深層パーミュテーション解決法の汎化性能に関する実験的評価*

☆蓮池郁也，北村大地（香川高専），渡辺瑠伊（北陸先端大）

1　はじめに

ブラインド音源分離（blind source separation: BSS）[1] とは，事前情報を用いることなく，複数の音源が混合した観測信号から混合前の各音源信号を推定する技術である．優決定条件下では，独立成分分析（independent component analysis: ICA）[2] に基づく手法として周波数領域 ICA（frequency-domain ICA: FDICA）[3] が提案されている．FDICA は観測信号の各周波数ビンに独立な ICA を適用することで BSS を行うが，ICA は一般に分離信号の順序が不定である．従って，Fig. 1 に示すように，FDICA には分離信号成分の順序が周波数間で不揃いになる問題が生じる．この問題はパーミュテーション問題と呼ばれ，劣決定条件下にも応用できる強力な BSS のフルランク共分散分析（full-rank spatial covariance analysis: FCA）[4] においても生じる典型的な問題である．

これまで，様々なパーミュテーション問題の解決法が提案されてきた（例えば [5–7] など）．その後，音源信号の時間周波数構造に関する仮定（音源モデル）を導入し，パーミュテーション問題の解決と周波数毎の BSS を同時に行う手法が提案された．例えば，独立ベクトル分析 [8,9] や独立低ランク行列分析 [10,11] がある．

近年では，音源モデルを仮定してパーミュテーション問題を回避するのではなく，分離信号の正しい並び替えのみを目的とする深層ニューラルネットワーク（deep neural network: DNN）の学習が検討されている [12,13]．これを深層パーミュテーション解決法（deep permutation solver: DPS）と呼ぶ．本稿では，文献 [13] で提案された DPS の汎化性能に関する実験的調査を行い，この手法の有用性を評価する．

2　FDICA とパーミュテーション問題

2.1　信号の定義

短時間 Fourier 変換（short-time Fourier transform: STFT）を適用して得られる時間周波数領域の複数音源信号と多チャネル観測信号を次式でそれぞれ表す．

$$\boldsymbol{s}_{ij} = [s_{ij1}, s_{ij2}, \cdots, s_{ijn}, \cdots, s_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N \quad (1)$$

$$\boldsymbol{x}_{ij} = [x_{ij1}, x_{ij2}, \cdots, x_{ijm}, \cdots, x_{ijM}]^{\mathrm{T}} \in \mathbb{C}^M \quad (2)$$

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，観測チャネルのインデクスを示す．また，FDICA を適用した後に得られる信号を次式で表す．

$$\boldsymbol{z}_{ij} = [z_{ij1}, z_{ij2}, \cdots, z_{ijn'}, \cdots, z_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N \quad (3)$$

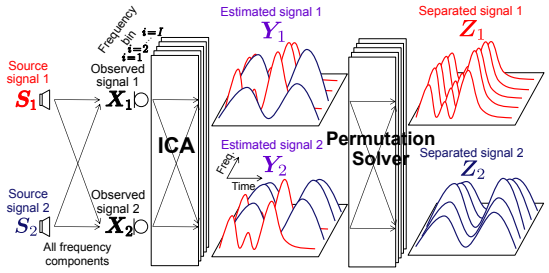

Fig. 1　Permutation problem in FDICA.

$n' = 1, 2, \cdots, N$ は分離信号のインデクスであり，音源の順序が必ずしも n と一致しているとは限らないため，n と n' を使い分ける．本稿では，以後 $M = N$ を仮定する．また，音源信号，観測信号，及び分離信号の複素スペクトログラムをそれぞれ $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$，及び $\boldsymbol{Z}_{n'} \in \mathbb{C}^{I\times J}$ と定義する．

2.2　BSS の定式化と FDICA

FDICA では，観測信号を次式で表す．

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

ここで，$\boldsymbol{A}_i \in \mathbb{C}^{M\times N}$ は周波数毎の時不変混合行列である．混合行列 $\boldsymbol{A}_i$ が正則であれば，周波数毎の分離行列 $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} \in \mathbb{C}^{N\times M}$ が存在し，これを用いて理想的な分離信号を次式で表せる．

$$\boldsymbol{z}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \quad (5)$$

従って FDICA は，観測信号 $\boldsymbol{x}_{ij}$ の各周波数ビンに対して独立に（複素数の）ICA を適用している．

2.3　パーミュテーション問題

ICA は，推定された分離信号成分の周波数毎のスケール及び順序が不定である．従って，FDICA の推定分離行列を $\hat{\boldsymbol{W}}_i \in \mathbb{C}^{N\times M}$ とすると，たとえ完全な推定が実現できたとしても，真の分離行列 $\boldsymbol{W}_i$ に対して次式のような不定性が残る．

$$\hat{\boldsymbol{W}}_i = \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{W}_i \quad (6)$$

ここで，$\boldsymbol{D}_i \in \mathbb{R}^{N\times N}$ は，$\boldsymbol{w}_{in}$ のスケールを変化させる可能性のある対角行列である．また，$\boldsymbol{P}_i \in \{0,1\}^{N\times N}$ は分離行列 $\boldsymbol{W}_i$ の行ベクトル $\boldsymbol{w}_{in}$ の順序を入れ変えうるパーミュテーション行列（置換行列）である．例えば，$N = 2$ であれば $\boldsymbol{P}_i$ は

$$\boldsymbol{P}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (7)$$

の 2 通りの内のいずれかを取る．そのため，FDICA で得られる信号は，次式のように推定信号成分の順

*Experimental evaluation of generalizability of deep permutation solver. By Fumiya HASUIKE, Daichi KITAMURA (NIT Kagawa), and Rui WATANABE (JAIST).

序やスケールが周波数間で不揃いである．

$$\boldsymbol{y}_{ij} = \hat{\boldsymbol{W}}_i \boldsymbol{x}_{ij} \tag{8}$$

$$= \left[y_{ij1}, y_{ij2}, \cdots, y_{ijn'_i}, \cdots, y_{ijN}\right]^{\mathrm{T}} \in \mathbb{C}^N \tag{9}$$

ここで，$n'_i = 1, 2, \cdots, N$ は周波数ビン i 毎に音源の順序が異なっている状態を表すための新たな音源インデクスである．$\boldsymbol{D}_i$ で生じる周波数間のスケールの不整合は，プロジェクションバック法 [14] で解析的に復元できる．しかし，$\boldsymbol{P}_i$ で生じる周波数間の音源順序の不整合を全周波数ビンにわたって復元（整列）すること（$\boldsymbol{P}_i^{-1}$ の推定）は容易ではなく，パーミュテーション問題と呼ばれる．パーミュテーション問題の概要を Fig. 1 に示す．ここで，FDICA で得られる推定信号 $\boldsymbol{y}_{ij}$ の n' 番目のスペクトログラムを $\boldsymbol{Y}_{n'} \in \mathbb{C}^{I \times J}$ と定義している．

理想的なパーミュテーション問題の解決は

$$\boldsymbol{z}_{ij} = \boldsymbol{P}_i^{-1} \boldsymbol{D}_i^{-1} \boldsymbol{y}_{ij} \tag{10}$$

と表せる．但し厳密には，周波数間の音源順序の整列後も，全周波数をまとめた音源信号全体の順序の不定性は残るため，分離信号は次式となる．

$$\boldsymbol{z}_{ij} = \boldsymbol{P}_{\mathrm{all}} \boldsymbol{P}_i^{-1} \boldsymbol{D}_i^{-1} \boldsymbol{y}_{ij} \tag{11}$$

ここで，$\boldsymbol{P}_{\mathrm{all}} \in \{0,1\}^{N \times N}$ は周波数に非依存なパーミュテーション行列である．本稿では，この音源信号全体の順序については復元しない．

3　深層パーミュテーション解決法

3.1　DPS の動機と本稿の目的

1 章で述べたように，パーミュテーション問題をできるだけ回避する BSS として，これまで音源モデルに基づく手法が検討された．しかし，仮定する音源モデルが音源信号に適合しない場合，分離精度は劣化する．その一方で，幅広い音源に適合する万能な音源モデルの構築は困難である．そこで，万能な音源モデルではなく可能な限り汎化性能の高いパーミュテーション解決法を目的として DPS が検討されている [12, 13]．特に文献 [13] で著者らが提案した DPS（以後，提案 DPS と呼ぶ）は一般の音源数に拡張可能であるが，文献 [13] では汎化性能に関して調査されなかった．本稿では，DPS の汎化性能に焦点を当て，より詳細な実験結果の報告を行う．

3.2　提案 DPS における DNN の入出力

FDICA からはパーミュテーション問題が生じた状態の推定信号の複素スペクトログラム $(\boldsymbol{Y}_{n'})_{n'=1}^N$ が得られる．提案 DPS ではまず，これらの信号を用いて正規化パワースペクトログラムを得る．

$$\overline{\boldsymbol{Y}}_{n'} = \frac{|\boldsymbol{Y}_{n'}|^{.2}}{\sum_{n'=1}^N |\boldsymbol{Y}_{n'}|^{.2}} \in [0,1]^{I \times J} \tag{12}$$

ここで，$|\cdot|^{.2}$ は行列の要素毎の絶対値の 2 乗を示す．この正規化は，同一音源に属する成分の相関を強調する [7]．次に，$(\overline{\boldsymbol{Y}}_{n'})_{n'=1}^N$ から，次式のように時間

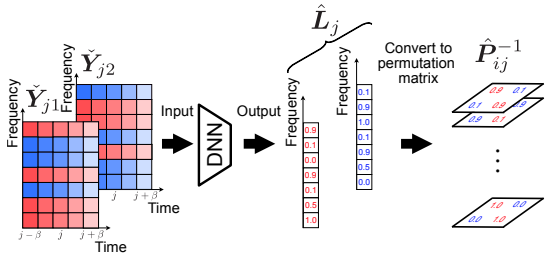

Fig. 2　Estimation of permutation matrix.

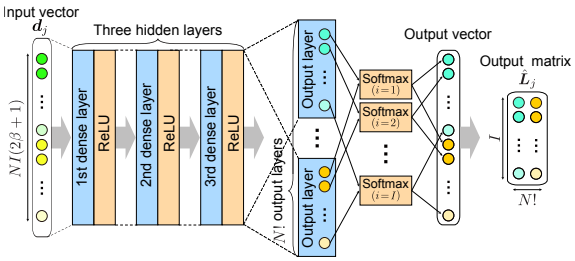

Fig. 3　DNN architecture.

フレーム j を中心とする局所時間パワースペクトログラムを抽出する．

$$\check{\boldsymbol{Y}}_{jn'} = [\overline{\boldsymbol{y}}_{(j-\beta)n'} \quad \cdots \quad \overline{\boldsymbol{y}}_{(j+\beta)n'}] \in [0,1]^{I \times (2\beta+1)} \tag{13}$$

ここで，$\overline{\boldsymbol{y}}_{jn'} \in [0,1]^I$ は $\overline{\boldsymbol{Y}}_{n'}$ の j 列目の列ベクトルを表す．また，β（0 以上の整数）は時間フレーム j の近傍時間フレームをどの程度 DNN に入力するかを決めるハイパーパラメータである．

提案 DPS の DNN の入力ベクトルは，$(\check{\boldsymbol{Y}}_{jn'})_{n'=1}^N$ を一次元にベクトル化したものであり次式で表す．

$$\boldsymbol{d}_j = \left[\mathrm{vec}(\check{\boldsymbol{Y}}_{j1})^{\mathrm{T}}, \cdots, \mathrm{vec}(\check{\boldsymbol{Y}}_{jN})^{\mathrm{T}}\right]^{\mathrm{T}} \in [0,1]^{NI(2\beta+1)} \tag{14}$$

ここで，$\mathrm{vec}(\cdot)$ は行列をベクトル化する処理である．Fig. 2 に示すように，DNN の出力は次式となる．

$$\hat{\boldsymbol{L}}_j = \mathrm{DNN}(\boldsymbol{d}_j) \in [0,1]^{I \times N!} \tag{15}$$

ここで，$\hat{\boldsymbol{L}}_j$ はパーミュテーション行列の確率値 $\hat{l}_{iqj} \geq 0$ から構成される行列であり，$q = 1, 2, \cdots, N!$ は N 個の音源に対する $N!$ 通りの順列のインデクスを表す．この時，$N = 2$ を例とすると予測パーミュテーション行列は $\hat{l}_{i1j}$ と $\hat{l}_{i2j}$ を用いて次式のように表せる．

$$\hat{\boldsymbol{P}}_{ij}^{-1} = \hat{l}_{i1j} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \hat{l}_{i2j} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in [0,1]^{N \times N} \tag{16}$$

また，確率値 $\hat{l}_{iqj}$ は $\sum_q \hat{l}_{iqj} = 1$ を満たす．式 (16) はその定義より二重確率行列であるため，Birkhoff–von Neumann の定理より，DNN はパーミュテーション行列の凸結合係数を予測しているとも捉えられる．

3.3　DNN の構造

Fig. 3 に，提案 DPS で用いる DNN の構造を示す．全て全結合層で構成され，隠れ層の 1 層目から 3 層目には rectified linear unit（ReLU）関数を用いている．出力層で得た $N!$ 個の I 次元のベクトルの同一インデクスの要素に softmax 関数を適用することで，$\hat{l}_{iqj} \geq 0$ かつ $\sum_q \hat{l}_{iqj} = 1$ となることを保証している．

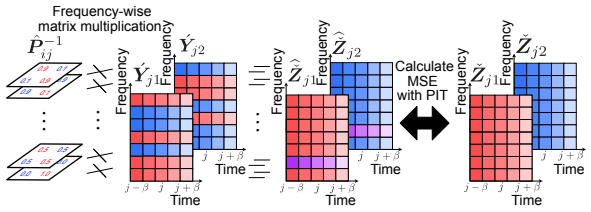

Fig. 4　Loss function using MSE with PIT.

3.4　損失関数

推定パーミュテーション行列 $\hat{\boldsymbol{P}}_{ij}^{-1}$ を求めた後の処理を Fig. 4 に示す．ここで，Fig. 4 中の $(\acute{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ は $(\boldsymbol{Y}_{n'})_{n'=1}^{N}$ の局所時間複素スペクトログラムである．DNN を用いて求めた予測分離信号 $(\widehat{\boldsymbol{Z}}_{n'})_{n'=1}^{N}$ と予測分離信号に対する正解ラベル $(\check{\boldsymbol{Z}}_{n'})_{n'=1}^{N}$（分離信号 $(\boldsymbol{Z}_{n'})_{n'=1}^{N}$ の局所時間複素スペクトログラム）を用意し，$(\widehat{\widehat{\boldsymbol{Z}}}_{n'})_{n'=1}^{N}$ と $(\check{\boldsymbol{Z}}_{n'})_{n'=1}^{N}$ の間で損失関数として平均二乗誤差（mean squared error: MSE）を用いる．ここで，2.3 節で述べた通り，提案 DPS は $\boldsymbol{P}_{\mathrm{all}}^{-1}$ の推定を目的としない．この不定性を許容しつつ予測とラベルの誤差を測るために，順序不変学習（permutation invariant training: PIT）[15] を導入した損失関数 $\mathcal{L}$ を用いる．

$$\mathcal{L} = \min(C_1, C_2, \cdots, C_{N!}) \tag{17}$$

$$C_q = \sum_{n'}^{N} ||\widehat{\widehat{\boldsymbol{Z}}}_{jn'} - \check{\boldsymbol{Z}}_{j\mathcal{P}(q,n')}||_2^2 \tag{18}$$

ここで，$\min(\cdot)$ は入力の最小値，$\mathcal{P}(q, n')$ は $N!$ 個の全てのありうる順列の内，q 番目の順列における n' 番目の値を返す処理を表す．

3.5　DNN のテストデータへの適用

DNN 学習後は，提案 DPS を FDICA の推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{N}$ に適用する．パーミュテーション問題は時不変な分離行列 $\boldsymbol{W}_i$ で生じることから，正しい音源順序は時間フレーム方向には常に一定である．そのため，テストデータへの適用時には，様々な時間 j の局所時間パワースペクトログラム $(\acute{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ を DNN に入力し，その出力 $(\hat{\boldsymbol{P}}_{ij}^{-1})_{j=1}^{J}$ を次式のように多数決処理することで，更なる精度向上が期待できる．

$$\hat{\boldsymbol{P}}_{i}^{-1} = \mathrm{round}\left(\frac{1}{J}\sum_{j=1}^{J}\hat{\boldsymbol{P}}_{ij}^{-1}\right) \in \{0,1\}^{N\times N} \tag{19}$$

ここで，$\mathrm{round}(\cdot)$ は入力行列の要素毎の四捨五入を表す．最終的な推定分離信号は次式で得られる．

$$\hat{\boldsymbol{z}}_{ij} = \hat{\boldsymbol{P}}_{i}^{-1}\boldsymbol{y}_{ij} \tag{20}$$

4　実験

4.1　実験条件

提案 DPS の汎化性能を評価するために，音声信号だけを用いて学習した DNN モデルと音楽信号だけを用いて学習した DNN モデルの 2 つを用意し，インドメイン（学習データとテストデータに同じ音源を用いる）とアウトドメイン（学習データとテストデータに異なる音源を用いる）に対する実験を行った．音源信号 $(\boldsymbol{S}_n)_{n=1}^{2}$ として Table 1 に示す男女の音声信号又はドラムとベースの音楽信号の 2 種類のペアを用いた．両信号のサンプリング周波数は 16 kHz である．

Table 1　Speech and music sources obtained from SiSEC2011 [16]

Signal type	Source	Data name	Length
Speech	Male speech	`dev2_male4_inst_src_2.wav`	10.0 s
	Female speech	`dev3_female4_inst_src_2.wav`	10.0 s
Music	Drums	`dev1_wdrums_src_3.wav`	11.0 s
	Bass	`dev1_wdrums_src_2.wav`	11.0 s

STFT における分析窓関数長（短時間信号長）は 2048 点（128 ms），シフト長は 1024 点（64 ms）と設定し，窓関数にはハン窓を用いた．各信号のスペクトログラムは $I = 1025$ 及び $J = 158$（音声信号）又は $J = 173$（音楽信号）となった．

本実験ではブロックパーミュテーション問題を模擬する．具体的には，各ブロックの周波数ビン数を 16 として $(\boldsymbol{S}_n)_{n=1}^{2}$ の全周波数を 64 ブロックに分割し，ブロック単位でランダムに周波数成分を入れ替えることで，ブロックパーミュテーション問題の残る推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{2}$ を模擬した．学習データには重複なしのランダム入れ替え 300 パターンで作成した $(\boldsymbol{Y}_{n'})_{n'=1}^{2}$ を用いた．性能評価に用いるテストデータには，学習データにはないランダム入れ替え 10 パターンで作成した $(\boldsymbol{Y}_{n'})_{n'=1}^{2}$ を用いた．提案 DPS に用いる DNN の隠れ層の次元は全て 4096 とし，式 (13) の β は 13 とした．最適化手法は Adam，ミニバッチサイズは 8，エポック数は 1000 とした．また，$N = 2$ の時にのみ適用できる文献 [12] の DPS（以後，従来 DPS と呼ぶ）と提案 DPS の比較実験も行った．従来 DPS の DNN モデルの条件や各パラメータ，最適化関数等は文献 [12] と同一に設定した．評価指標には，信号対歪み比（source-to-distortion ratio: SDR）[17] の改善量を用いた．

4.2　実験結果

Figs. 5 及び 6 は，従来 DPS と提案 DPS におけるインドメインのデータセットに対する DNN の入力信号と並び替えた信号を示している．従来 DPS，提案 DPS のどちらも高精度でパーミュテーション問題を解決している．しかし，従来 DPS は，SDR や主観的聴取評価に強い影響を与える低周波帯域の並び替えによく失敗している．Figs. 7 及び 8 は，ドメイン別の SDR を示している．それぞれの手法において，2 種類の音源の平均を取った SDR を示している．提案 DPS では，インドメインデータセットに対して平均で 20 dB 以上の SDR の改善が見られた．さらに，提案 DPS はアウトドメインデータセットに対しても平均で 10 dB 以上の SDR の改善がある．今回学習に用いた音源は音声信号及び音楽信号であり，これらは全く異なる時間周波数構造を持つ信号であるが，音声信号で学習した DPS が音楽信号にもある程度の性能を発揮でき，その逆も言えることから，提案 DPS が

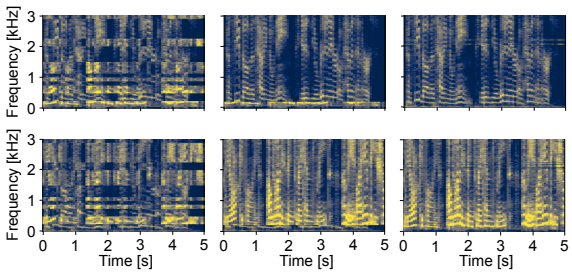

Fig. 5　Input (left) and estimated signals of conventional (center) and proposed DPSs (right): male (top) and female (bottom) speech.

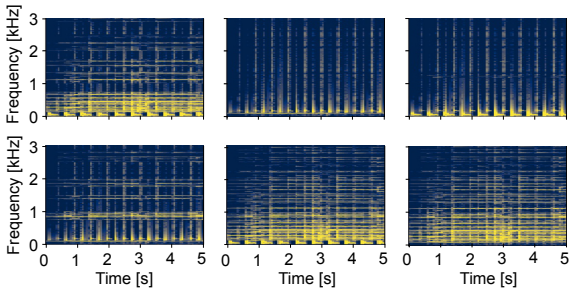

Fig. 6　Input (left) and estimated signals of conventional (center) and proposed DPSs (right): drums (top) and bass (bottom).

学習データのドメインに対する頑健性を持ち，汎化性能の高いアプローチであることが示唆された．

5　まとめ

本稿では，提案 DPS の汎化性能に関する評価を行った．実験結果より，提案 DPS はアウトドメインのデータセットに対して平均で 10 dB 以上の SDR の改善があり，汎化性能が高いことを示した．

謝辞　本研究の一部は JSPS 科研費 19K20306 及び 22H03652 の助成を受けたものである．

参考文献

[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 TSIP*, vol. 8, no. e12, pp. 1–14, 2019.

[2] P. Comon, “Independent component analysis, a new concept?,” *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

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

[4] N. Q. K. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[5] 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.

[6] 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 TASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[7] H. Sawada, S. Araki, and S. Makino, “Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS,” *Proc. ISCAS*, pp. 3247–3250, 2007.

[8] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE TASLP*, vol. 15, no. 1, pp. 70–79, 2007.

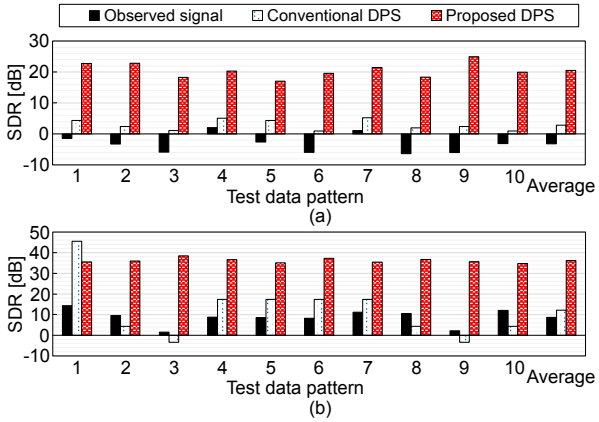

Fig. 7　Average SDR for in-domain test data: (a) speech mixtures with speech models and (b) music mixtures with music models.

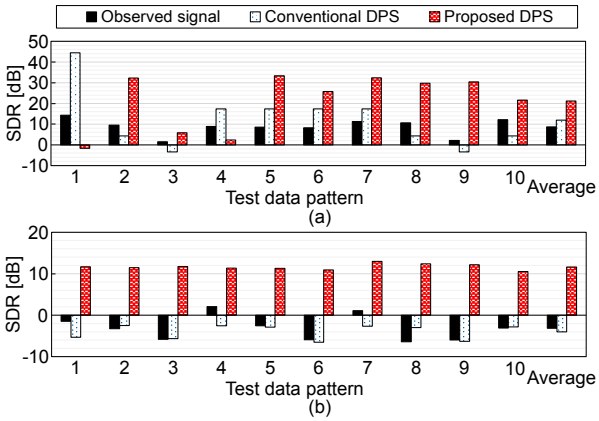

Fig. 8　Average SDR for out-domain test data: (a) music mixtures with speech models and (b) speech mixtures with music models.

[9] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. WASPAA*, pp. 189–192, 2011.

[10] 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 TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[11] 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., pp. 125–155. Springer, Cham, 2018.

[12] S. Yamaji and D. Kitamura, “DNN-based permutation solver for frequency-domain independent component analysis in two-source mixture case,” *Proc. APSIPA ASC*, pp. 781–787, 2020.

[13] 蓮池郁也, 渡辺瑠伊, 北村大地, “深層ニューラルネットワークに基づくパーミュテーション解決法の基礎的検討,” **信学技報**, EA2022-13, vol. 122, no. 20, pp. 62–67, 2022.

[14] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” *Proc. ICA*, pp. 722–727, 2001.

[15] D. Yu, M. Kolbak, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” *Proc. ICASSP*, pp. 241–245, 2017.

[16] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, 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 TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.