一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2022-13(2022-05)

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

蓮池　郁也†　渡辺　瑠伊†　北村　大地†

† 香川高等専門学校，〒761–8058 香川県高松市勅使町 355

あらまし　本稿では，優決定条件ブラインド音源分離手法である周波数領域独立成分分析（FDICA）において生じるパーミュテーション問題について取り扱う．FDICA は，周波数毎に独立成分分析を適用することで音源成分の分離を行うが，パーミュテーション問題と呼ばれる周波数毎の分離信号成分の並び替え問題が生じる．そのため，ポスト処理としてパーミュテーション解決法の適用が必要となる．本稿では，深層ニューラルネットワークに基づく新しいパーミュテーション解決法を提案する．提案手法では，DNN は周波数毎の分離信号成分を並び替えるパーミュテーション行列を予測する．また，基礎的な実験として，パーミュテーション問題を DNN で解くことの妥当性について調査する．
キーワード　独立成分分析，パーミュテーション問題，深層ニューラルネットワーク，ブラインド音源分離

Basic study for permutation solver based on deep neural networks

Fumiya HASUIKE†, Rui WATANABE†, and Daichi KITAMURA†

† National Institute of Technology, Kagawa College, 355 Chokushi, Takamatsu, Kagawa, 761–8058 Japan

Abstract　This paper focuses on a permutation problem associated with frequency-domain independent component analysis (FDICA) that is a technique for (over-)determined blind audio source separation. In FDICA, independent component analysis is applied to each of frequencies, and FDICA encounters the so-called permutation problem, which is a frequency-wise reordering problem of separated source components. Thus, FDICA requires a permutation solver as post processing to obtain the separated source signals. In this paper, we propose a new permutation solver based on a deep neural network (DNN), where DNN predicts a frequency-wise permutation matrix that aligns the order of estimated source components. The validity of using DNN for solving the permutation problem is investigated via basic experiments.
Key words　independent component analysis, permutation problem, deep neural network, blind audio source separation

1. はじめに

音源分離とは，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問題である．特に，マイクロホンや音源の空間的な位置に関する事前情報を用いない音源分離手法をブラインド音源分離（blind source separation: BSS）[1] という．また，マイクロホン数（観測チャネル数）が音源数以上となる収録条件のことを優決定条件と呼び，優決定条件における BSS には独立成分分析（independent component analysis: ICA）[2] に基づく様々な手法がこれまで提案されている．

複数音源の混合は，時間領域では畳み込み混合となるため，ICA を時間領域の観測信号に適用しても BSS は実現されない．そのため，観測信号に短時間 Fourier 変換（short-time Fourier transform: STFT）を適用して時間周波数領域に変換することで，混合系を周波数毎の音源成分の瞬時混合としてモデル化し，周波数毎に ICA を適用することで BSS を達成する手法が提案された．この手法は周波数領域 ICA（frequency-domain ICA: FDICA）[3] と呼ばれ，音響信号の高品質な優決定条件 BSS の達成が期待されるアプローチである．しかしながら，ICA は一般に分離信号の順序が不定であるため，FDICA は分離信号成分の順序が周波数間で不揃いな状態になってしまう問題が生じる．この問題はパーミュテーション問題と呼ばれており，周波数毎の分離信号を正しい順序に並び替えるパーミュテーション解決法が各種検討されてきた [4]〜[6].

その後，混合前の音源信号の時間周波数構造に関する仮定（音源モデル）を導入することで，周波数毎の BSS とパーミュテーション問題の解決を同時に実現する手法が登場した．例えば，音源モデルにグループスパース性を仮定した独立ベクトル分析（independent vector analysis: IVA）[7] やその高速安定アルゴリズムの補助関数法に基づく IVA [8]，非負値行列因子分解 [9] に基づく低ランク良近似性を仮定した独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[10], [11], 音源の

種類毎に音源モデルを深層ニューラルネットワーク（deep neural network: DNN）で事前学習する独立深層学習行列分析 [12]，時間周波数マスクに基づく BSS [13] 等が提案されている．これらの手法は，仮定する音源モデルが実際に混合している各音源によく適合する場合に高い BSS 性能を発揮することが知られている．一方で，文献 [14] では，FDICA に対して理想的な（正解の）パーミュテーション解決を施した BSS の性能が，IVA や ILRMA を大きく上回る例を実験的に示している．これは即ち，音源モデルの良し悪しがパーミュテーション問題の解決精度を極端に左右することを示唆しており，音声や楽器音や雑音等，可能な限り様々な種類の音源信号を表現できる万能な音源モデルが望まれる．但し，そのような万能な音源モデルの構築・学習は決して容易ではない．

そこで近年では，音源モデルではなくパーミュテーション問題の解決そのものを DNN で実現するアプローチ（以後，深層パーミュテーション解決法と呼ぶ）が検討されている．文献 [15] では，局所周波数（サブバンド）毎に，参照周波数の分離信号成分と他の分離信号成分が同じ音源か否かを 2 値分類問題として予測する DNN を学習し，これを用いたパーミュテーション解決法が提案された．この手法はサブバンド毎に予測を行うため，その後のパーミュテーション解決が複雑なアルゴリズムとなっており，特に 3 音源以上の BSS ではさらに極端に複雑化してしまう問題がある．本稿では，3 音源以上でもアルゴリズムが極端に複雑化しないよう，既存手法よりもシンプルな深層パーミュテーション解決法の構築を目指す．

2. FDICA とパーミュテーション問題

2.1 変数の定義

今，N 個の音源信号 $(\tilde{s}_n[t])_{n=1}^N$ が混合し，M 個のマイクロホンで録音された観測信号 $(\tilde{x}_m[t])_{m=1}^M$ を考える．ここで，$n = 1, 2, \cdots, N$，$m = 1, 2, \cdots, M$，及び $t = 1, 2, \cdots, T$ はそれぞれ音源信号，観測信号，及び離散時間のインデクスである．また，BSS によって推定された分離信号を $(\tilde{z}_{n'}[t])_{n'=1}^N$ と表記する．ここで，$n' = 1, 2, \cdots, N$ は分離信号のインデクスであり，$(\tilde{s}_n[t])_{n=1}^N$ と $(\tilde{z}_{n'}[t])_{n'=1}^N$ の音源の順序が必ずしも一致しているとは限らない（$\tilde{s}_1[t]$ の推定成分が $\tilde{z}_2[t]$ となる可能性がある）ことを表すために n と n' を使い分けている．なお，本稿では優決定条件 BSS を取り扱うため，以後 $M = N$ を仮定する[1]．

次に，各信号に STFT を適用して得られる時間周波数領域の信号を次式で表す．

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

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

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

ここで，$i = 1, 2, \cdots, I$ 及び $j = 1, 2, \cdots, J$ はそれぞれ周波数ビン及び時間フレームのインデクスを示す．また，$\cdot^{\mathrm{T}}$ は転置を示す．式 (1)–(3) はいずれも複数音源又は複数チャネルをまとめたベクトルであるが，音源又はチャネルではなく時間周波数でまとめた行列も定義しておく．即ち，n 番目の音源信号の複素スペクトログラム，m 番目の観測信号の複素スペクトログラム，及び n' 番目の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

音響信号の混合は，収録環境の残響の影響等が要因となり，各音源から各マイクロホンまでのインパルス応答が畳み込まれた状態での加算となる．この現象は次式で表される．

$$\tilde{\boldsymbol{x}}[t] = \sum_{n=1}^{N}\sum_{t'=0}^{T'-1} \tilde{\boldsymbol{a}}_n[t']\tilde{s}_n[t-t'] \tag{4}$$

ここで，$\tilde{\boldsymbol{x}}[t] = [\tilde{x}_1[t], \tilde{x}_2[t], \cdots, \tilde{x}_M[t]]^{\mathrm{T}} \in \mathbb{R}^M$ は観測信号ベクトル，T' は残響長，$\tilde{\boldsymbol{a}}_n[t] = [\tilde{a}_{n1}[t], \tilde{a}_{n1}[t], \cdots, \tilde{a}_{nM}[t]]^{\mathrm{T}} \in \mathbb{R}^M$ は音源 n から全マイクロホンまでのインパルス応答ベクトルである．式 (4) で混合される各音源を観測信号から推定する問題は，逆畳み込みフィルタの推定となる．通常の ICA は瞬時混合の逆系を推定する為，式 (4) に対する BSS は不可能である．

そこで，各信号を時間周波数領域で表現して「時間領域での畳み込み混合」を「時間周波数領域での周波数毎の瞬時混合」に変換し，周波数毎に ICA を適用する FDICA [3] が提案された．FDICA では，時間周波数領域の観測信号を次式で表す．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{5}$$

ここで，$\boldsymbol{A}_i = [\boldsymbol{a}_{i1}\ \boldsymbol{a}_{i2}\ \cdots\ \boldsymbol{a}_{iN}] \in \mathbb{C}^{M\times N}$ は周波数毎の時不変混合行列である．この混合モデルは，STFT の短時間区間長が式 (4) の残響長より十分長い場合に良く成立する．

混合行列 $\boldsymbol{A}_i$ が全周波数において正則であれば，周波数毎の分離行列 $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = [\boldsymbol{w}_{i1}\ \boldsymbol{w}_{i2}\ \cdots\ \boldsymbol{w}_{iN}]^{\mathrm{H}} \in \mathbb{C}^{N\times M}$ 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．

$$\boldsymbol{z}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{6}$$

ここで，$\cdot^{\mathrm{H}}$ はエルミート転置を示す．分離行列 $\boldsymbol{W}_i$ の行ベクトルである $\boldsymbol{w}_{in} \in \mathbb{C}^M$ は，i 番目の周波数ビンにおいて，観測信号から n 番目の音源の分離信号を推定する分離フィルタである．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 を適用することで，周波数毎の分離行列 $\boldsymbol{W}_i$ を全周波数にわたって推定し，BSS の実現を目指す．

2.3 パーミュテーション問題とその解決

FDICA 中で周波数毎に適用している ICA は，その分離原理より，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．

$$\hat{\boldsymbol{W}}_i = \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{W}_i \tag{7}$$

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

(注1)：但し，各成分の物理的な意味を明らかにするために，インデクス n（又は n'）及び m や N 及び M は常に使い分けて表記する．

（置換行列）である．例えば，$N = 2$ であれば $\boldsymbol{P}_i$ は

$$\boldsymbol{P}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{8}$$

の 2 通りの内のいずれかを取り，$N = 3$ であれば

$$\boldsymbol{P}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \text{ or } \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{9}$$

の 6 通りの内のいずれかを取る．パーミュテーション行列はその性質から，各行の総和及び各列の総和が全て 1 となるため，二重確率行列（doubly stochastic matrix: DSM）[16] である．

従って，FDICA で得られる信号は，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な状態である．

$$\boldsymbol{y}_{ij} = \hat{\boldsymbol{W}}_i \boldsymbol{x}_{ij} \tag{10}$$

$$= \left[y_{ij1}, y_{ij2}, \cdots, y_{ijn'_i}, \cdots, y_{ijN} \right]^{\mathrm{T}} \in \mathbb{C}^N \tag{11}$$

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．このうち，$\boldsymbol{D}_i$ 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] で解析的に復元可能である．一方で，$\boldsymbol{P}_i$ 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）することは，組み合わせ爆発が生じるため容易ではなく，一般に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（パーミュテーション問題が生じている状態の）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．FDICA 直後の $\boldsymbol{Y}_{n'}$ は，理想的には周波数毎での音源分離が達成できており，周波数間の音源順序の不整合のみ生じている．そのため，ポスト処理として，音源順序を全周波数ビンにわたって整列する必要がある．

理想的なパーミュテーション問題の解決は次式で表される．

$$z_{ij} = \boldsymbol{P}_i^{-1} \boldsymbol{D}_i^{-1} \boldsymbol{y}_{ij} \tag{12}$$

従って，パーミュテーション問題の解決とは，全周波数ビンにわたって $\boldsymbol{P}_i^{-1}$ を求める問題と解釈できる．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．

$$z_{ij} = \boldsymbol{P}_{\mathrm{all}} \boldsymbol{P}_i^{-1} \boldsymbol{D}_i^{-1} \boldsymbol{y}_{ij} \tag{13}$$

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不定性については問題とせず，$\boldsymbol{P}_{\mathrm{all}}^{-1}$ は推定しない．

3. 提 案 手 法

3.1 提案手法の動機

文献 [15] で提案された既存の深層パーミュテーション解決法は，あるサブバンド内でパーミュテーション問題を解決する DNN を学習し，これを様々なサブバンドに適用した結果を用いて全周波数のパーミュテーション問題を解決する手法である．サブバンド単位の処理では，サブバンドの中心周波数を参照周波数ビンと定義したうえで，その他の（サブバンド内の）周波数ビンが参照周波数ビンの推定信号成分と同一音源の成分か否かを DNN で予測する．音源数が $N = 2$ であれば，この「同一音源の成分か否か」という 2 クラス分類問題は「どちらの音源の成分か」に一致する．しかし，音源数が $N \geq 3$ となった場合，仮に「同一音源の成分ではない」と DNN が予測した際に，残りのどの音源の成分かが確定しない．従って，この場合に推定信号成分の音源順序を確定させるためには，前述の 2 クラス分類 DNN を音源数 N 個の中から 2 つ選ぶ組み合わせ数（${}_N C_2$）回適用しなければならない．さらに，既存手法におけるサブバンド間のパーミュテーション問題の解決処理（全サブバンドの予測結果のスティッチングによる統合）を考えると，そのアルゴリズムは音源数の増加に伴って極端に複雑化してしまう．

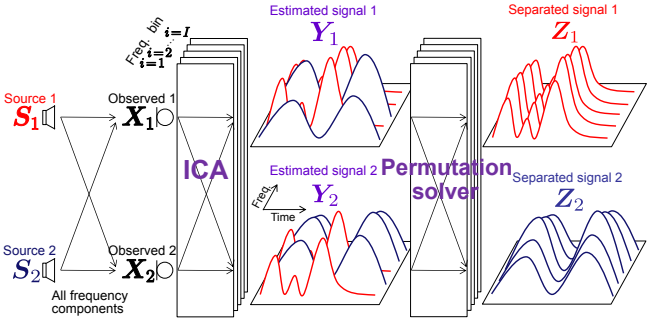

Fig. 1: Permutation problem in FDICA ($N = M = 2$).

そこで本稿では，より簡潔なアルゴリズムとして，2 クラス分類ではなく N クラス分類の DNN を学習し，サブバンド単位ではなく全周波数の音源順序を一度に予測する手法を検討する．次節以降で述べる通り，この N クラス分類 DNN は式 (13) 中のパーミュテーション行列 $\boldsymbol{P}_i^{-1}$ を直接予測することに対応しており，既存手法よりも見通しの良いアルゴリズムとなっている．なお提案手法は，音源数 N の増加に対してアルゴリズムが極端に複雑化しない手法として提案するが，本稿は基礎的検討に終始するため，以後 $N = 2$ の場合のみを取り扱う．次節以降の議論は，$N \geq 3$ についても一般性を失うことなく応用できる．

3.2 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 $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ が得られる．提案手法ではまず，これらのパワースペクトログラム $(|\boldsymbol{Y}_1|^{\cdot 2}, |\boldsymbol{Y}_2|^{\cdot 2})$ に次式を適用し，正規化パワースペクトログラム $(\overline{\boldsymbol{Y}}_1, \overline{\boldsymbol{Y}}_2)$ を得る．

$$\overline{\boldsymbol{Y}}_{n'} = \frac{|\boldsymbol{Y}_{n'}|^{\cdot 2}}{|\boldsymbol{Y}_1|^{\cdot 2} + |\boldsymbol{Y}_2|^{\cdot 2}} \in [0, 1]^{I \times J} \tag{14}$$

ここで，行列に対する絶対値記号，ドット付き指数乗，及び分数はそれぞれ要素毎の絶対値，要素毎の指数乗，及び要素毎の商を表す．この正規化は，同一音源に属する成分の相関を強調する利点がある [6] だけでなく，推定信号の値が区間 $[0, 1]$ の範囲に制限され，DNN の学習が安定する効果も期待できる．

次に，正規化パワースペクトログラム $(\overline{\boldsymbol{Y}}_1, \overline{\boldsymbol{Y}}_2)$ 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 $(\check{\boldsymbol{Y}}_{j1}, \check{\boldsymbol{Y}}_{j2})$ を抽出する．

$$\check{\boldsymbol{Y}}_{jn'} = [\overline{\boldsymbol{y}}_{(j-\beta)n'} \ \ \overline{\boldsymbol{y}}_{(j-\beta+1)n'} \ \ \cdots \ \ \overline{\boldsymbol{y}}_{(j+\beta)n'}] \in [0,1]^{I\times(2\beta+1)} \tag{15}$$

ここで，$\overline{\boldsymbol{y}}_{jn'} \in [0,1]^I$ は正規化パワースペクトログラム $\overline{\boldsymbol{Y}}_{n'}$ 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．

提案手法の DNN の入力ベクトルは，式 (15) で得られる両信号の局所時間パワースペクトログラム $(\check{\boldsymbol{Y}}_{j1}, \check{\boldsymbol{Y}}_{j2})$ を一次元にベクトル化したものである．入力された行列をベクトル化する処理を vec(·) と表記すると，DNN の入力ベクトルは次式となる．

$$\boldsymbol{d}_j = \begin{bmatrix} \mathrm{vec}(\check{\boldsymbol{Y}}_{j1}) \\ \mathrm{vec}(\check{\boldsymbol{Y}}_{j2}) \end{bmatrix} \in [0,1]^{2I(2\beta+1)} \tag{16}$$

DNN による予測を次式で表す．

$$\hat{\boldsymbol{l}}_j = \mathrm{DNN}(\boldsymbol{d}_j) \in [0,1]^{2I} \tag{17}$$

ここで，$\hat{\boldsymbol{l}}_j = [\hat{l}_{11j}, \hat{l}_{21j}, \cdots, \hat{l}_{I1j}, \hat{l}_{12j}, \hat{l}_{22j}, \cdots, \hat{l}_{I2j}]^{\mathrm{T}}$ は DNN の出力（予測ベクトル）を表す．入力されたベクトルを適切に行列化する処理を mat(·) と表記すると，予測ベクトルは次式で $I \times N$ の行列に再成型される．

$$\hat{\boldsymbol{L}}_j = \mathrm{mat}(\hat{\boldsymbol{l}}_j) \in [0,1]^{I\times 2} \tag{18}$$

式 (18) で得られる行列 $\hat{\boldsymbol{L}}_j$ は，Fig. 2 に示すように，パーミュテーション問題が生じている入力信号 $(\check{\boldsymbol{Y}}_{j1}, \check{\boldsymbol{Y}}_{j2})$ の各周波数成分の順序に対して，「順序をそのままにすべきである真の確率 l_{i1}」と「順序を反転させるべきである真の確率 l_{i2}」$(l_{i1}, l_{i2} \in \{0,1\})$ を $\boldsymbol{d}_j$ から予測したものと定義する[2]．即ち，式 (7) 中の $\boldsymbol{P}_i$ で生じてしまうパーミュテーション問題を解決するような真のパーミュテーション行列 $\boldsymbol{P}_i^{-1}$ は，真値 l_{i1} 及び l_{i2} を係数とする凸結合で次式のように表せる．

$$\boldsymbol{P}_i^{-1} = l_{i1}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + l_{i2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in \{0,1\}^{N\times N} \tag{19}$$

なお，音源数が $N = 3$ のときは，式 (9) の 6 通りのパーミュテーション行列に対する真の確率値を $(l_{in'_i})_{n'_i=1}^{6}$ と定義すれば，一般性を失うことなく同様に議論できる．音源数が $N \geq 4$ でも同様である．真値 (l_{i1}, l_{i2}) は確率値であるため，それらの予測値である $(\hat{l}_{i1j}, \hat{l}_{i2j})$ もまた $\hat{l}_{i1j}, \hat{l}_{i2j} \in [0,1]$ かつ $\hat{l}_{i1j} + \hat{l}_{i2j} = 1$ を満たすように DNN の中で制約する．この予測値 $(\hat{l}_{i1j}, \hat{l}_{i2j})$ から，推定パーミュテーション行列を次式のように構成できる．

$$\hat{\boldsymbol{P}}_{ij}^{-1} = \hat{l}_{i1j}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \hat{l}_{i2j}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in [0,1]^{N\times N} \tag{20}$$

(注2)：実際は l_{i1} 及び l_{i2} の定義は逆でもよい．この定義の順序は分離音源全体の順序にのみ影響するため，$\boldsymbol{P}_{\mathrm{all}}^{-1}$ を推定しない本稿の問題設定では任意である．また，3.4 節に示す方法により，この任意性を残したまま DNN を学習できる．

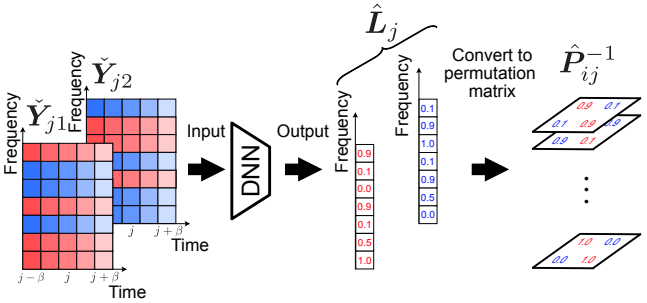

Fig. 2: Calculation of predicted permutation matrix.

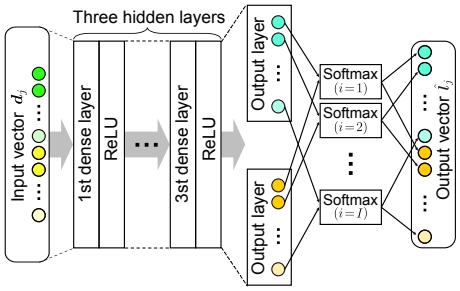

Fig. 3: DNN architecture.

従って，提案手法の DNN は，推定音源の各周波数成分を正しい順序に並び替えるようなパーミュテーション行列を予測する多クラス分類モデルと解釈できる．さらに，式 (20) はその定義より DSM であるため，Birkhoff–von Neumann の定理（付録参照）より，DNN は式 (8) や式 (9) のような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 に提案手法で用いる 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．また，隠れ層の 3 層目から出力層に変換する際には，Fig. 3 に示すように 2 つの I 次元ベクトルに分岐させている．この時の各ベクトルへの変換パラメータは独立している[3]．その後，2 つの I 次元の同一インデクスの要素に softmax 関数を適用することで，予測ベクトルの全要素が閉区間 [0, 1] 内の値かつ同一インデクスの要素の和が 1 となることを保証している．これは，前節で説明した $\hat{l}_{i1} + \hat{l}_{i2} = 1$ の制約を保証することに対応する．

3.4 DNN の損失関数

式 (20) で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 $(\check{\boldsymbol{Y}}_{j1}, \check{\boldsymbol{Y}}_{j2})$ は式 (15) と同様の手法で抽出した $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ の局所時間複素スペクトログラムである．まず，式 (12) と同様の処理で $(\check{\boldsymbol{Y}}_{j1}, \check{\boldsymbol{Y}}_{j2})$ の音源パーミュテーションを並び替えた予測分離信号 $(\widehat{\check{\boldsymbol{Z}}}_{j1}, \widehat{\check{\boldsymbol{Z}}}_{j2})$ を求める．次に，予測分離信号に対する正解ラベル $(\check{\boldsymbol{Z}}_{j1}, \check{\boldsymbol{Z}}_{j2})$（分離信号 $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ の局所時間複素スペクトログラム）を用意する．提案手法では，予測分離信号 $(\widehat{\check{\boldsymbol{Z}}}_{j1}, \widehat{\check{\boldsymbol{Z}}}_{j2})$ とラベル $(\check{\boldsymbol{Z}}_{j1}, \check{\boldsymbol{Z}}_{j2})$ の間の平均二乗誤差（mean squared error: MSE）を損失関数に用いる．ここで，2.3 節で述べた通り，提案手法は音源信号全体の順序の不定性の解決（$\boldsymbol{P}_{\mathrm{all}}^{-1}$ の推定）を目的としない．即ち，提案手法を適用した結果が，$(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ 及び $(\boldsymbol{Z}_2, \boldsymbol{Z}_1)$ のどちらで

(注3)：即ち $2I$ 次元への全結合層による変換と等価だが，明示的に分岐させる．

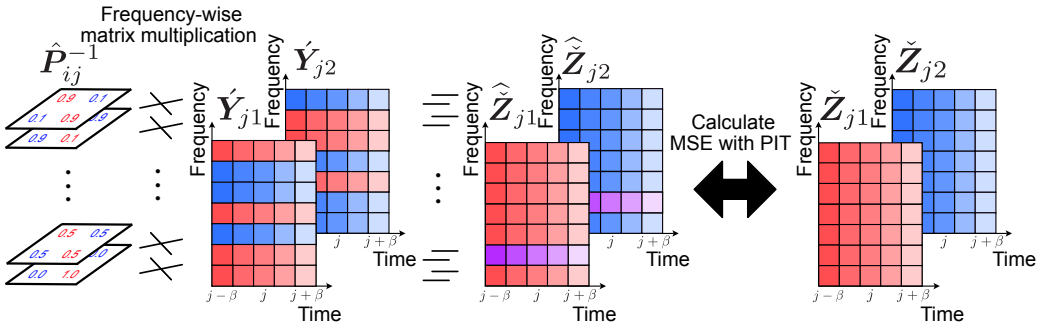

Fig. 4: Calculation of MSE with PIT.

出力されようとも構わない．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）[18] を導入した損失関数 $\mathcal{L}$ を用いる．

$$\mathcal{L} = \min\left(L_1, L_2\right) \tag{21}$$

$$L_1 = ||\hat{\tilde{\boldsymbol{Z}}}_{j1} - \check{\boldsymbol{Z}}_{j1}||_2^2 + ||\hat{\tilde{\boldsymbol{Z}}}_{j2} - \check{\boldsymbol{Z}}_{j2}||_2^2 \tag{22}$$

$$L_2 = ||\hat{\tilde{\boldsymbol{Z}}}_{j1} - \check{\boldsymbol{Z}}_{j2}||_2^2 + ||\hat{\tilde{\boldsymbol{Z}}}_{j2} - \check{\boldsymbol{Z}}_{j1}||_2^2 \tag{23}$$

ここで，$\min(\cdot,\cdot)$ は複数のスカラー引数の中で最小値を返す処理を表す．この PIT の導入により，DNN は周波数間の音源順序の整列にのみ注力し，分離信号全体の順序には依存しない学習が実現できる．

3.5 学習済の DNN のテストデータへの適用

DNN 学習後は，提案手法を FDICA の推定信号 $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ に適用できる．このとき，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（$\boldsymbol{P}_i^{-1}$ は j に非依存）．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 $(\check{\boldsymbol{Y}}_{j1}, \check{\boldsymbol{Y}}_{j2})$ を DNN に入力し，その各々から予測される推定局所時間パーミュテーション行列 $(\hat{\boldsymbol{P}}_{ij}^{-1})_{j=1}^{J}$ を多数決処理して時不変な行列 $\hat{\boldsymbol{P}}_i$ に変換することで，更なる精度向上が期待できる．この時間方向の多数決処理は次式で表せる．

$$\hat{\boldsymbol{P}}_i^{-1} = \mathrm{round}\left(\frac{1}{J}\sum_{j=1}^{J}\hat{\boldsymbol{P}}_{ij}^{-1}\right) \in \{0,1\}^{N\times N} \tag{24}$$

ここで，$\mathrm{round}(\cdot)$ は入力された行列を要素毎に四捨五入する処理を表す．最終的な推定分離信号は次式で得られる．

$$\hat{\boldsymbol{z}}_{ij} = \hat{\boldsymbol{P}}_i^{-1}\boldsymbol{y}_{ij} \tag{25}$$

なお，推定分離信号 $\hat{\boldsymbol{z}}_{ij}$ を時間周波数でまとめた行列を $\hat{\boldsymbol{Z}}_{n'} \in \mathbb{C}^{I\times J}$ と定義する．

4. 実　　験

4.1 実験条件

本稿では，基礎的な実験として，パーミュテーション問題の解決に DNN を用いることの妥当性について調査する．この実験では，音源信号 $(\boldsymbol{S}_1, \boldsymbol{S}_2)$ として Table 1 に示す男女の音声信号（Female speech 及び Male speech）又はドラムとピアノの音楽信号（Piano 及び Drums）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．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．式 (15) で抽出する局所時間フレーム数は $2\beta + 1 = 27$ と設定した．

Table 1: Speech and music sources obtained from SiSEC2011

Signal type	Data name	Length [s]
Female speech	`dev3_female4_inst_src_2`	10.0
Male speech	`dev2_male4_inst_src_2`	10.0
Piano	`dev2_nodrums_liverec_250ms_src_3`	11.0
Drums	`dev2_wdrums_liverec_250ms_src_3`	11.0

本実験では，IVA や ILRMA 等で生じることの多いブロックパーミュテーション問題 [19] を模擬した．これは，ある程度まとまった周波数帯域（ブロック）単位でパーミュテーション問題が発生する現象である．具体的には，各ブロックの周波数ビン数を 16 として $(\boldsymbol{S}_1, \boldsymbol{S}_2)$ 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 $\boldsymbol{S}_1$ 及び $\boldsymbol{S}_2$ の成分を入れ替えることで FDICA の推定信号 $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ を模擬した．DNN の学習データは，各実験条件の推定信号 $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$（入力）及び分離信号 $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$（ラベル）の局所時間スペクトログラムを用いた．但し，入力はパーミュテーション問題を模擬するランダム入れ替えを重複無しの 300 パターンで生成したものであり，学習エポック数は 1000 回とした．性能評価に用いるテストデータは同一の入力だが，学習データには含まれていないランダム入れ替え 1 パターンで生成した $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ を用いた．評価指標には，テストデータに対する周波数毎の並び替えの正答率及び信号対歪み比（source-to-distortion ratio: SDR）[20] の改善量を用いた．

4.2 実験結果

Figs. 5 及び 6 はそれぞれ音声信号及び音楽信号でブロックパーミュテーション問題を模擬したスペクトログラム $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ 及び提案手法でパーミュテーション問題を解決した結果 $(\hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2)$ を示している．$(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ はいずれも周波数方向に一貫性のないスペクトログラムであるが，$(\hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2)$ は各周波数成分に連続性が確認でき，高い精度でパーミュテーション問題を解決できていることが分かる．

DNN のテストデータに対する周波数毎の並び替えの正答率は，音声信号では 92.5%，音楽信号では 97.5% であった．また，SDR の改善量は Female speech 及び Male speech がそれぞれ 23.8 dB 及び 24.7 dB，Piano 及び Drums がそれぞれ 27.0 dB 及び 17.0 dB であり，この結果は提案手法におけるパーミュテーション問題の解決精度を客観的に示している．

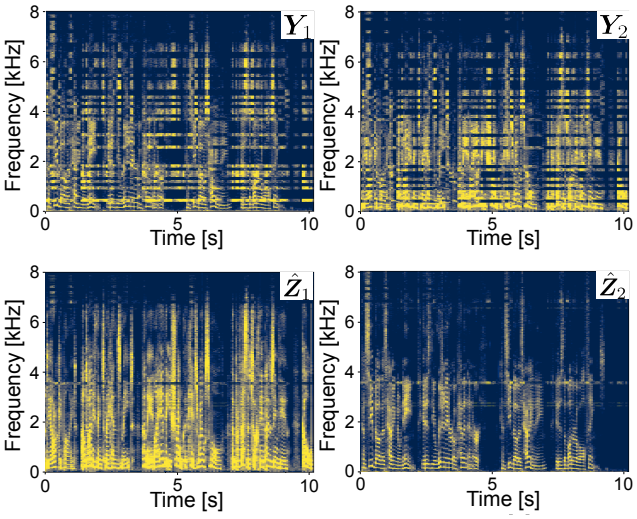

Fig. 5: Input speech spectrograms with simulated permutation problem (upper) and permutation-aligned speech spectrograms (bottom).

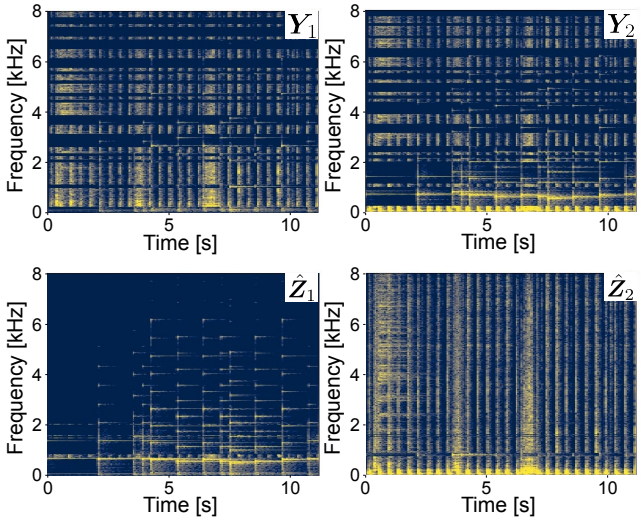

Fig. 6: Input music spectrograms with simulated permutation problem (upper) and permutation-aligned music spectrograms (bottom).

5. ま と め

本稿では，FDICA におけるパーミュテーション問題を DNN で解決する手法を提案し，その妥当性について基礎的な検討を示した．実験結果より，学習データとテストデータが同一信号という理想的な条件では，DNN はブロックパーミュテーション問題を解決するようなパーミュテーション行列を十分な精度で予測できることが明らかとなった．今後の課題として，様々な音響信号から学習した DNN で未知の音響信号のパーミュテーション問題を解決できるかの調査や，周波数方向の関連を明確に学習するための双方向再帰型 DNN の活用等が挙げられる．

付 録

［定理 1］ Birkhoff–von Neumann の定理

二重確率行列 $\boldsymbol{A}$ が与えられたとき，$\boldsymbol{A}$ は同じサイズのパーミュテーション行列 $\{\boldsymbol{P}_n\}_{n=1}^{N!}$ の凸結合で表せる．即ち，凸結合係数 σ_n（$\sigma_n \geq 0$ かつ $\sum_n \sigma_n = 1$）を用いて $\boldsymbol{A} = \sum_n \sigma_n \boldsymbol{P}_n$ と書ける．

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。

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 Trans. Signal and Info. Process.*, 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. 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.

[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. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 666–678, 2006.

[6] 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.

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

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

[9] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[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 Trans. Audio, Speech, Lang. Process.*, 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] 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 Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, 2019.

[13] K. Yatabe and D. Kitamura, “Determined BSS based on time-frequency masking and its application to harmonic vector analysis,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1609–1625, 2021.

[14] D. Kitamura, N. Ono, and H. Saruwatari, “Experimental analysis of optimal window length for independent low-rank matrix analysis,” *Proc. EUSIPCO*, pp. 1210–1214, 2017.

[15] 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.

[16] A. Horn, “Doubly stochastic matrices and the diagonal of a rotation matrix,” *Am. J. Math.*, vol. 76, no. 3, pp. 620–630, 1954.

[17] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” *Proc. ICA*, pp. 722–727, 2001. *Proc. ICML*, 2010.

[18] 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.

[19] 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.

[20] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.