, d_{OR}, is the proportion of the overlapped region of them. Kullback-Leibler divergence (KLD), d_{KLD}, and density power divergence (DPD), d_{DPD}, calculate their KLD and DPD, which is an efficient and robust divergence [19], respectively.

C. Permutation detection and alignment

The occurrence of the source permutations can be detected by comparing the sum of source MUSIC spectrum $\boldsymbol{S}_l^{all}$ with that at each frequency bin, $\boldsymbol{S}_{i,l}$, on the basis of the above metrics $d_\cdot$. If the permutations of the index i are aligned, the inequation

$$d_\cdot(\boldsymbol{S}_l^{all}, \boldsymbol{S}_{i,l}) \leq d_\cdot(\boldsymbol{S}_l^{all}, \boldsymbol{S}_{i,l'}) \text{ (for } {}^\forall l' \neq l \text{ and } {}^\forall i), \tag{10}$$

holds. For example, as in Fig. 1, the l-th source MUSIC spectrum $\boldsymbol{S}_l^{all}$ is compared with that at the certain frequency bin (here, $i = 10$), $\boldsymbol{S}_{10,l}$. In the case of the peak metric d_{PK}, the peak of $\boldsymbol{S}_1^{all}$, ϕ_1, is nearer to the peak of $\boldsymbol{S}_{10,2}$, $\phi_{10,2}$, than that of $\boldsymbol{S}_{10,1}$, $\phi_{10,1}$. The inequation, $d_{PK}(\boldsymbol{S}_1^{all}, \boldsymbol{S}_{10,1}) = |\phi_1 - \phi_{10,1}| > d_{PK}(\boldsymbol{S}_1^{all}, \boldsymbol{S}_{10,2}) = |\phi_1 - \phi_{10,2}|$, holds and this indicates that a permutation of the source index occurs at $i = 10$. For other metrics, this inequation 10 holds.

As a result, the permutations can be aligned as

$$\arg\min_{\Pi_i} \sum_l d_\cdot\left(\boldsymbol{S}_l^{all}, \boldsymbol{S}_{i,\Pi_i(l)}\right), \tag{11}$$

where Π_i is a permutation of L sound sources at the i-th frequency bin as $\Pi_i = \{1, 2, ..., L\} \rightarrow \{1, 2, ..., L\}$. The above

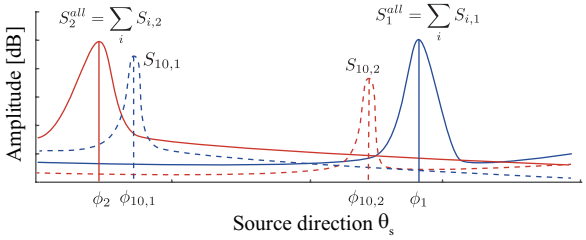
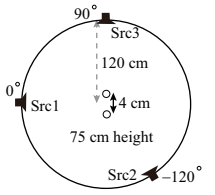


Fig. 1. Source MUSIC spectrum summed over all reliable frequency bins, $S_{\{1,2\}}^{all}$, and that at the frequency bin $i(=10)$, $S_{i,\{1,2\}}$.



Sampling frequency f_s	16 kHz
Frame size and shift	1024 and 256
# bases K	30
# iterations of MNMF	500
# iterations of ILRMA	200

Fig. 2. Source and microphone positions and BSS settings.

example compares the peak of the MUSIC spectrum of the l-th source, ϕ_l, with that for the MUSIC spectrum at each frequency bin, $\phi_{i,l}$, and permutation is aligned to make these peaks nearer for all is as (11). After the permutation Π has been obtained, $\boldsymbol{H}$ or $\boldsymbol{w}$ is aligned as

$$\boldsymbol{H}_{i,l} \leftarrow \boldsymbol{H}_{i,\Pi_i(l)} \quad \text{(for mixing system)}, \tag{12}$$

$$\boldsymbol{w}_{i,l} \leftarrow \boldsymbol{w}_{i,\Pi_i(l)} \quad \text{(for demixing system)}. \tag{13}$$

D. Schedule of permutation alignment

BSS algorithms begin with no prior knowledge about the source positions; thus, it is impossible to align permutations in the initial iterations. The permutation alignment starts after some iterations, i.e., in the middle of the BSS algorithm.

We devised two types of scheduling for starting the permutation alignment: fixed iteration count and fixed threshold. The first schedule starts the permutation alignments after a fixed number of iterations. Preliminary experiments show that it is better to align permutations at several times with some intervals than at a single time. The second schedule uses a fixed threshold, wherein if the total discrepancy metrics D^1 or D^2 in (9) exceed certain fixed threshold, the permutation alignment starts. For permutation alignment, stable and distinct MUSIC spectra are needed. Initially, the MUSIC spectra of all sources overlap and all total discrepancy metrics take their minimum values. The results of the experiments depicted in Fig. 4 show that all metrics almost monotonically increase with the number of iterations as MUSIC spectra become separable.

IV. BSS EXPERIMENTS

A. Experimental setups

The effectiveness of the proposed permutation alignment was assessed on BSS of musical pieces[2] in underdetermined ($M < L$) settings. Fig. 2 shows the source and microphone positions along with other parameters. Mixed signals were picked up by two microphones 4 cm apart ($M = 2$). These

[2]Musical pieces are publicly available at http://sisec2010.wiki.irisa.fr

TABLE II
MUSICAL PIECES COMPOSED OF THREE SOURCES.

ID	Title	Snip [s]	Source
1	Ultimate Nz Tour	43-61	guitar, synth, drums
2	The Ones We Love	69-94	drums, guitar, vocals
3	Remember the Name	54-78	violin+synth, vocals, drums
4	Roads	85-99	piano, vocals, ambient

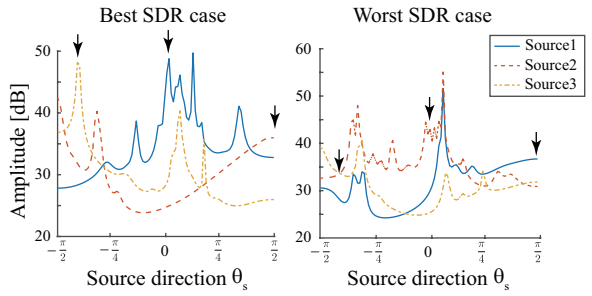


Fig. 3. Source MUSIC spectra of $\boldsymbol{H}$ corresponding to the best and worst SDRs among ten trials with different random initializations.

settings were similar to those of [12]. Table II shows the musical pieces, which was composed of three sources ($L = 3$) convolved with the impulse responses from three different directions, 0, -120, and 90 [deg] ($T_{60} = 100$ [ms]).

For MNMF, $\boldsymbol{T}$, $\boldsymbol{V}$, and $\boldsymbol{Z}$ were randomly initialized by ten different random-value seeds. As in [10], [12], $\boldsymbol{H}_{i,l}$ was initialized to a unit matrix with $\boldsymbol{H}_{i,l}$ being constant in the first 20 iterations. To obtain demixed signals, multi-channel Wiener filter was applied. For ILRMA, $\boldsymbol{W}_i$ was initialized as $[1\ 0; 0\ 1; 0.5\ 0.5]$. The permutation was aligned for the fixed count schedule at iterations 40, 45, 50, 55, and 60 for MNMF and at 70, 75, and 80 for ILRMA. The permutation alignment started for the fixed threshold schedule if D^1 exceeded the half of the maximum value (i.e., $\bar{D}^1 > 0.5$ in Fig. 4). After that, alignments performed at five times with five intervals. The reliable frequency band was from $f_{min} = 500$ to $f_{max} = \frac{c}{2\delta} = 4250$ [Hz]. The separation performance averaged over ten trials was evaluated in terms of the signal-to-noise ratio (SDR) [dB] [20].

B. MUSIC spectra for the best and worst SDR cases

The separation performance significantly varied with the random initial values [12]. Fig. 3 shows the source MUSIC spectra, S_l^{all}, derived from the final $\boldsymbol{H}$ for the best SDR case ($= 14.63$ [dB]) and the worst SDR case ($= 1.63$ [dB]) where three downward arrows indicate the true source directions. The peaks overlapped more in the worst case. This shows that the permutation alignment is problematic even for MNMF.

C. Permutation metrics

Fig. 4 shows the normalized $D_\cdot^1$ in (9) with the number of iterations for two different musical pieces. Although the speed of convergence differed between the pieces, the metrics had almost converged by 100 iterations. All metrics almost monotonically increased [3]. For the case of $D_\cdot^2$, the trends were similar.

[3]The metrics did not change in the first 20 iterations because $\boldsymbol{H}$ was not updated in the first 20.

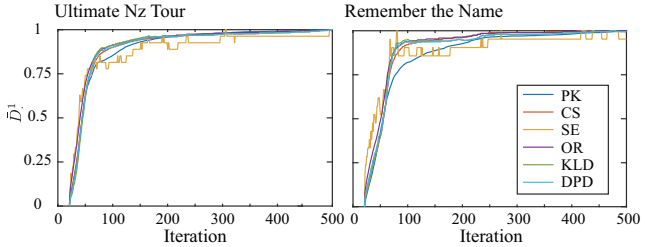


Fig. 4. Normalized total discrepancy $\bar{D}^1$ and the number of iterations where $\bar{D}_\cdot^1$ were normalized as $(D_\cdot^1 - \min(D_\cdot^1))/(\max(D_\cdot^1) - \min(D_\cdot^1))$.

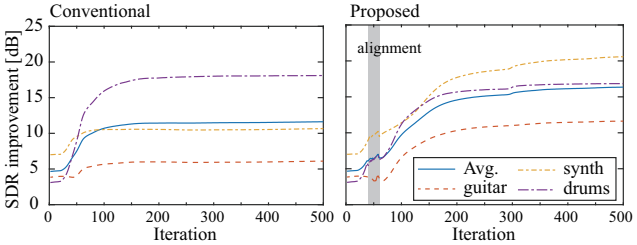


Fig. 5. SDR improvement [dB] at each iteration comparing the proposed method with the conventional MNMF.

D. BSS results (MNMF)

Fig. 5 shows the SDR improvement at each iteration of MNMF where alignment in the middle of MNMF iterations improved the SDR. Fig. 6 shows the average SDR improvement [dB] in the case of the fixed count schedule. "Based on Eq. (5)" in the figure used the MUSIC spectra in (5) and "Based on Eq. (6)" used the power-normalized MUSIC spectra in (6). The performance improved for all metrics. Power normalization slightly improved the performance. In average, "PK" and "CS" criterion achieved the best separation performance.

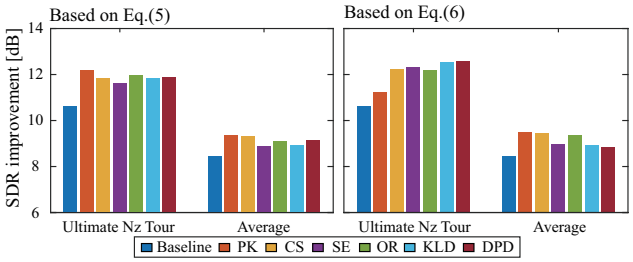


Fig. 6. Average SDR improvement [dB] of MNMF for "Ultimate Nz Tour" and all four pieces with fixed iteration count schedule.

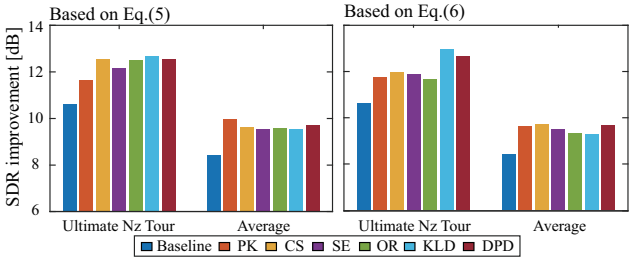


Fig. 7. Average SDR improvement [dB] of MNMF with fixed threshold schedule.

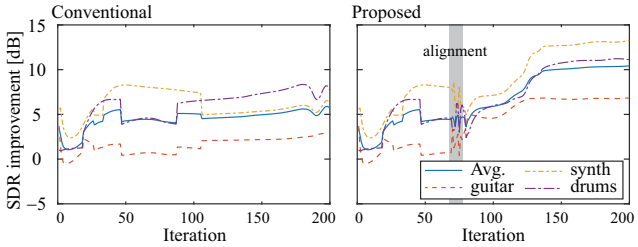


Fig. 8. SDR improvement [dB] at each iteration comparing the proposed method with the conventional ILRMA.

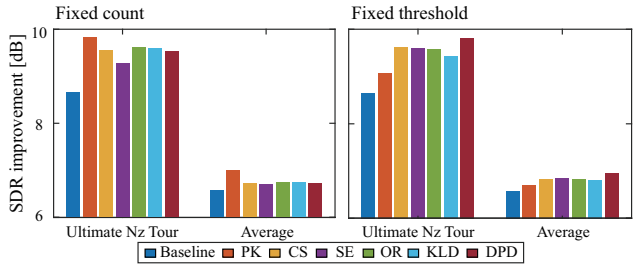


Fig. 9. Average SDR improvement [dB] of ILRMA for each piece with fixed iteration count and fixed threshold schedule.

Fig. 7 shows the average SDR improvement for the fixed threshold schedule. This schedule was better than the fixed count schedule because it can be adjusted to the speed of convergence, which was different for each piece.

E. BSS results (ILRMA)

Fig. 8 shows the SDR improvement at each iteration of ILRMA. Although the separation performance of the conventional ILRMA did not improve after 100 iterations, the proposed permutation alignment in the middle of iterations also did improve the SDR. Fig. 9 shows the average SDR improvement. The trends were similar to those of MNMF and the proposed method was also effective on ILRMA.

V. Conclusion

We proposed a permutation alignment method for BSS algorithms. The method uses the discrepancy of the source MUSIC spectra obtained from a spatial correlation matrix as a cue of permutation alignment directly. BSS experiments on musical pieces show that permutation alignments inherent in MNMF were insufficient for the poor separation case. The proposed method improved the BSS performance by aligning permutations in the middle of the iterations of the BSS algorithms. The proposed method was effective in the mixing system approach, MNMF, and in the demixing system approach, ILRMA. In addition we devised fixed count and fixed threshold schedule. Although both two schedules were effective, the latter was better than the former because the alignment schedule can be adjusted to the speed of convergence.

References

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

[2] P. Smaragdis, "Blind separation of convolved mixture in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[3] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 516–527, 2011.

[4] S. Araki, T. Nakatani, and H. Sawada, "Sparse source separation based on simultaneous clustering of source locational and spectral features," *Acoustical Science & Technology*, vol. 32, no. 4, pp. 161–164, 2011.

[5] R. Mazur and A. Mertins, "An approach for solving the permutation problem of convolutive blind source separation based on statistical signal models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 117–126, 2009.

[6] S. Saito, K. Oishi, and T. Furukawa, "Convolutive blind source separation using an iterative least-squares algorithm for non-orthogonal approximate joint diagonalization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2434–2448, 2015.

[7] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proceedings of International Workshop on Independent Component Analysis and Source Separation (ICA)*, 2006, pp. 601–608.

[8] T. Kim, T. Eltoft, and T.W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proceedings of International Workshop on Independent Component Analysis and Source Separation (ICA)*, 2006, pp. 165–172.

[9] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[10] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[11] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[12] I. Miura, Y. Tachioka, T. Narita, J. Ishii, F. Yoshiyama, S. Uenohara, and K. Furuya, "Analysis of initial-value dependency in multichannel nonnegative matrix factorization for blind source separation and speech recognition," *IEICE Transactions on Information and Systems*, vol. J100-D, no. 3, pp. 376–384, 2017.

[13] T. Uramoto, Y. Tachioka, T. Narita, I. Miura, S. Uenohara, and K. Furuya, "Sequential initialization of multichannel nonnegative matrix factorization for sound source separation," in *Proceedings of IEEE 6th Global Conference on Consumer Electronics (GCCE 2017)*, 2017, DOI: 10.1109/GCCE.2017.8229207.

[14] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas Propagation*, vol. 34, pp. 276–280, 1986.

[15] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proceedings of ICASSP*, 2013, pp. 3238–3242.

[16] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[17] N. Mitianoudis and M. Davies, "Permutation alignment for frequency domain ICA using subspace beamforming methods," in *Proceedings of International Workshop on Independent Component Analysis and Source Separation (ICA)*, 2004, pp. 127–132.

[18] K. Zhang and L. Chan, "Convolutive blind source separation by efficient blind deconvolution and minimal filter distortion," *Neurocomputing*, vol. 73, no. 1315, pp. 2580–2588, 2010.

[19] A. Basu, I.R. Harris, N.L. Hjort, and M.C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, 2006.