

ユーザーからの補助情報を用いるインタラクティブ音源分離システム*

☆中野将生, 北村大地 (香川高専)

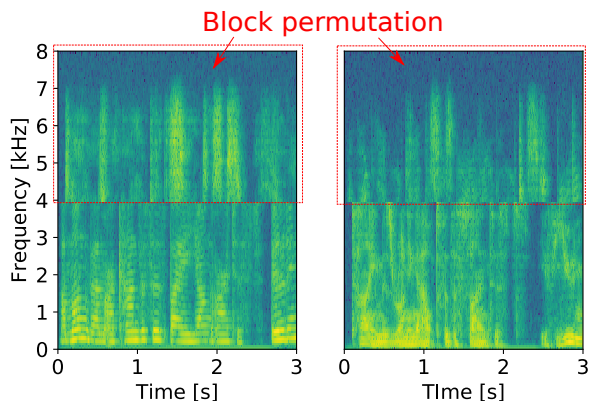


Fig. 1 Example of block permutation.

1 はじめに

ブラインド音源分離 (blind source separation: BSS) とは、複数の音源が混合された観測信号から混合系に関する事前情報を用いず混合前の音源を推定する技術である。観測チャンネル数が音源数以上となる優決定条件では、音源信号間の独立性を利用して音源分離を行う独立成分分析 (Independent component analysis: ICA) [1] に基づく手法が盛んに研究されている。特に、音響信号の混合時に生じる残響の影響を考慮するため、時間周波数領域での線形時不変な瞬時混合を仮定して周波数ビン毎に ICA を適用する周波数領域 ICA (frequency-domain ICA: FDICA) [2] が提案されている。

FDICA では、周波数毎に独立な ICA による推定が行われることから、分離信号の順番が周波数毎にばらばらになってしまう問題がある。これはパーミュテーション問題 [3] と呼ばれ、これまでに様々なパーミュテーション解決法が提案されてきた [3–5]。近年では、パーミュテーション問題を極力起こさずに音源分離を行う手法として、同一音源の分散 (パワー) が全周波数ビンで共起性を持つことを仮定する独立ベクトル分析 (independent vector analysis: IVA) [6, 7] や、同一音源の時間周波数の共起性が低ランク構造を持つことを仮定する独立低ランク行列分析 (independent low-rank matrix analysis: ILRMA) [8, 9] が提案されている。特に ILRMA は、音源の時間周波数構造の低ランク性仮定が良く合致する音楽信号において、頑健かつ高精度な BSS を実現しているが、音声信号の分離問題においては IVA と同程度の性能となる場合も多く [10]、実用においては課題が残る。

IVA や ILRMA が音声信号の BSS において分離に失敗する主な原因は、ブロックパーミュテーション問題と呼ばれる現象が起こることである [6, 11, 12]。これは、Fig. 1 のように、隣接する周波数ビン間でのパーミュテーション問題が解決されているにもかかわらず、まとまった周波数帯域がブロックとしてパーミュテーション不整合を起こす問題である。

本稿では、ILRMA に基づく BSS に対してユーザーとのインタラクションを導入することで、高精度な音源分離を達成するシステムを新たに提案する。具体的には、ILRMA の最適化の途中で現状の分離結果をユーザーに提示し、ブロックパーミュテーション問題が生じていた場合はユーザーによるアノテーション情報を利用することで、パーミュテーション不整合を修正しつつ最適化を続けるアルゴリズムを構築する。

2 従来手法

2.1 定式化

音源数と観測チャンネル数をそれぞれ N 及び M とし、各時間周波数における音源信号、観測信号及び分離信号をそれぞれ

$$\mathbf{s}_{ij} = (s_{ij,1} \cdots s_{ij,N})^T \in \mathbb{C}^{N \times 1} \quad (1)$$

$$\mathbf{x}_{ij} = (x_{ij,1} \cdots x_{ij,M})^T \in \mathbb{C}^{M \times 1} \quad (2)$$

$$\mathbf{y}_{ij} = (y_{ij,1} \cdots y_{ij,N})^T \in \mathbb{C}^{N \times 1} \quad (3)$$

と表す。ここで、 $i = 1, \dots, I$ は周波数インデックス、 $j = 1, \dots, J$ は時間インデックス、 $n = 1, \dots, N$ は音源インデックス、 $m = 1, \dots, M$ はチャンネルインデックスを示し、 T は転置を表す。また、各信号の複素スペクトログラム行列を $\mathbf{S}_n \in \mathbb{C}^{I \times J}$, $\mathbf{X}_m \in \mathbb{C}^{I \times J}$, $\mathbf{Y}_n \in \mathbb{C}^{I \times J}$ で表す。これらの行列の要素はそれぞれ $s_{ij,n}$, $x_{ij,m}$, $y_{ij,n}$ である。混合系が線形時不変であり、時間周波数領域での複素瞬時混合で表現できると仮定すると、周波数毎の時不変な複素混合行列 $\mathbf{A}_i = (\mathbf{a}_{i,n} \cdots \mathbf{a}_{i,N})^T \in \mathbb{C}^{M \times N}$ ($\mathbf{a}_{i,n} = (a_{i,n1} \cdots a_{i,nM})$ は各音源のステアリングベクトル) が定義でき、観測信号を次式で表現できる。

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{s}_{ij} \quad (4)$$

この混合モデルは、時不変混合系の残響時間が短時間フーリエ変換 (short time fourier transform: STFT) の窓長よりも十分に短い場合に成立する。この時、 $M = N$ かつ $\mathbf{A}_i$ が正則であれば、分離ベクトル $\mathbf{w}_{i,n} = (w_{i,n1} \cdots w_{i,nM})^T$ で構成される分離行列 $\mathbf{A}_i^{-1} \approx \mathbf{W}_i = (\mathbf{w}_{i,1} \cdots \mathbf{w}_{i,N})^H \in \mathbb{C}^{N \times M}$ が存在し、分離信号は次式で与えられる。

$$\mathbf{y}_{ij} = \mathbf{W}_i \mathbf{x}_{ij} \quad (5)$$

ここで、 H はエルミート転置を表す。線形時不変を仮定する優決定条件下での BSS では、この分離行列 $\mathbf{W}_i$ を求めることが目標である。本稿では、観測チャンネル数と音源数が等しいと仮定 ($M = N$) する。

2.2 ILRMA の生成モデルと更新式

FDICA におけるパーミュテーション問題を避けるため、IVA では同一音源の全周波数ビンのパワーの共起性を仮定しつつ、複数音源間の独立性を最大化している。このような音源モデル (同一音源の時間周波数

*Interactive audio source separation informed by user annotation by Masaki Nakano and Daichi Kitamura (NIT Kagawa).

構造に関する仮定)をより一般化するため, IVA の音源モデルに非負値行列因子分解 (nonnegative matrix factorization: NMF) [13] を導入した手法が ILRMA [8, 9] である.

ILRMA では次式の複素ガウス分布を音源信号の生成モデルとして仮定する.

$$\begin{aligned} p(\mathbf{Y}_1, \dots, \mathbf{Y}_n) &= \prod_n p(\mathbf{Y}_n) \\ &= \prod_{n,i,j} p(y_{ij,n}) \\ &= \prod_{n,i,j} \frac{1}{\pi r_{ij,n}} \exp\left(-\frac{|y_{ij,n}|^2}{r_{ij,n}}\right) \end{aligned} \quad (6)$$

ここで, $r_{ij,n}$ は音源 n の時間周波数毎のパワーであり, $r_{ij,n} = \mathbb{E}[|y_{ij,n}|^2]$ である. さらに, $r_{ij,n}$ を要素に持つ時間周波数行列を $\mathbf{R}_n \in \mathbb{R}_{\geq 0}^{I \times J}$ とおくと, この分散行列が次式のように NMF でモデル化される低ランク構造を持っていることを仮定する.

$$\mathbf{R}_n = \mathbf{T}_n \mathbf{V}_n \quad (7)$$

$$r_{ij,n} = \sum_k t_{ik,n} v_{kj,n} \quad (8)$$

ここで, $\mathbf{T}_n \in \mathbb{R}_{\geq 0}^{I \times K}$ 及び $\mathbf{V}_n \in \mathbb{R}_{\geq 0}^{K \times J}$ はそれぞれ音源 n の分散行列をモデル化する非負行列 (基底行列及びアクティベーション行列) であり, $t_{ik,n}$ 及び $v_{kj,n}$ はこれらの要素である. また, $k = 1, \dots, K$ は基底行列 $\mathbf{T}_n$ 中の基底ベクトルのインデックスである. 従って ILRMA では, 同一音源の分散の時間周波数構造が高々 K 個の基底で表される低ランク構造を持つことが仮定されている.

ILRMA では, 式 (6) の生成モデルに基づいて, 変数 $\mathbf{W}_i$, $\mathbf{T}_n$, 及び $\mathbf{V}_n$ の最尤推定を行う. 式 (6) の負対数尤度は次式となる.

$$\mathcal{L} = -2J \sum_i \log |\det \mathbf{W}_i| + \sum_{i,j,n} \left(\frac{|y_{ij,n}|^2}{r_{ij,n}} + \log r_{ij,n} \right) \quad (9)$$

ILRMA では, 分離信号のパワースペクトログラム $|\mathbf{Y}_n|^2$ (ここで, 行列に対する演算 $|\cdot|^2$ は, 要素毎に絶対値を取って二乗した行列を返す演算を表す) を低ランク分散行列 $\mathbf{R}_n$ (ランク K 行列) でモデル化し, その時間周波数の共起性を加味しながら分離行列 $\mathbf{W}_i$ を推定する. 混合前の音源のパワースペクトログラム $|\mathbf{S}_n|^2$ が低ランクであれば, 混合信号のパワースペクトログラム $|\mathbf{X}_n|^2$ のランクは基本的に増加することから, ILRMA は分離信号を低ランクに誘導することでパーミュテーション問題を避けつつ, 互いに独立となる分離信号を推定している.

LRMA の最尤推定問題は式 (9) の最小化問題と等価である. 分離行列 $\mathbf{W}_i$ は, その行ベクトルである $\mathbf{w}_{i,n}$ を次に示す反復射影法 (iterative projection: IP) [7] で全ての n について更新することで最適化される.

$$\mathbf{U}_{i,n} = \frac{1}{J} \sum_j \frac{1}{r_{ij,n}} \mathbf{x}_{ij} \mathbf{x}_{ij}^H \quad (10)$$

$$\mathbf{w}_{i,n} \leftarrow (\mathbf{W}_i \mathbf{U}_{i,n})^{-1} \mathbf{e}_n \quad (11)$$

$$\mathbf{w}_{i,n} \leftarrow \mathbf{w}_{i,n} (\mathbf{w}_{i,n}^H \mathbf{U}_{i,n} \mathbf{w}_{i,n})^{-\frac{1}{2}} \quad (12)$$

ここで, $\mathbf{e}_n \in \mathbb{R}_{\{0,1\}}^{N \times 1}$ は n 番目の要素が 1, それ以外が 0 となるようなベクトルである. 分離ベクトル更新後の分離信号は次式で更新する.

$$y_{ij,n} \leftarrow \mathbf{w}_{i,n}^H \mathbf{x}_{ij} \quad (13)$$

分散行列の更新式は板倉斎藤擬距離 [14] に基づく NMF [15] と同様に, 乗算型反復更新式で最適化できる.

$$t_{ik,n} \leftarrow t_{ik,n} \sqrt{\frac{\sum_j |y_{ij,n}|^2 v_{kj,n} (\sum_{k'} t_{ik',n} v_{k'j,n})^{-2}}{\sum_j v_{kj,n} (\sum_{k'} t_{ik',n} v_{k'j,n})^{-1}}} \quad (14)$$

$$v_{kj,n} \leftarrow v_{kj,n} \sqrt{\frac{\sum_i |y_{ij,n}|^2 t_{ik,n} (\sum_{k'} t_{ik',n} v_{k'j,n})^{-2}}{\sum_i t_{ik,n} (\sum_{k'} t_{ik',n} v_{k'j,n})^{-1}}} \quad (15)$$

$\mathbf{T}_n$ 及び $\mathbf{V}_n$ の更新後は, 推定分散を次式で更新する.

$$\mathbf{R}_n \leftarrow \mathbf{T}_n \mathbf{V}_n \quad (16)$$

3 提案手法

3.1 動機

IVA や ILRMA で複数音声を分離する場合, Fig. 1 に示すようなブロックパーミュテーション問題が発生する場合がある. これは, 音声信号の周波数帯域間の高次相関を推定する際に, 共起性や低ランク性を仮定しても誤った局所最適解に陥り, 別の音源の成分を同一音源の成分とみなす事が主な原因と考えられる. このため, 最適化の途中でブロックパーミュテーション問題が生じた場合, そのまま反復計算による最適化を継続しても, ブロック帯域内での分離は進むものの, ブロックパーミュテーションが解決されることは殆どない.

しかし, Fig. 1 から確認できるように, 音声信号や音楽信号等のブロックパーミュテーションは (各周波数ビンで音源が十分分離できていれば) 分離信号のパワースペクトログラムに対する目視で確認・判別できる可能性が高い. そこで本稿では, web フロントエンドを用いたインタラクティブなアノテーション付与機能を作成し, ユーザからのアノテーション情報を ILRMA にフィードバックするシステムを構築する.

3.2 アノテーションの種類と指定方法

Fig. 2 は, 音源数 $N = 2$ の場合の Web フロントエンドのインターフェース画面を示している. ユーザはスペクトログラム上をマウスでドラッグすることで部分的な領域を指定することができる. Fig. 2 では, ブロックパーミュテーション問題が生じている周波数ビンのみを含むような領域指定をしている. また, 指定した領域が何番目の音源に属するかをテキストボックスに記述できる. 最後に, 画面下部の Frequency のラジオボタンをクリックし submit を押すことで, ILRMA にブロックパーミュテーション問題が生じている周波数ビンの情報が送信される. また, 提案するインターフェースでは, 周波数の範囲だけでなく時間の範囲指定も可能である. 例えば, 2 人の話者の内一方しか発話していない時間範囲を指定し, 画面下部の Time のラジオボタンをクリックし submit を押

すことで、ILRMA に 1 音源のみ沈黙している時間範囲の情報が送信される（この場合テキストボックスの内容は破棄される）。なお、いずれのスペクトログラムも play ボタンを押すことで分離音を聞くことができ、ユーザが領域を指定する際の参考にできる。

3.3 ブロックパーミュテーションの周波数帯の直接指定

今、パーミュテーションが誤っている周波数範囲の開始点 $i = p$ と終了点 $i = q$ （ここで $0 \leq p < q$ ）及び領域指定中の音源インデクス $n = b$ と正しいパーミュテーションとなる交換先の音源インデクス $n = c$ （ブロックの移動先）がユーザから与えられた状況を考える。この場合、以後の ILRMA の更新では該当周波数ピンの分離フィルタ $w_{i,n}$ と基底成分 $t_{ik,n}$ について入れ替え処理を行えばよいので、次式のような処理を行う。

$$w_{p,b} \cdots w_{q,b} \Leftrightarrow w_{p,c} \cdots w_{q,c} \quad (17)$$

$$t_{pk,b} \cdots t_{qk,b} \Leftrightarrow t_{pk,c} \cdots t_{qk,c} \quad \forall k \quad (18)$$

ここで、 $\Leftrightarrow$ は入れ替えを意味する。式 (17) 及び (18) で成分を入れ替えた後は、一度全ての n について分散行列を式 (16) で更新し、ILRMA の反復最適化を再開する。

3.4 1 音源が沈黙している時間範囲の指定

今、 n 番目の音源のみ沈黙している時間範囲の開始点 $j = p$ と終了点 $j = q$ （ここで $0 \leq p < q$ ）がユーザから与えられた状況を考える。この場合、以後の ILRMA の更新では、次式のように、該当時間フレームのアクティベーション成分 $v_{kj,n}$ に非常に小さな値を乗じる処理を行う。

$$v_{kp,n} \cdots v_{kq,n} \leftarrow \varepsilon v_{kp,n} \cdots \varepsilon v_{kq,n} \quad \forall k \quad (19)$$

ここで、 ε は適当に定めた微小な値である。式 (19) の更新後は、一度全ての n について分散行列を式 (16) で更新し、ILRMA の反復最適化を再開する。

4 実験

4.1 実験条件

比較のための信号には SiSEC2011 [16] の UND タスクの 6 信号を用いた。Table. 1 に信号名を示す。STFT は 128 ms のハミング窓を 64 ms のシフトで行った。比較対象及び提案手法での NMF 部分の基底数は、ILRMA の基底数は全て $K = 4$ とした。提案手法の式 (19) 中の ε は 10^{-15} とした。評価値には source-to-distortion ratio (SDR) [17] の改善量を用いた。実験では ILRMA を 50 回反復した後、従来手法ではそのまま残りの 50 回反復を継続し、提案手法ではユーザからのアノテーション情報を与えた上で同様に 50 回反復を継続した。

4.2 実験結果

Fig. 2 の状態において、ブロックパーミュテーションの周波数範囲を指定し ILRMA を 50 回反復したときの結果が Fig. 3 である。これを見ると、ブロックパーミュテーション問題を起こしていた周波数帯域が修復されている様子が分かる。Fig. 4 にブロックパーミュテーションを直接指定した場合の SDR 改善量の

Table 1 Sources used in experiment.

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

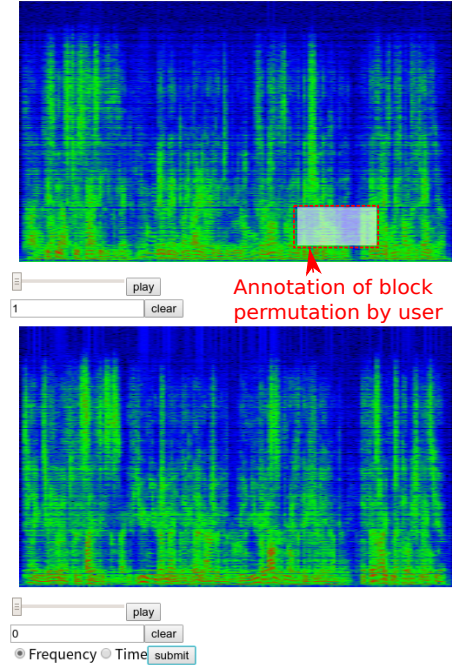


Fig. 2 User interface (before annotated).

推移を示す。Nos. 1 及び 2 の混合音源において、アノテーションを用いた補正後に SDR が改善する現象が確認できる。とくに no. 2 の例の改善量は大きく、提案システムが上手く動いた例といえる。一方で、従来の ILRMA ですでに十分な SDR 改善量を達成している no. 1 の混合音源については、アノテーション情報を用いても大きな変化は得られなかった。また、実際に本システムを利用した印象として、STFT の窓長が長くなるとブロックパーミュテーションの境界線が目視で確認しづらくなった。従って、残響の長い環境での混合信号の場合は、周波数範囲指定よりも沈黙時間範囲の指定の方が幾分容易と思われる。

Fig. 5 は、沈黙時間範囲を指定した場合の SDR 改善量の推移である。アノテーションによる補正後に SDR が上昇する例も確認できるが、全体的に大きな変化は見られなかった。今回用いた音源では 2 人の話者が継続的に発話しているため、今後は通常の会話のように沈黙時間が明確に存在する例での効果を確認する必要がある。

5 まとめ

本稿では、ユーザーとのインタラクションを利用した音源分離手法として、ILRMA に対してアノテーション情報を与えるシステムを提案した。本システムでは、ブロックパーミュテーション問題が生じる帯域や特定の音源が沈黙している時間範囲をアノテーションとして与えることができる。実験の結果、音源分離精度の向上に一定の効果があることを確認した。

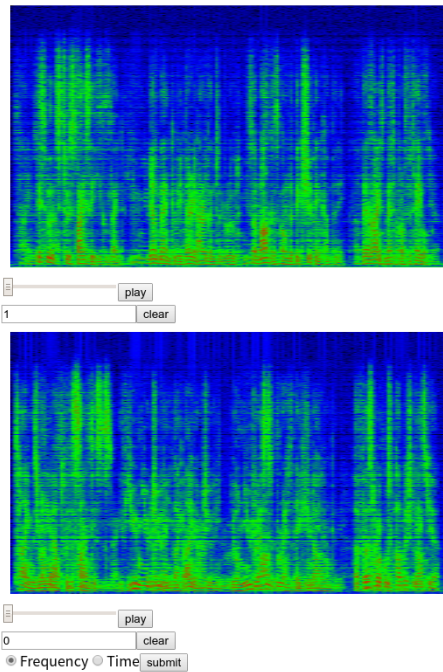


Fig. 3 Spectrograms of block-permutation-fixed estimated signals.

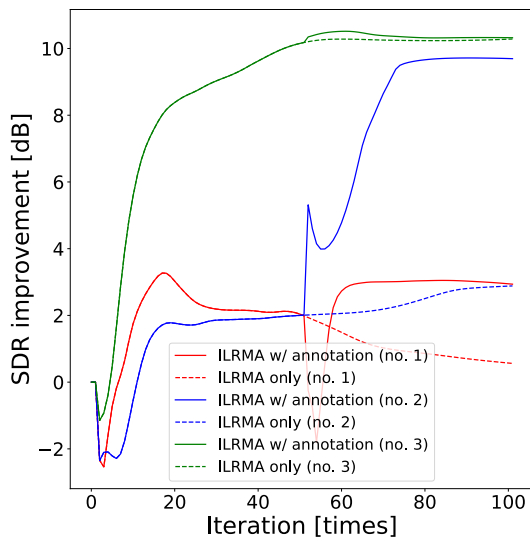


Fig. 4 SDR improvement by first method.

一方で、アノテーションの作成難度が残響長に依存する問題が残るため、より適切なユーザインタフェースを作成することが課題となった。

謝辞 本研究の一部はJSPS 科研費 19K20306 の助成を受けたものである

参考文献

- [1] P. Comon, “Independent component analysis, a new concept?” *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.
- [2] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.
- [3] 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. SAP*, vol. 12, no. 5, pp. 530–538, 2004.
- [4] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of

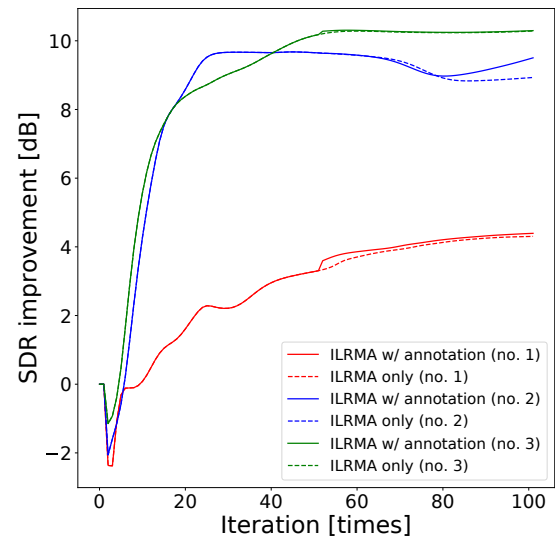


Fig. 5 SDR improvement by second method.

speech signals,” *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

- [5] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, “Blind source separation based on a fast-convergence algorithm combining ICA and beamforming,” *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.
- [6] T. Kim, H. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.
- [7] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” in *Proc. WASPAA*, 2011, pp. 189–192.
- [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 Trans. ASLP*, vol. 24, no. 9, pp. 1622–1637, 2016.
- [9] K. Daichi, O. Nobutaka, S. Hiroshi, K. Hirokazu, and S. Hiroshi, “Determined blind source separation with independent low-rank matrix analysis,” in *Audio Source Separation*, 2018, pp. 125–155.
- [10] D. Kitamura, N. Ono, and H. Saruwatari, “Experimental analysis of optimal window length for independent low-rank matrix analysis,” in *Proc. EUSIPCO*, 2017, pp. 1170–1174.
- [11] Y. Liang, S. Naqvi, and J. Chambers, “Overcoming block permutation problem in frequency domain blind source separation when using auxiva algorithm,” *Electronics Letters*, vol. 48, no. 8, pp. 460–462, 2012.
- [12] Y. Mitsui, D. Kitamura, N. Takamune, H. Saruwatari, Y. Takahashi, and K. Kondo, “Independent low-rank matrix analysis based on parametric majorization-equalization algorithm,” in *Proc. CAMSAP*, 2017, pp. 1–5.
- [13] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [14] F. Itakura and S. Saito, “Analysis synthesis telephony based on the maximum likelihood method,” *Proc. ICA*, pp. C-17–C-20, 1968.
- [15] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis,” *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.
- [16] 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,” in *Proc. LVA/ICA*, 2012, pp. 414–422.
- [17] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.