深層パーミュテーション解決法の基礎的検討

Basic study for deep permutation solver

蓮池 郁也
Fumiya Hasuike

1　はじめに

音源位置やマイクロフォン位置等が未知の条件で音源分離を達成する技術はブラインド音源分離（BSS）と呼ばれる．観測信号のチャネル数（マイクロフォン数）と混合されている音源数が等しい条件下では，観測信号を時間周波数領域に変換し周波数毎に独立成分分析（ICA）[1] を適用する時間周波数領域 ICA（FDICA）[2] が提案されている．ICA は一般に推定分離信号の順番が不定であり，FDICA は Fig. 1 に示すように周波数毎に独立な ICA による BSS を行うため，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生じる．この問題は一般に「パーミュテーション問題」と呼ばれている．過去様々な解決法が提案されたが，あらゆる音源に適用可能な手法は未だ存在しない．

本稿では，様々な音源に適用可能なパーミュテーション解決法の構築を目的として，DNN の活用を検討する．また，提案する深層パーミュテーション解決法を音声及び音楽信号に適用し，深層パーミュテーション解決法の実現可能性を調査する．

2　FDICA とパーミュテーション問題

ICA は音源間の統計的独立性のみに基づいて分離行列を推定するため，周波数毎に独立な ICA を適用している FDICA で推定される分離信号

$$\boldsymbol{y}_{ij} = \hat{\boldsymbol{W}}_i \boldsymbol{x}_{ij} \tag{1}$$

は Fig. 1 に示すように，推定音源の順番が周波数毎にばらばらになっている状態である．ここで，$i = 1, 2, \cdots, I$ 及び $j = 1, 2, \cdots, J$ はそれぞれ周波数及び時間のインデクス，$\boldsymbol{y}_{ij}$，$\boldsymbol{x}_{ij}$，及び $\boldsymbol{W}_i$ はそれぞれ分離信号，観測信号，及び分離行列を示す．また Fig. 1 に示すように，音源信号，観測信号，FDICA の推定信号，及びパーミュテーション問題解決後の信号の各時間周波数行列をそれぞれ $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$，$\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$，及び $\boldsymbol{Z}_n \in \mathbb{C}^{I\times J}$ と定義する（$n = 1, 2, \cdots, N$ 及び $m = 1, 2, \cdots, M$ はそれぞれ音源及びマイクロホンのインデクス）．FDICA で得られる $\boldsymbol{Y}_n$ は，周波数毎の音源分離はできており，周波数間の推定音源の順序が統一されていない状態である．これがパーミュテーション問題であり，全周波数の推定音源が同じ順序となるよう並び替える処理が必要となる．

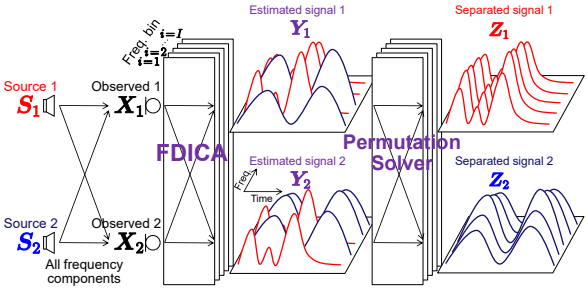

Fig. 1　Permutation problem in FDICA.

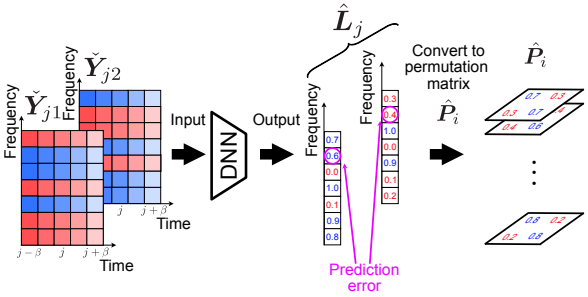

Fig. 2　Calculatation of predicted permutation matrix.

3　提案手法

3.1　DNN の入出力

本稿では $N = M = 2$ の状況のみを取り扱う．DNN の入力の前処理として，FDICA の推定信号 $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ にパワー比の正規化 [3] と時間フレーム j 近傍のスペクトルの抽出を施す．この前処理を適用した後の信号を $(\check{\boldsymbol{Y}}_{j1}, \check{\boldsymbol{Y}}_{j2})$ と定義し，これらをベクトル化し DNN に入力する．DNN の出力は周波数毎の確率値 $(\hat{l}_{i1j}, \hat{l}_{i2j})$ であり．$\hat{l}_{i1j} + \hat{l}_{i2j} = 1$ を満たす．この $\hat{l}_{inj}$ は入力の各周波数成分が n 番目の音源の成分である確率を表しており，2 音源であれば次式のようにパーミュテーション行列 $\hat{\boldsymbol{P}}_i$ に変換できる．

$$\hat{\boldsymbol{P}}_i = \begin{bmatrix} \hat{l}_{i1j} & \hat{l}_{i2j} \\ \hat{l}_{i2j} & \hat{l}_{i1j} \end{bmatrix} \tag{2}$$

上記の処理を Fig. 2 に示す．このとき，パーミュテーション問題は次式で解決される．

$$\hat{\boldsymbol{z}}_{ij} = \hat{\boldsymbol{P}}_i \boldsymbol{y}_{ij} \tag{3}$$

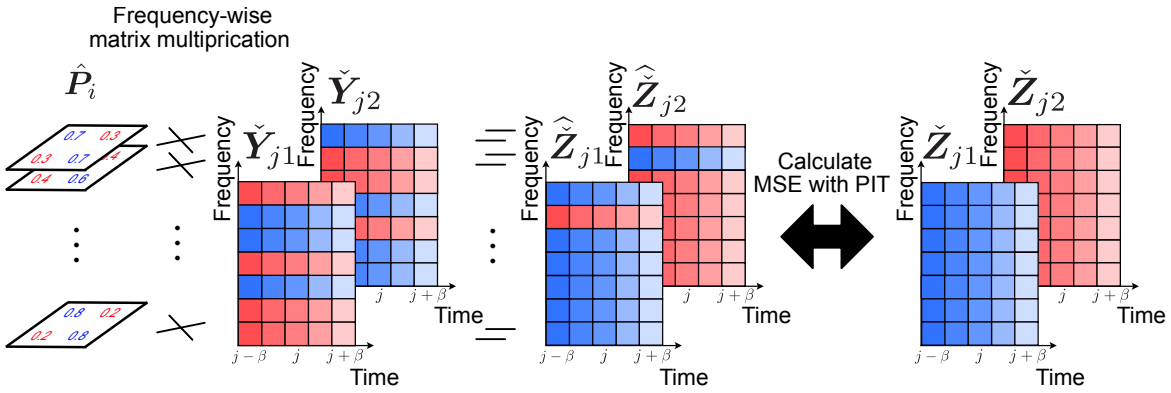

Fig. 3　Calculation of MSE with PIT.

なお，式 (3) で得られる予測分離信号の時間周波数行列を $(\hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2)$ と定義する．

3.2　DNN の構造と損失関数

提案手法で用いる DNN の構造は，入力層，隠れ層 3 層，及び出力層の計 5 層の全結合層からなる多層パーセプトロンである．また損失関数は，$\hat{\boldsymbol{P}}_i$ に従って $(\check{\boldsymbol{Y}}_{j1}, \check{\boldsymbol{Y}}_{j2})$ を並び替えた局所時間分離信号 $(\check{\boldsymbol{Z}}_{j1}, \check{\boldsymbol{Z}}_{j2})$ と局所時間予測分離信号 $(\widehat{\check{\boldsymbol{Z}}}_{j1}, \widehat{\check{\boldsymbol{Z}}}_{j2})$ 間の平均二乗誤差を順序不変学習（PIT）[4] として定義している．この一連の流れを Fig. 3 に示す．DNN はこの処理により得た損失値を用いて誤差逆伝播を行い，最適なモデルを学習する．

4　実験

4.1　実験条件

本稿では，実際の音声及び音楽信号の時間周波数行列にパーミュテーション問題を起こし，提案手法でどの程度解決できるか調査した．このとき，学習データとして 300 パターンのパーミュテーション問題を生じさせた $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ を用意し，検証及びテストデータは学習データに含まれないパーミュテーション問題を用いた．音声及び音楽信号はそれぞれ男女の英語発話及びドラムとピアノの音楽を用いた．パーミュテーション問題の解決性能は信号対歪み比（SDR）の改善量で評価する．SDR とは，音源分離の度合と分離音の歪みの少なさの両方を加味した客観評価尺度である．

4.2　実験結果

Fig. 4 は実際の男女の音声信号に対するパーミュテーション問題が生じているスペクトログラム $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ 及び予測結果 $(\hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2)$ を示している．$(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ は，各周波数毎に男女の信号が混ざっており一貫性のないスペクトログラムとなっているが，$(\hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2)$ は，各周波数成分に連続性が見られる．即ち，男女のどちらの信号も高い精度でスペクトログラムを予測できていることが分かる．DNN の学習データ及び検証データに対する正答率は 90% を超えており，ほとんどの周波数成分を正しく並び替えで

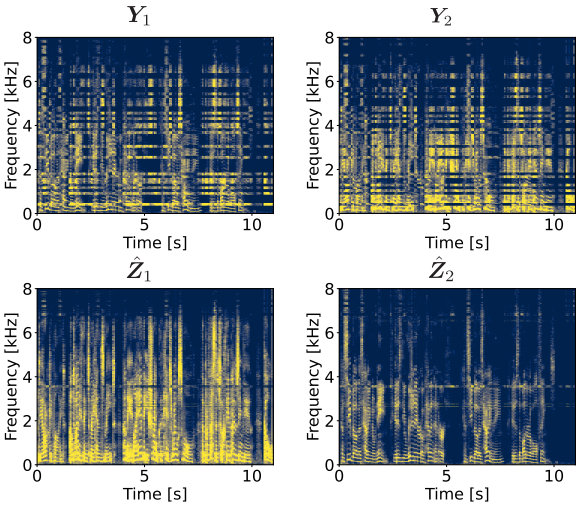

Fig. 4　Input spectrograms with permutation problem (upper) and permutation-aligned spectrograms using predicted results (bottom).

きていることが分かる．SDR の改善量は，$(\hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2)$ に対してそれぞれ 26.7 dB，及び 31.0 dB であり，提案手法による分離精度が高いことを示している．

5　まとめ

本稿では，FDICA に伴うパーミュテーション問題の解決を目的とし，深層パーミュテーション解決法を新たに提案した．実験結果より，DNN はパーミュテーション問題の解決に有効である可能性が示唆された．

参考文献

[1] P. Comon, “Independent component analysis, a new concept?,” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
[2] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.
[3] H. Sawada, S. Araki, and S. Makino, “Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS,” *Proc. IEEE International Symposium on Circuits and Systems*, pp. 3247–3250, 2007.
[4] D. Yu, M. Kolbak, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 241-245, 2017.