ユーザーからの補助情報を用いる独立低ランク行列分析*

☆大島風雅 (香川高専), 中野将生 (筑波大), 北村大地 (香川高専)

1 はじめに

ブラインド音源分離（blind source separation: BSS）とは，混合系が不明な観測信号から混合前の音源を推定する技術である．観測チャネル数が音源数以上となる優決定条件では，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．実環境下では，音響信号の混合が残響によって畳み込みになることを考慮し，時間周波数領域で ICA を適用する周波数領域 ICA（frequency-domain ICA: FDICA）[2] が基本となる．

FDICA では，分離信号の周波数ごとの順番とスケールが不定になるという問題が存在する．後者の問題は projection back 法 [3] を用いることで解決が可能だが，前者の問題はパーミュテーション問題 [4] と呼ばれ，解決のために様々な方法が提案されている [3, 4, 5]．特に，音源のパワーが全周波数で共起性を持つことを仮定した独立ベクトル分析（independent vector analysis: IVA）[6, 7] や，音源の時間周波数の共起性が低ランク行列で表せると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．ILRMA では音楽信号に対して頑健かつ高精度な分離を実現しているが，音声信号に対しては IVA と同程度の性能となる場合もあり，実用においては課題が残る．

IVA や ILRMA が音声信号の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にある [11, 12]．これは，Fig.1 のように，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．解決策として，文献 [13] では，ユーザーからのアノテーション情報を用いた BSS システムを開発した．このシステムでは，ブロックパーミュテーションが生じている周波数帯域や，他音源が沈黙し特定の音源のみが発話している時間等のアノテーション情報をユーザーから受け取ることで，ILRMA における最適化をより良い解に誘導するアルゴリズムを提案している．本稿では，従来のアノテーション情報を ILRMA で活用する方法を変更し効果を実験的に調査し，ブロックパーミュテーション問題解決のためのより良いアルゴリズムについて考察する．

2 従来手法

2.1 周波数領域の音源分離

音源数及びマイクロホン数をそれぞれ N 及び M と定義する．また，混合前の音源信号，観測信号，及び分離信号に対して短時間フーリエ変換（short-time Fourier transform: STFT）したものをそれぞれ次のように定義する．

$$\boldsymbol{s}_{ij} = (s_{ij,1}, \cdots, s_{ij,n}, \cdots, s_{ij,N})^{\mathrm{T}} \in \mathbb{C}^{N} \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij,1}, \cdots, x_{ij,m}, \cdots, x_{ij,M})^{\mathrm{T}} \in \mathbb{C}^{M} \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij,1}, \cdots, y_{ij,n}, \cdots, y_{ij,N})^{\mathrm{T}} \in \mathbb{C}^{N} \tag{3}$$

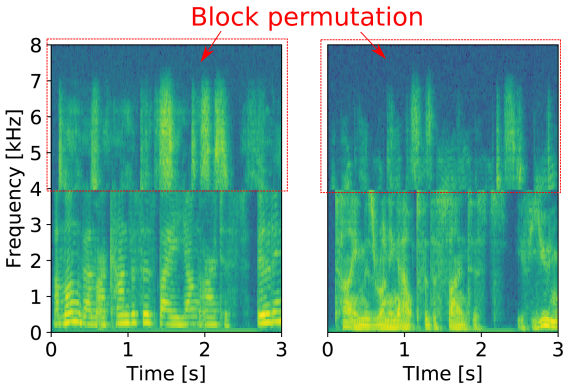

Fig. 1 Example of block permutation problem.

ここで，$i = 1, 2, \cdots, I$，$j = 1, 2, \cdots, J$，$n = 1, 2, \cdots, N$，及び $m = 1, 2, \cdots, M$ はそれぞれ周波数ビン，時間フレーム，音源，及びマイクロホンのインデクスを表し，$\cdot^{\mathrm{T}}$ は行列またはベクトルの転置を表す．周波数毎の混合行列を $\boldsymbol{A}_i \in \mathbb{C}^{M \times N}$ と定義すると，観測信号が次式で表せることを仮定する．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{4}$$

以後，本稿では決定的な系（$M = N$）を取り扱う．$\boldsymbol{A}_i$ がフルランクの場合は，混合行列の逆行列である分離行列 $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1}, \cdots, \boldsymbol{w}_{i,n}, \cdots, \boldsymbol{w}_{i,N})^{\mathrm{H}}$ が存在し，分離信号を次式で表せる．

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{5}$$

ここで，$\cdot^{\mathrm{H}}$ は行列またはベクトルのエルミート転置を表す．BSS は分離行列 $\boldsymbol{W}_i$ を推定する問題である．

2.2 ILRMA

ILRMA[8, 9] は，パーミュテーション問題を回避するために，FDICA の音源モデルとして非負値行列因子分解（nonnegative matrix factorization: NMF）[14, 15] を導入した BSS である．空間モデル（分離行列）$\boldsymbol{W}_i$ の最適化と同時に，分離信号のパワースペクトログラムを音源モデル（NMF 変数）$\boldsymbol{T}_n \boldsymbol{V}_n$ でモデル化する最適化問題として定式化される．ここで，$\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I \times L}$ 及び $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{L \times J}$ は NMF の基底行列及びアクティベーション行列である．

ILRMA の目的関数は次式で定義される．

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

ここで，$\mathsf{W} = \{\boldsymbol{W}_i\}_{i=1}^{I}$，$\mathsf{T} = \{\boldsymbol{T}_n\}_{n=1}^{N}$，及び $\mathsf{V} = \{\boldsymbol{V}_n\}_{n=1}^{N}$ は最適化パラメタの集合である．また，$t_{il,n}$ 及び $v_{lj,n}$ はそれぞれ $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の非負要素であり，

*Indepndent low-rank matrix analysis informed by user annotation. By Fuga OSHIMA (NIT Kagawa), Masaki NAKANO (Tsukuba Univ.), and Daichi KITAMURA (NIT Kagawa).

$l = 1, 2, \cdots, L$ は NMF の基底ベクトルのインデクスである．ILRMA の最適化問題は，式 (6) を最小化する W，T，及び V を求める問題である．

空間モデル $\boldsymbol{W}_i$ の最適化には反復射影法 [7] を用いて次式で表される．

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

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

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

ここで，$\boldsymbol{e}_n \in \mathbb{R}_{\{0,1\}}^{N\times 1}$ は n 番目の要素が 1，他要素が 0 のベクトルである．

音源モデル $\boldsymbol{T}_n \boldsymbol{V}_n$ の更新式は ISNMF [15] における乗算型反復更新式で最適化できる．

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

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

2.3 インタラクティブ音源分離システム

ILRMA や IVA は，音声信号を分離する場合にしばしば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 [11, 12]．これは，最適化の過程で各パラメタが悪い局所最適解に陥り，別の音源の成分を同一の音源の成分とみなしたまま更新されなくなるために起こると推測される．そこで文献 [13] では，ブロックパーミュテーション問題が生じて失敗した状態から復帰し，より良い分離結果を与える解へと誘導するために，最適化途中の分離信号のスペクトログラムを一度ユーザーに提示してアノテーション情報を要求するシステムを開発している．具体的には，ユーザーがブロックパーミュテーション問題の生じている周波数帯域を認める場合に，(a) その周波数帯域を指定した情報及び (b) 他音源が沈黙し特定の音源のみが発話している時間を指定した情報を ILRMA にフィードバックし，再度最適化を行うインタラクティブな音源分離システムである．Fig.2 に開発されたシステムのユーザーインターフェースを示す．次項で，アノテーション情報の ILRMA への活用方法を具体的に説明する．

2.3.1 ブロックパーミュテーション発生周波数帯域の指定

今，周波数ビン $i = i_s$ から $i = e$ $(1 \le i_s < i_e \le I)$ の範囲でブロックパーミュテーション問題が発生し，それに該当する音源インデクスが $n = n_s$，ブロックの移動先となる正しいパーミュテーションの音源が $n = n_t$ であるというアノテーション情報が，ユーザー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ビンの分離フィルタ $\boldsymbol{w}_{i,n}$ と基底成分 $t_{ik,n}$ について入れ替え処理を行えば良いので，次式のような処理を行う．

$$\boldsymbol{w}_{i_s,n_s}, \boldsymbol{w}_{i_s+1,n_s}, \cdots, \boldsymbol{w}_{i_e,n_s} \Leftrightarrow \boldsymbol{w}_{i_s,n_t}, \boldsymbol{w}_{i_s+1,n_t}, \cdots, \boldsymbol{w}_{i_e,n_t} \tag{12}$$

$$t_{i_s k,n_s}, t_{(i_s+1)k,n_s}, \cdots, t_{i_e k,n_s} \Leftrightarrow t_{i_s k,n_t}, t_{(i_s+1)k,n_t}, \cdots, t_{i_e k,n_t} \quad \forall k \tag{13}$$

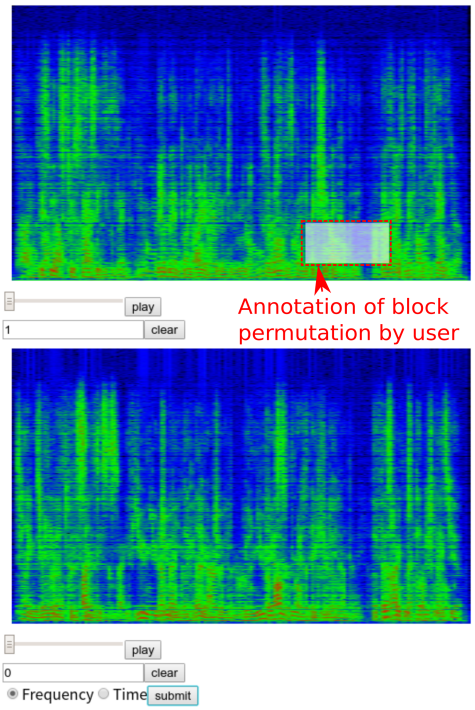

Fig. 2 User interface of interactive audio source separation system.

ここで $\Leftrightarrow$ は左辺と右辺の変数を入れ替えすることを意味する．式 (12) 及び (13) で成分を入れかえた後は，ILRMA の反復最適化を再開する．

2.3.2 他音源が沈黙している時間範囲の指定

今，時間 $j = j_s$ から $j = j_e$ $(1 \le j_s < j_e \le J)$ の範囲で音源 $n = n_t$ の音源が沈黙しているというアノテーション情報がユーザー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 $v_{kj,n}$ に微小値を乗じる．

$$v_{kj_s,n_t}, v_{k(j_s+1),n_t}, \cdots, v_{kj_e,n_t} \leftarrow \varepsilon v_{kj_s,n_t}, \varepsilon v_{k(j_s+1),n_t}, \cdots, \varepsilon v_{kj_e,n_t} \quad \forall k \tag{14}$$

ここで，$\varepsilon > 0$ は適当な微小値である．式 (14) の更新後は，ILRMA の反復最適化を再開する．

3 アノテーション情報の新しい活用法

3.1 動機

前章の 2.3.1 及び 2.3.2 項の手法を用いて ILRMA の最適化をより良い解へと誘導できる可能性が文献 [13] で示されている．しかしながら，この手法がブロックパーミュテーション問題の解決にとって最良である保証は無く，他のアノテーション情報の活用方法とも比較して調査する必要がある．特に，従来手法では一部のパラメタのみに操作を加えているため効果は薄く，最適化を再開しても悪い局所解付近から抜け出せない可能性がある．そこで本稿では，ユーザーから受け取るアノテーション情報を ILRMA の最適化に活用する際の方法を変更し，従来手法との比較実験を通してより良いインタラクティブ音源分離アルゴリズムについて考察する．

3.2 ブロックパーミュテーション発生周波数帯域の指定情報の新しい活用

2.3.1 項と同様に，周波数 $i = i_s$ から $i = i_e$ の範囲でブロックパーミュテーション問題が発生しているというアノテーション情報が，ユーザーから与えられた状況を考える．このとき，従来手法で行われた該当周波数ビンの分離フィルタと基底成分の入れ替え処理 (12) 及び (13) に加えて，アクティベーション行列を以下のように乱数でリセットする処理を行う．

$$v_{kj,n} \leftarrow \rho \quad \forall k, j, n \tag{15}$$

ここで，ρ は区間 $(0, 1)$ の一様乱数である．この処理は，アクティベーション行列 $\boldsymbol{V}_n$ をリセットすることで，現在捕らわれている悪い局所解から一度抜け出すことを目的としている．分離行列 $\boldsymbol{W}_i$ 及び基底行列 $\boldsymbol{T}_n$ は，式 (12) 及び (13) で入れ替えたうえでそのまま引き継いで ILRMA 最適化の反復更新を再開するため，ブロックパーミュテーション問題を回避しながらより高精度な音源分離ができる解へと誘導されることを期待している．

3.3 他音源が沈黙している時間範囲の指定情報の新しい活用 (a)

2.3.2 項と同様に，時間 $j = j_s$ から $j = j_e$ の範囲で音源 $n = n_t$ が沈黙しているというアノテーション情報がユーザーから与えられた状況を考える．このとき，従来手法で行われた該当時間フレームのアクティベーション成分の修正を次式のように一様な微小値で置き換え，同時に分離フィルタについても乱数でリセットする処理を行う．

$$v_{kj_s,n_t},\, v_{k(j_s+1),n_t}, \cdots, v_{kj_e,n_t} \leftarrow \varepsilon, \varepsilon, \cdots, \varepsilon \quad \forall k \tag{16}$$

$$\boldsymbol{w}_{i,n} \leftarrow \rho \quad \forall i, n \tag{17}$$

この処理も前項の手法と同様に，分離行列 $\boldsymbol{W}_i$ をリセットすることで，現在捕らわれている悪い局所解から一度抜け出すことを目的としている．この場合は基底行列 $\boldsymbol{T}_n$ とアクティベーション行列 $\boldsymbol{V}_n$ の一部は引き継いで ILRMA 最適化の反復更新を再開するため，より良い解へと誘導されることを期待している．

3.4 他音源が沈黙している時間範囲の指定情報の新しい活用 (b)

前項の処理 (16) 及び (17) では，沈黙音源の当該時間フレームのアクティベーションに微小値を代入していたが，同時にアクティベーション行列のその他の要素を次式のようにリセットする手法も考えられる．

$$v_{k1,n_t}, v_{k2,n_t}, \cdots, v_{k(j_s-1),n_t} \leftarrow \alpha, \alpha, \cdots, \alpha \quad \forall k \tag{18}$$

$$v_{k(j_s+1),n_t}, v_{k(j_s+2),n_t}, \cdots, v_{kJ,n_t} \leftarrow \alpha, \alpha, \cdots, \alpha \quad \forall k \tag{19}$$

$$v_{kj,n} \leftarrow \alpha \quad \forall k, j, n \neq n_t \tag{20}$$

ここで，α は区間 $[1.0 \times 10^5, 1.1 \times 10^5]$ の一様乱数（ε と比較して十分大きな一様乱数）である．前項の処理 (16) 及び (17) のみの場合と比べ，本項の処理 (16)–(20) はより多くのパラメタをリセットしている．

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

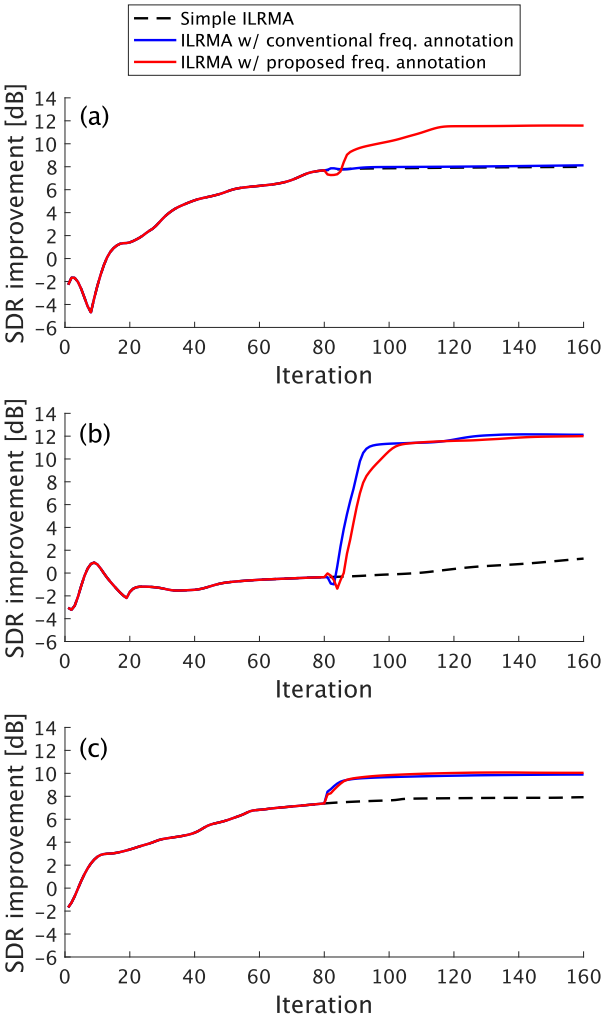

Fig. 3　SDR improvements by frequency annotation: (a) no. 1, (b) no. 2, and (c) no.3.

4　評価実験

4.1　実験条件

比較のための信号には，文献 [13] と同様に，SiSEC2011 [16] の UND タスクの 6 信号を用いる．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．比較対象と提案手法での NMF 基底数は全て $L = 3$ とした．提案手法では，微小値として $\varepsilon = 10^{-15}$ を用いた．評価値には音源対歪み比（source-to-distortion ratio: SDR）[17] の改善量を用いた．実験では通常の ILRMA を 160 回反復した結果と，通常の ILRMA を 80 回反復したタイミングでユーザーからのアノテーション情報を与えた従来手法及び提案手法の結果を比較する．

4.2　実験結果

Fig. 3 は信号 nos. 1–3 について周波数の指定による SDR 改善量の推移を示したものである．信号 no. 1

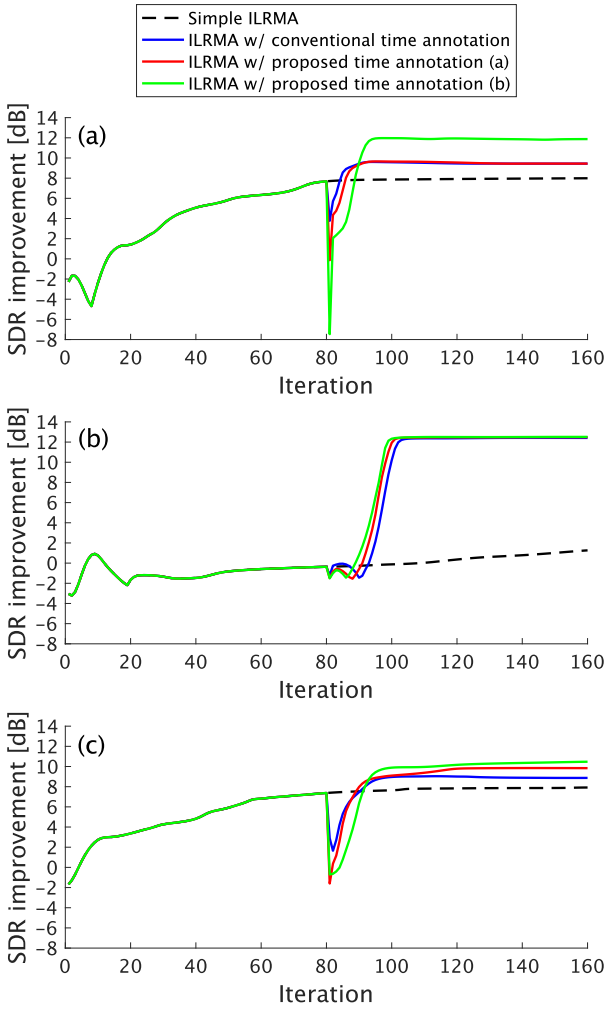

Fig. 4　SDR improvements by time annotation: (a) no. 1, (b) no. 2, and (c) no.3.

に関して，従来手法ではアノテーションを与えたにも関わらず，SDR 改善量がほとんど変わってない．これは，式 (12) 及び (13) での変数更新を行っても，結局同じような局所解から抜け出せないことを示唆している．一方で，式 (12) 及び (13) に加えて式 (15) でアクティベーション行列を初期化する提案手法では，悪い局所解を抜けだし，大きな改善が得られている．残りの 2 つの信号については，従来手法と提案手法で同程度の SDR の改善があった．以上より，周波数アノテーションに関する提案手法は従来手法と比較して，より良い解へと誘導できると考えられる．

Fig. 4 は信号 nos. 1–3 について沈黙時間の指定による SDR 改善量の推移を示したものである．信号 nos. 1 及び 3 では，提案手法が従来手法よりも SDR 改善量の上昇幅が大きい．特に，信号 no. 1 での提案手法 (b) での上昇量は他の手法と比較すると，大きく上昇している．この事実から，時間アノテーションに関する提案手法も，従来手法と同程度またはそれ以上の改善が得られることが分かる．

また，再分離がより良い分離を実現できた時の特徴として，SDR 改善量がアノテーションを与えたときに 0 dB 以下へ下がっている点が挙げられる．これはアノテーションの処理によって一度パラメタが適切にリセットされて再分離したためと考えられる．つまり，SDR 改善量が下がることがブロックパーミュテーション問題の解決に繋がっていると言える．

5　まとめ

本稿では，ユーザーが分離途中のスペクトログラムを見てパラメタを誘導することで，より高精度な BSS を提供するためのアノテーション処理方法を提案した．提案手法は，従来手法よりも安定かつ高精度であることが実験で示された．

しかし，本システムの問題としてアノテーション情報を与えるにあたって，ユーザーが分離精度低下の原因となるブロックパーミュテーションを発見しにくいという問題がある．特に SDR 改善量が 7 dB 前後のとき，ユーザーがスペクトログラムからブロックパーミュテーションが発生している領域を発見するのは困難である．このことから，正確なアノテーションを与えるために，ユーザーに向けた新たな指標が必要であると考える．

謝辞　本研究の一部は JSPS 科研費 19K20306 及び 19H01116 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] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol.41, no. 1–4 pp.1–24, , 2001.
[4] 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.
[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. T. 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,” *Proc. WASPAA*, 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,” *IEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 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,” *Audio Source Separation*, S. Makino, Ed., pp. 125–155. Springer, Cham, 2018.
[10] D. Kitamura, N. Ono, and H. Saruwatari, “Experimental analysis of optimal window length for independent low-rank matrix analysis,” *Proc. EUSIPCO*, pp.1170–1174, 2017.
[11] Y. Liang, S. M. Naqvi, and J. A. Chambers, “Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm,” in *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 majorizaion-equalization algorithm,” *Proc. CAMSAP*, pp.98–102, 2007.
[13] 中野将生, 北村大地, “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” **日本音響学会** *2020* **年春季研究発表会講演論文集**, pp. 421–424, 2020.
[14] 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.
[15] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the β-divergence,” *Neural Comput.*, vol.23, no.9, pp.2421–2456, 2011.
[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,” *Proc. LVA/ICA*, pp. 414-422, 2012.
[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.