周波数双方向再帰に基づく深層パーミュテーション解決法

Deep Permutation Solver Based on Frequency Bidirectional Recursion

蓮池郁也 † 北村大地 † 渡辺瑠伊 ‡ 川口翔也 †
† 香川高等専門学校
‡ 北陸先端科学技術大学院大学

Fumiya HASUIKE†, Daichi KITAMURA†, Rui WATANABE‡, and Shoya KAWAGUCHI†
†National Institute of Technology, Kagawa College
‡Japan Advanced Institute of Science and Technology

アブストラクト 本稿では，周波数領域ブラインド音源分離で発生する周波数領域のパーミュテーション問題について取り扱う．我々は過去に深層ニューラルネットワーク（DNN）に基づくパーミュテーション解決法を提案したが，この手法ではブロック単位で周波数成分が入れ替わるブロックパーミュテーション問題しか解けない問題があった．そこで本稿では DNN のネットワーク構造を改良し，周波数ビン単位のパーミュテーション問題を解決できる手法の構築を目指す．提案法では，双方向再帰型 DNN を周波数方向に適用することで，従来手法では解けなかった周波数ビン単位のパーミュテーション問題が解決できることを実験的に確認する．

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る．

BSS の歴史では，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 [4] 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 [5,6]．近年では，音源モデルを事前学習する手法も提案されている [7,8]．

しかしながら，これらの手法をもってしても，特定のまとまった帯域で分離信号成分の順序を間違えるブロックパーミュテーション問題が生じることが報告されている [9–11]．このブロックパーミュテーション問題の解決法も研究されており，ユーザとのインタラクションを用いる手法 [10] や音源モデルを用いてコスト行列を設計しハンガリー法を適用する手法 [11,12] が提案されている．

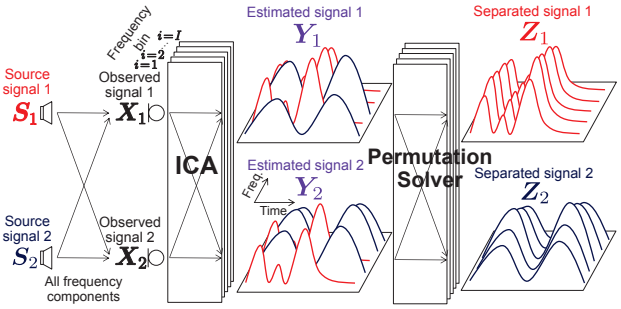

Fig. 1 Permutation problem in FDICA ($N = 2$).

一方で，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）の学習も検討されている [13–16]．これを深層パーミュテーション解決法（deep permutation solver: DPS）と呼ぶ．特に文献 [14–16] では，多層パーセプトロン（multilayer perceptron: MLP）を用いた DNN に基づく DPS（以後，MLP-DPS と呼ぶ）が検討され，ある程度のブロックパーミュテーション問題の解決ができることが示されたが，本稿の実験で確認する通り，MLP-DPS は周波数ビン単位のパーミュテーション問題を解決できない．そこで本稿では，周波数ビン単位のパーミュテーション問題を解決する DPS の構築を目指し，周波数双方向再帰を用いた DNN を DPS に活用する．

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音源信号，

観測信号，及び分離信号を次式でそれぞれ表す．

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

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

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

ここで，$i = 1, 2, \cdots, I$，$j = 1, 2, \cdots, J$，$n = 1, 2, \cdots, N$，$m = 1, 2, \cdots, M$，及び $n' = 1, 2, \cdots, N$ はそれぞれ周波数ビン，時間フレーム，音源信号，観測チャネル，及び分離信号のインデクスを示す[1]．また，$\cdot^{\mathrm{T}}$ は転置を表す．さらに，分離信号の複素スペクトログラムを $\boldsymbol{Z}_{n'} \in \mathbb{C}^{I \times J}$ と定義する．本稿では，以後 $M = N$ を仮定する．

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．$\boldsymbol{A}_i$ 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，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，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り，$N = 3$ であれば

$$\boldsymbol{P}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \text{ or } \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad (8)$$

[1]分離信号は，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' を使い分けている．

の 6 通りの内のいずれかを取る．そのため，FDICA で得られる信号を $\boldsymbol{y}_{ij}$ とすると，次式のように推定信号成分の順序やスケールが周波数間で不揃いである．

$$\boldsymbol{y}_{ij} = \hat{\boldsymbol{W}}_i \boldsymbol{x}_{ij} \quad (9)$$

$$= \left[y_{ij1}, y_{ij2}, \cdots, y_{ijn'_i}, \cdots, y_{ijN}\right]^{\mathrm{T}} \in \mathbb{C}^N \quad (10)$$

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 [17] 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

$$\boldsymbol{z}_{ij} = \boldsymbol{P}_i^{-1} \boldsymbol{D}_i^{-1} \boldsymbol{y}_{ij} \quad (11)$$

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．

$$\boldsymbol{z}_{ij} = \boldsymbol{P}_{\mathrm{all}} \boldsymbol{P}_i^{-1} \boldsymbol{D}_i^{-1} \boldsymbol{y}_{ij} \quad (12)$$

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の復元は対象としない．

3 提案手法

3.1 提案手法の動機

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，音源モデルに基づく手法が広く検討されてきたが，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．著者らは，汎化性能の高いパーミュテーション解決モデルの構築を目的として，DPS を検討している [13–16]．文献 [14–16] で提案した DPS は，ブロックパーミュテーション問題を解決できることを確認したが，本稿の実験で示す通り，周波数ビン単位のパーミュテーション問題の解決は困難であった．本稿では，各音源の周波数方向の関係性を明確に学習するために，長・短期記憶（long-short term memory: LSTM）ユニット [18] を用いた双方向再帰型ニューラルネットワーク（bidirectional recurrent neural network using LSTM: BiLSTM）に基づく DPS（以後，BiLSTM-DPS と呼ぶ）を提案し，周波数ビン単位のパーミュテーション問題の解決を目指す．

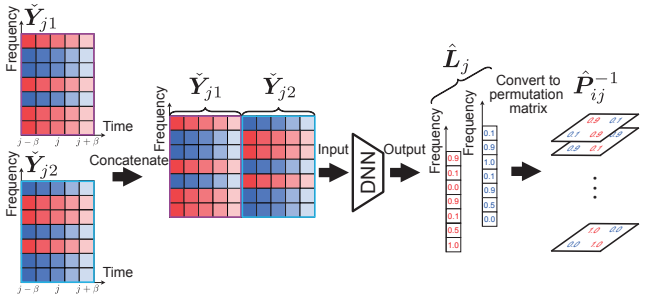

Fig. 2　Estimation of permutation matrix ($N = 2$).

3.2　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．MLP-DPS 及び BiLSTM-DPS ではまず，これらの信号を次式で正規化パワースペクトログラムに変換する．

$$\overline{\boldsymbol{Y}}_{n'} = \frac{|\boldsymbol{Y}_{n'}|^{\cdot 2}}{\sum_{n'=1}^{N} |\boldsymbol{Y}_{n'}|^{\cdot 2}} \in [0,1]^{I\times J} \tag{13}$$

ここで，$|\cdot|^{\cdot 2}$ 及び括線はそれぞれ行列の要素毎の絶対値の 2 乗及び要素毎の割り算を示す．この正規化は，同一音源に属する成分の相関を強調する [4]．次に，$(\overline{\boldsymbol{Y}}_{n'})_{n'=1}^{N}$ から，次式のように時間フレーム 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．

$$\check{\boldsymbol{Y}}_{jn'} = [\overline{\boldsymbol{y}}_{(j-\beta)n'} \ \cdots \ \overline{\boldsymbol{y}}_{(j+\beta)n'}] \in [0,1]^{I\times(2\beta+1)} \tag{14}$$

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．MLP-DPS では $(\check{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ をベクトル化し MLP に入力するが，BiLSTM-DPS では $(\check{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ を時間方向に結合した行列 $[\check{\boldsymbol{Y}}_{j1} \ \cdots \ \check{\boldsymbol{Y}}_{jN}] \in [0,1]^{I\times N(2\beta+1)}$ を BiLSTM に入力する（Fig. 2 参照）．

MLP-DPS 及び BiLSTM-DPS は，予測結果として行列 $\hat{\boldsymbol{L}}_j \in [0,1]^{I\times N!}$ を出力する．$\hat{\boldsymbol{L}}_j$ はパーミュテーション行列の予測確率値 $\hat{l}_{iqj} \geq 0$ 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．また，$\hat{l}_{iqj}$ は確率値なので $\sum_q \hat{l}_{iqj} = 1$ を満たす．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．

$$\hat{\boldsymbol{P}}_{ij}^{-1} = \hat{l}_{i1j}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \hat{l}_{i2j}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \in [0,1]^{N\times N} \tag{15}$$

3.3　BiLSTM-DPS における DNN の構造

BiLSTM-DPS では，周波数ビン単位のパーミュテーション問題を解くうえで重要となる各音源の周波数方向の関係性を明確に学習するため，周波数方向に対して BiLSTM を適用する．BiLSTM は，時間や周波数等の連続的な系列の次元をもつ入力に対して，その系列の順方向及び逆方向の再帰性を考慮した学習ができる DNN である．

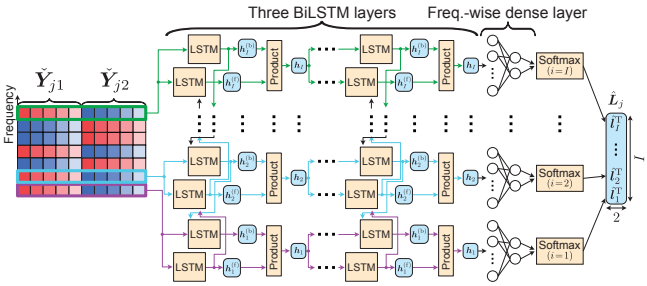

Fig. 3　DNN architecture of BiLSTM-DPS ($N = 2$).

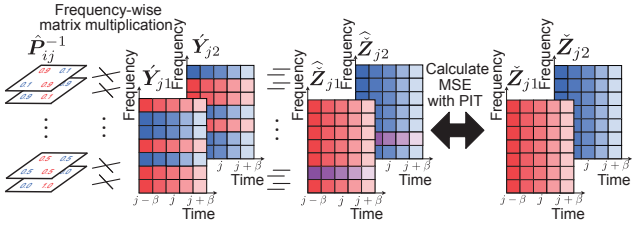

Fig. 4　Loss function using MSE with PIT ($N = 2$).

Fig. 3 に BiLSTM-DPS で用いる DNN の構造を示す．最初に BiLSTM を 3 層適用する[2]．各 BiLSTM 層では，周波数ビンの順方向の特徴量 $\boldsymbol{h}_i^{(\mathrm{f})} \in \mathbb{R}^{N(2\beta+1)}$ と逆方向の特徴量 $\boldsymbol{h}_i^{(\mathrm{b})} \in \mathbb{R}^{N(2\beta+1)}$ を出力し，次式のように同一周波数ビンで要素毎に乗算したベクトルを出力する．

$$\boldsymbol{h}_i = \boldsymbol{h}_i^{(\mathrm{f})} \odot \boldsymbol{h}_i^{(\mathrm{b})} \in \mathbb{R}^{N(2\beta+1)} \tag{16}$$

ここで，$\odot$ は要素毎の積を表す．3 層の BiLSTM の後は，特徴量 $\boldsymbol{h}_i$ を $N(2\beta+1)$ から $N!$ に次元圧縮するために，周波数毎の全結合層を通し Softmax 関数を適用する．

$$\hat{\boldsymbol{l}}_i = \mathrm{Softmax}(\mathrm{Dense}(\boldsymbol{h}_i)) \in [0,1]^{N!} \tag{17}$$

ここで，Dense(·) 及び Softmax(·) はそれぞれ全結合層と Softmax 関数を表す．DNN の出力である $\hat{\boldsymbol{L}}_j$ は，$\hat{\boldsymbol{l}}_i^{\mathrm{T}}$ を行ベクトルに持つ行列である．Softmax 関数により，$\hat{\boldsymbol{L}}_j$ の要素は $\hat{l}_{iqj} \geq 0$ かつ $\sum_q \hat{l}_{iqj} = 1$ が保証されている．

3.4　DPS における DNN の損失関数

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 $(\check{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ 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 $(\widehat{\check{\boldsymbol{Z}}}_{n'})_{n'=1}^{N}$ 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，$(\widehat{\check{\boldsymbol{Z}}}_{n'})_{n'=1}^{N}$ 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，提案 DPS は $\boldsymbol{P}_{\mathrm{all}}^{-1}$ の推定を目

[2] ここでは 1 つの BiLSTM の中で層を増やす構造（multilayer BiLSTM）ではなく，BiLSTM 層そのものを複数重ねる構造（stacked BiLSTM）を採用している．

Table 1 Speech and music sources obtained from SiSEC2011 [20]

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
	Guitar	dev1_wdrums_src_2.wav	11.0 s

的としないため，順序不変学習（permutation invariant training: PIT）[19] を導入した損失関数 $\mathcal{L}$ を用いる．

$$\mathcal{L} = \min(C_1, C_2, \cdots, C_q, \cdots, C_{N!}) \tag{18}$$

$$C_q = \sum_{n'}^{N} ||\hat{\tilde{\boldsymbol{Z}}}_{jn'} - \check{\boldsymbol{Z}}_{j\mathcal{P}(q,n')}||_2^2 \tag{19}$$

ここで，$\min(\cdot)$ は入力の最小値を返す関数であり，$\mathcal{P}(q, n')$ 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 DPS のテストデータへの適用

DNN 学習後は，DPS を推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{N}$ に適用できる．パーミュテーション問題は時不変な分離行列 $\hat{\boldsymbol{W}}_i$ 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は，様々な時間 j の局所時間パワースペクトログラム $(\check{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ を DPS に入力し，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．

$$\hat{\boldsymbol{P}}_i^{-1} = \mathrm{round}\left(\frac{1}{J}\sum_{j=1}^{J}\hat{\boldsymbol{P}}_{ij}^{-1}\right) \in \{0,1\}^{N\times N} \tag{20}$$

ここで，round(·) 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．

$$\hat{\boldsymbol{z}}_{ij} = \hat{\boldsymbol{P}}_i^{-1}\boldsymbol{y}_{ij} \tag{21}$$

4 実験

4.1 実験条件

従来 DPS [13]，MLP-DPS [14–16]，及び BiLSTM-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，in-domain（学習データとテストデータに全く同じ音源を用いる）と out-of-domain（学習データとテストデータに異なる種類の音源を用いる）に対する実験を行った[3]．具体的には，音源信号 $(\boldsymbol{S}_1, \boldsymbol{S}_2)$ として Table 1 の男女の音声信号又はドラムとギターの音楽信号の 2 種類のペアを用いた．信号のサンプリング周波数はいずれも 16 kHz である．STFT における窓長は 2048 点（128 ms），シフト長は 1024 点（64 ms）に設定し，窓関数には Hann 窓を用いた．本実験では，2 つの音源信号 $(\boldsymbol{S}_1, \boldsymbol{S}_2)$ を周波数ビン単位でランダムに入れ替えることで，パーミュテーション問題が残る推定信号 $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ 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 150 パターンで作成した $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ を用い，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 $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ を用いた．

従来 DPS の DNN の条件等は文献 [13] と同一とした．MLP-DPS 及び BiLSTM-DPS では，局所時間長を $\beta = 13$ に設定し，最適化手法は Adam，ミニバッチサイズは 8，エポック数は 500 とした．その他の設定は文献 [15, 16] と同一である．評価指標には，信号対歪み比（source-to-distortion ratio: SDR）[21] の改善量を用いた．

4.2 実験結果

4.2.1 In-domain のテストデータに対する結果

Tables 2 及び 3 は，それぞれ音楽信号及び音声信号の in-domain のテストデータに対する SDR を示している．In-domain では学習データとテストデータで同じ音源信号 $(\boldsymbol{S}_1, \boldsymbol{S}_2)$ を用いていることから，高い性能であっても過学習を起こしている可能性がある点に留意する．

結果より，従来 DPS では観測の SDR 値から一定量の改善があるものの，音楽信号で平均 3 dB 程度，音声信号で平均 10 dB 程度となった．また，MLP-DPS はほとんど全てのデータに対して SDR の改善が得られなかった．文献 [15, 16] の実験結果と合わせて考察すると，MLP-DPS ではブロックパーミュテーション問題を解決するモデルの学習は可能だが，本実験の学習方法では周波数ビン単位のパーミュテーション問題を解決するモデルの学習が困難なことが分かる．一方，BiLSTM-DPS ではほぼ全てのデータで 50 dB 以上の SDR の改善があり，周波数ビン単位のパーミュテーション問題を解決できている．

4.2.2 Out-of-domain のテストデータに対する結果

Tables 4 及び 5 は，それぞれ音楽信号及び音声信号の out-of-domain のテストデータに対する SDR を示している．即ちこれらの結果は「音声信号で学習した DPS が音楽信号のパーミュテーション問題を解決できるか（Table 4）」及び「音楽信号で学習した DPS が音声信号のパーミュテーション問題を解決できるか（Table 5）」をそれぞれ表しており，各モデルの汎化性能を比較している．また，観測信号及び各手法の out-of-domain における推定結果の一例（female speech 及び drums）を Figs. 5 及び 6 に示す．

MLP-DPS は，Tables 4 及び 5 の両条件において，in-domain と同様にパーミュテーション問題の解決に失敗し

[3] 本実験は文献 [15, 16] の実験条件を踏襲しているが，ブロックパーミュテーション問題ではなく周波数ビン単位のパーミュテーション問題を対象とする点のみが異なる実験条件になっている．

Table 2 SDRs [dB] for music test data using DPS trained with music signals (in-domain evaluation)

Test data pattern	Observed signal	Conventional DPS	MLP-DPS	BiLSTM-DPS
1	-0.95	2.95	1.80	**64.75**
2	2.00	2.95	-0.20	**64.75**
3	0.55	2.95	2.75	**155.00**
4	1.25	2.95	2.25	**64.75**
5	-1.00	2.95	-1.25	**66.65**
6	-1.00	2.95	-1.40	**61.15**
7	-0.85	2.95	-1.95	**66.65**
8	-0.15	2.95	2.10	**64.75**
9	0.60	2.95	0.70	**64.75**
10	-0.35	2.95	-0.80	**61.15**

Table 3 SDRs [dB] for speech test data using DPS trained with speech signals (in-domain evaluation)

Test data pattern	Observed signal	Conventional DPS	MLP-DPS	BiLSTM-DPS
1	-6.25	3.60	-8.45	**44.5**
2	-6.85	4.65	-7.45	**44.5**
3	-5.40	3.60	-9.10	**44.5**
4	-6.45	3.55	-6.20	**44.5**
5	-6.60	4.70	-7.95	**44.5**
6	-6.45	4.65	-8.50	**44.5**
7	-6.35	3.60	-6.80	**44.5**
8	-5.50	4.65	-8.45	**44.5**
9	-5.85	3.60	-7.65	**44.5**
10	-5.55	4.65	-7.70	**44.5**

Table 4 SDRs [dB] for speech test data using DPS trained with music signals (out-of-domain evaluation)

Test data pattern	Observed signal	Conventional DPS	MLP-DPS	BiLSTM-DPS
1	-6.25	-8.00	-4.45	**33.55**
2	-6.85	-5.85	-5.00	**22.85**
3	-5.40	-7.20	-6.25	**33.85**
4	-6.45	-7.60	-6.60	**23.50**
5	-6.60	-7.40	-5.90	**22.00**
6	-6.45	-7.25	-4.95	**24.05**
7	-6.35	-1.40	-5.65	**23.60**
8	-5.50	-7.65	-6.70	**26.65**
9	-5.85	-6.40	-4.75	**25.15**
10	-5.55	-7.90	-5.45	**24.05**

Table 5 SDRs [dB] for music test data using DPS trained with speech signals (out-of-domain evaluation)

Test data pattern	Observed signal	Conventional DPS	MLP-DPS	BiLSTM-DPS
1	-0.95	**5.05**	4.85	3.35
2	2.00	**5.05**	-0.50	1.75
3	0.55	**5.05**	2.55	3.35
4	1.25	**11.35**	0.40	3.35
5	-1.00	**11.35**	0.50	3.35
6	-1.00	**11.35**	2.05	3.35
7	-0.85	**11.35**	0.30	3.35
8	-0.15	**5.05**	-0.40	3.35
9	0.60	**5.05**	-0.35	3.35
10	-0.35	**5.05**	-1.25	1.75

ている．従来 DPS は，Table 4 の条件では改善が得られなかったが，Table 5 では一定の改善を確認でき，音声信号で学習した従来 DPS がある程度の汎化性能を獲得していることが分かる．一方，BiLSTM-DPS は Table 5 で性能が低下したものの，Table 4 において 24 dB 以上の改善が得られており，従来 DPS とは逆に音楽信号で学習した BiLSTM-DPS が汎化性能を獲得している．

Tables 3 及び 5 の結果から，音声信号で学習した BiLSTM-DPS は過学習を起こしていることが予想される．参考として，BiLSTM-DPS を音楽信号又は音声信号で学習した際の損失関数値の遷移を Fig. 7 に示す．音声信号での学習時は音楽信号での学習時と比べて損失関数値が小さいことから，パーミュテーション問題を完全に解決した信号（即ち音源信号 $(\boldsymbol{S}_1, \boldsymbol{S}_2)$）を記憶するような過学習が起こったと予想される．本実験条件は，Table 1 の音声信号ペア又は音楽信号ペアの音源信号のみで DNN を学習しており，非常に少ないサンプルからモデルを構築する few-shot learning となっているため，このような過学習を避ける対処が重要となる．それでも，音楽信号で学習した BiLSTM-DPS が音声信号のパーミュテーション問題をある程度解決できたことから，few-shot learning でも実用的な DPS が学習できる可能性が示唆された．

5 まとめ

本稿では，BSS における DPS の検討として，音源の周波数方向の関係性を明確にとらえるために双方向再帰に基づく手法を提案し，パーミュテーション問題の解決性能に関する評価を行った．実験結果より，周波数ビン単位のパーミュテーション問題を解決できる汎用な DPS が省サンプルの音響信号で構築できる可能性が示唆された．

謝辞 本研究の一部は JSPS 科研費 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] 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.

[5] 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.

[6] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying

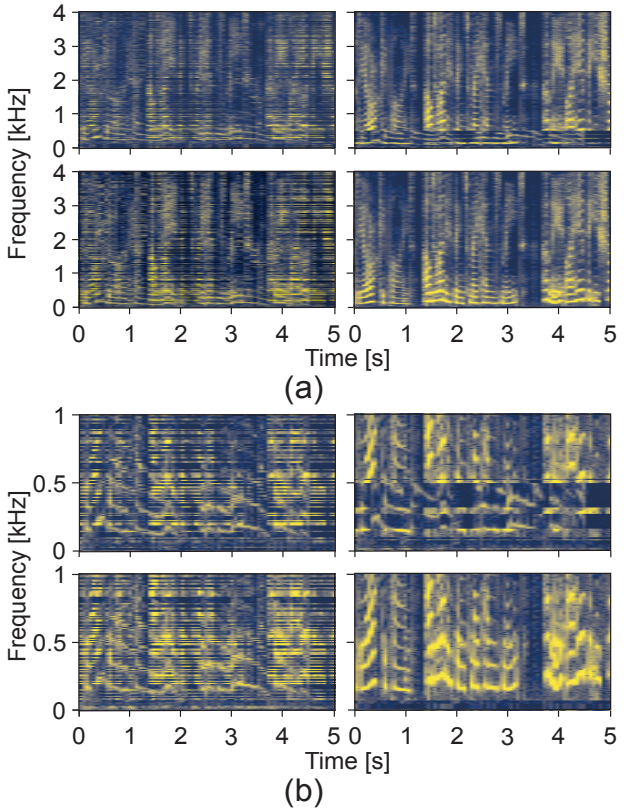

Fig. 5 Spectrograms of female speech signal: input (top-left), conventional (top-right), MLP- (bottom-left), and BiLSTM- (bottom-right) DPSs trained with music signals. (a) and (b) show 0–4 and 0–1 kHz, respectively.

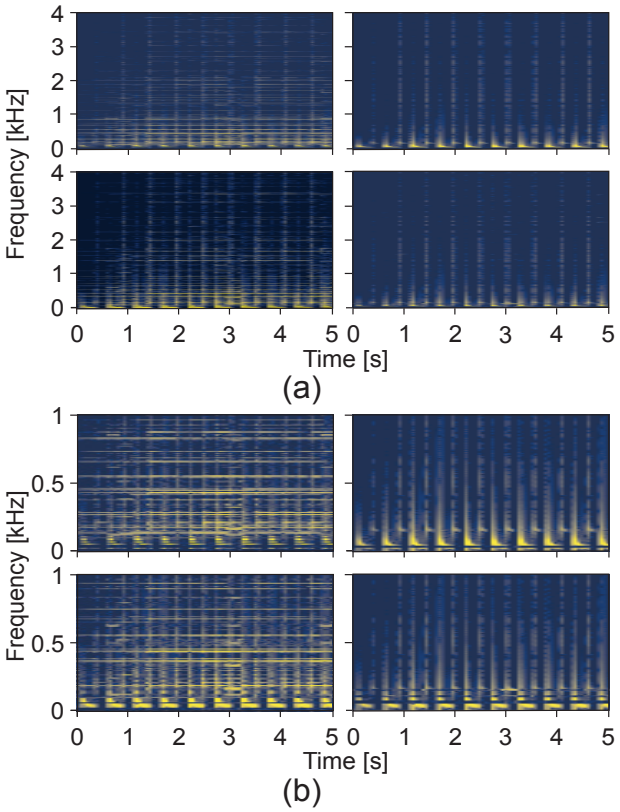

Fig. 6 Spectrograms of drums signal: input (top-left), conventional (top-right), MLP- (bottom-left), and BiLSTM- (bottom-right) DPSs trained with speech signals. (a) and (b) show 0–4 and 0–1 kHz, respectively.

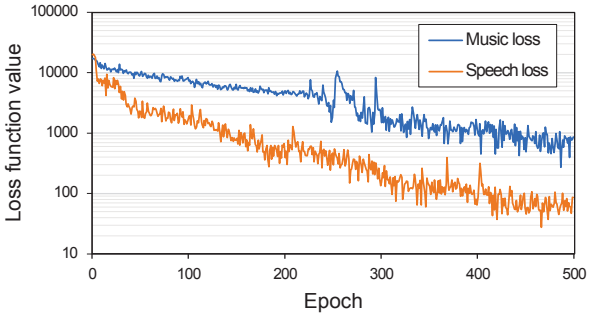

Fig. 7 Behaviors of loss values for BiLSTM-DPS.

independent vector analysis and nonnegative matrix factorization," *IEEE/ACM TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[7] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM TASLP*, vol. 27, no. 10, pp. 1601–1615, 2019.

[8] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[9] 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.*, vol. 48, no. 8, pp. 460–462, 2012.

[10] F. Oshima, M. Nakano, and D. Kitamura, "Interactive speech source separation based on independent low-rank matrix analysis," *Acoustical Science and Technology*, vol. 42, no. 4, pp. 222–225, 2021.

[11] L. Li, H. Kameoka, and S. Seki, "HBP: An efficient block permutation solver using Hungarian algorithm and spectrogram inpainting for multichannel audio source separation," *Proc. ICASSP*, pp. 516–520, 2022.

[12] 山地修平, 中嶋大志, 若林佑幸, 小野順貴, "ハンガリー法を用いたパーミュテーション解法に基づくブラインド音源分離," **日本音響学会** *2021* **年秋季研究発表会講演論文集**, pp. 305–306, 2021

[13] 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.

[14] 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.

[15] 蓮池郁也, 北村大地, 渡辺瑠伊, "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," **日本音響学会** *2022* **年秋季研究発表会講演論文集**, pp. 351–354, 2022.

[16] F. Hasuike, D. Kitamura, and R. Watanabe, "DNN-based frequency-domain permutation solver for multichannel audio source separation," *Proc. APSIPA ASC*, 2022 (in press).

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

[18] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE TNNLS*, vol. 28, no. 10, pp. 2222–2232, 2016.

[19] 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.

[20] 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.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.