局所時間周波数構造に基づく深層パーミュテーション解決法*

☆山地修平, 北村大地 (香川高専)

1 はじめに

ブラインド信号源分離（blind source separation: BSS）とは，複数の信号源が混合した観測信号から，混合前の信号源を推定する技術である．特に独立成分分析（independent component analysis: ICA）[1] の登場以降，音源信号 ≥ マイクロホン数である優決定条件下の信号源分離問題に広く適用されている．

音響信号の混合問題では一般的に残響がかかることから，周波数領域での瞬時混合を仮定し ICA を周波数領域に拡張した手法が周波数領域 ICA（frequency-domain ICA: FDICA）[2] である．しかし，ICA は一般に推定分離信号の順番が不定であり，FDICA は周波数毎に分離を行うため，周波数毎の分離信号の順番がバラバラになってしまう. これを正しい順番に並び替える問題は一般にパーミュテーション問題と呼ばれており，過去には隣接周波数の時系列強度（音源アクティベーション）の相関を用いたパーミュテーション解決法 [3]，マイクロホンの相対的な位置情報を既知として音源到来方位を計算し，パーミュテーション解決の手掛かりとする手法 [4]，及びその両者を組み合わせた手法 [5] が提案されている．また，近年では FDICA に対して音源の時間周波数成分の共起関係を新たに仮定して，パーミュテーション問題を起こさず分離信号を推定する手法が登場している．例えば，独立ベクトル分析（independent vector analysis: IVA）[6, 7] は，同一音源の周波数成分の共起を仮定しており，非負値行列因子分解 [8] と IVA を組み合わせた独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[9, 10] は同一音源の時間周波数成分の共起が低ランク構造を持つことを仮定している．さらに，深層学習（deep neural networks: DNN）を用いて音源の時間周波数構造の仮定を学習し，FDICA に適用する独立深層学習行列分析 [11] も提案されている．

しかしながら，パーミュテーション問題の解は組み合わせ爆発を起こすことから，上記いずれの手法を用いても完璧にパーミュテーション問題を解くことは非常に難しく，とくに複数音声の混合信号における高精度なパーミュテーション問題の解決はいまだできていない．例えば文献 [12] では，複数音声の混合信号の分離時に正解のパーミュテーションを与えた FDICA が，ブラインドな IVA や ILRMA よりも非常に高い分離精度を達成することを実験的に示している．

本稿では，近年目覚ましい発展を遂げている DNN を用いてパーミュテーション問題が解決できるか否かについて検討する．ここでは，優決定条件下での複数音声の混合を対象とし，FDICA で周波数毎の分離信号を求めた後に DNN に基づくパーミュテーション解決法を適用することを考える．なお，劣決定音源分離において周波数毎のフルランク空間相関行列を推定する BSS [13] 等でもパーミュテーション問題を解く必要が生じるため，提案手法は FDICA だけでなく，他の手法におけるパーミュテーション問題にも適用できる．

2 FDICA とパーミュテーション問題

2.1 定式化

音源数と観測チャネル数（マイクロホン数）をそれぞれ N 及び M とし，各時間周波数における音声信号，混合信号，及び分離信号をそれぞれ

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

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

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

と表す．ここで，$i = 1, \cdots, I$ は周波数ビンインデクス，$j = 1, \cdots, J$ は時間フレームインデクス，$n = 1, \cdots, N$ は音源インデクス，$m = 1, \cdots, M$ はチャネルインデクスを示し，$^{\mathrm{T}}$ は転置を表す．また，各信号の複素スペクトログラム行列を $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$，及び $\boldsymbol{Z}_n \in \mathbb{C}^{I\times J}$ で表す．混合信号の混合系が線形時不変であり，時間周波数領域での複素瞬時混合で表現できる場合，周波数毎の時不変な複素混合行列 $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \ \cdots \ \boldsymbol{a}_{i,N}) \in \mathbb{C}^{M\times N}$ を定義することで，混合信号を次式で表現できる．

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

ここで，$\boldsymbol{a}_{i,n}$ は音源 n のステアリングベクトルである．この混合モデルは，混合信号の残響時間が 短時間フーリエ変換（short-time Fourier transform: STFT）の窓長よりも十分短い場合に成立する．このとき，$\boldsymbol{A}_i$ が正方（音源数 = チャネル数）かつ正則であれば, 分離行列 $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} \in \mathbb{C}^{N\times M}$ が存在し，分離信号は次式で表現できる．

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

従って，式 (5) 中の分離行列 $\boldsymbol{W}_i$ を全周波数において推定することで音源分離が達成できる．

2.2 パーミュテーション問題

FDICA 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，分離音源のスケール及び順番に関しては推定することができない．従って，FDICA で推定される分離行列を $\hat{\boldsymbol{W}}_i$ とすると，たとえ完全に推定されたとしても，次式のような不定性が残る．

$$\hat{\boldsymbol{W}}_i = \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{W}_i \tag{6}$$

$$= \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{A}_i^{-1} \tag{7}$$

ここで，$\boldsymbol{P}_i$ は任意のパーミュテーション行列（$\boldsymbol{A}_i^{-1}$ の行の順番を入れ替えうる行列），$\boldsymbol{D}_i$ は任意の対角行列（$\boldsymbol{A}_i^{-1}$ の各行を任意定数倍しうる行列）である．すなわち，FDICA で推定される分離信号

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

$$= (y_{ij,1}, \cdots, y_{ij,N})^{\mathrm{T}} \in \mathbb{C}^N \tag{9}$$

*Deep permutation solver based on local time-frequency structure by Shuhei Yamaji (NIT Kagawa), Daichi Kitamura (NIT Kagawa).

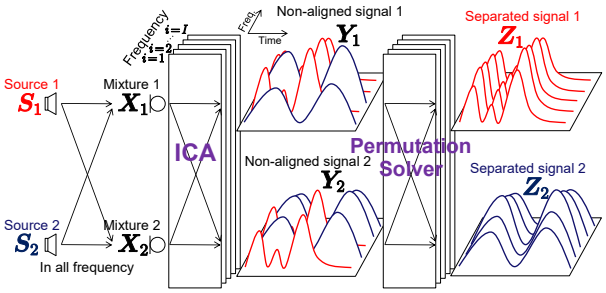

Fig. 1　Permutation problem in FDICA.

は，周波数毎に分離信号の順番やスケールがばらばらになっている状態である．このうち，行列 $\boldsymbol{D}_i$ によって生じるスケールの任意性に関しては，プロジェクションバック法 [3] により解析的に復元することができる．一方で，行列 $\boldsymbol{P}_i$ によって生じる分離信号の順番の任意性（パーミュテーション）を復元することは容易ではない．これを一般にパーミュテーション問題と呼び，その概念図を Fig. 1 に示す．ここで，FDICA で推定される分離信号 $\boldsymbol{y}_{ij}$ の音源毎の複素スペクトログラム行列を $\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$ で表している．このパーミュテーション問題を解決することで得られる分離信号は次式となる．

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

$$= \boldsymbol{P}_i^{-1}\boldsymbol{D}_i^{-1}\hat{\boldsymbol{W}}_i\boldsymbol{x}_{ij} \tag{11}$$

本稿では式 (11) の $\boldsymbol{P}_i^{-1}$ を全周波数において推定することが課題となる．

3　提案手法

3.1　動機

従来のパーミュテーション解決法 [3] では，FDICA で推定される（パーミュテーションがばらばらの状態の）N 個の分離信号 $\boldsymbol{Y}_n$ のパワースペクトログラム $|\boldsymbol{Y}_n|^{.2}$ において，隣接周波数ビン間の音源アクティベーションの相関を調べ，相関が高い組み合わせは同一音源であるという仮定の下並び替えを行い，分離信号 $\boldsymbol{Z}_n$ を出力する．ここで，行列に対する演算 $|\cdot|^{.2}$ は要素毎の 2 乗を施した行列を表す．事実として，同一音源の隣接周波数ビンでは相関が高くなる傾向があり [5]，上記の仮定は妥当である．但し，文献 [3] では，より頑健なパーミュテーション解決のため，$|\boldsymbol{Y}_n|^{.2}$ の各周波数ビンの音源アクティベーションを平滑化する処理を施している．

提案手法では，より頑健かつ高精度なパーミュテーション解決のために，学習データを用いて隣接周波数ビンのアクティベーションから音源を並び替える DNN モデルを構築する．即ち提案手法は，既存手法 [3] における平滑化処理及びクラスタリングを，学習データにとって最適なモデルに置き換えたパーミュテーション解決法である．

3.2　DNN の入出力

Fig. 2 は，提案手法の DNN の入力ベクトルの概略図である．今，パーミュテーションがばらばらの状態の分離信号のパワースペクトログラム $|\boldsymbol{Y}_n|^{.2}$ において，隣接する 2 つの周波数ビンの音源アクティベーション（i 及び $i+1$）の長さ τ の短時間区間の時系

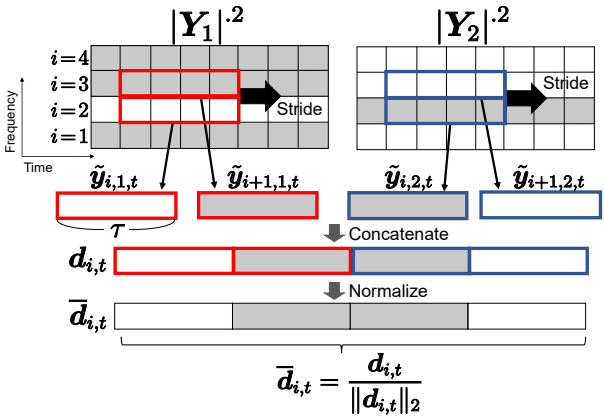

Fig. 2　Input vector for DNN.

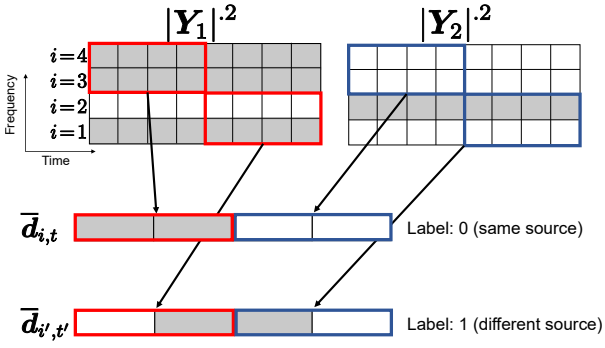

Fig. 3　Input vectors and their labels.

列ベクトルを次式のように表す．

$$\tilde{\boldsymbol{y}}_{i,n,t} = \left(|y_{ij,n}|^2, |y_{i(j+1),n}|^2, \cdots, |y_{i(j+\tau-1),n}|^2\right) \tag{12}$$

$$\tilde{\boldsymbol{y}}_{i+1,n,t} = \left(|y_{(i+1)j,n}|^2, |y_{(i+1)(j+1),n}|^2, \cdots, |y_{(i+1)(j+\tau-1),n}|^2\right) \tag{13}$$

ここで，$t = 1, \cdots, T$ は $|\boldsymbol{Y}_n|^{.2}$ から取り出す短時間区間のインデクスであり，その個数 T は短時間区間の長さ τ とそのストライド幅に依存して決まる．この短時間区間の隣接周波数ベクトルを Fig. 2 のように結合したベクトルを $\boldsymbol{d}_{i,t}$ と定義し，さらに $\boldsymbol{d}_{i,t}$ の正規化ベクトルを $\overline{\boldsymbol{d}}_{i,t}$ とすると，それぞれ次式のようになる．

$$\boldsymbol{d}_{i,t} = (\tilde{\boldsymbol{y}}_{i,1,t},\ \tilde{\boldsymbol{y}}_{i+1,1,t},\ \tilde{\boldsymbol{y}}_{i,2,t},\ \tilde{\boldsymbol{y}}_{i+1,2,t}) \tag{14}$$

$$\overline{\boldsymbol{d}}_{i,t} = \frac{\boldsymbol{d}_{i,t}}{\|\boldsymbol{d}_{i,t}\|_2} \tag{15}$$

ここで，$\|\cdot\|_2$ は L_2 ノルムを表す．式 (15) を DNN の入力ベクトルとする．

DNN の出力については，入力した 2 つの隣接周波数ビンの短時間区間の音源アクティベーションが同一音源か否かを表す 2 値とする．即ち，Fig. 3 に示すように，入力ベクトルを構成する $\tilde{\boldsymbol{y}}_{i,n,t}$ と $\tilde{\boldsymbol{y}}_{i+1,n,t}$ が同一音源であれば 0，異なる音源であれば 1 を出力する 2 値分類モデルである．

3.3　DNN の構造

Fig. 4 に提案手法の DNN の構造を示す．入力層，隠れ層 5 層，及び出力層の計 7 層からなる全結合構成となっており，1〜4 番目の隠れ層には ReLU 関数，

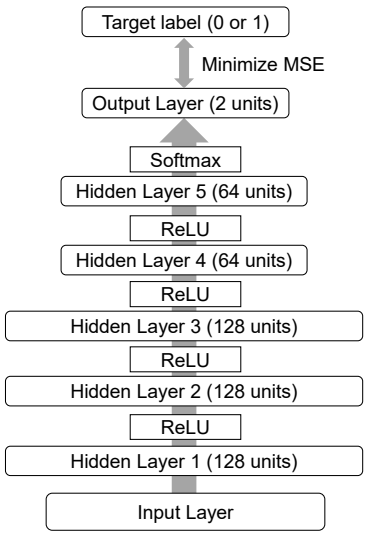

Fig. 4　DNN architechture.

最終隠れ層にはSoftmax関数を適用することで，2値分類問題へ対応している．誤差関数には平均二乗誤差（mean squared error: MSE）を使用する．

3.4　テストデータに対する多数決推論

音声信号のパワースペクトルでは，息継ぎ等で生じる無音区間が多く存在する．式 (15) のように短時間区間を切り出した入力ベクトルが無音区間に該当した場合，DNN の推定は困難となり，誤った出力をしてしまう可能性が高い．

この無音区間の悪影響を抑えるため，提案手法では，DNN をテストデータに適用する際に，周波数毎に多数決を行う．具体的には，各周波数の全短時間区間（$t = 1, \cdots, T$）の DNN の推定結果を多数決することで，一つのパーミュテーション推定結果を決定する．

4　実験

4.1　実験条件

本実験では JVS コーパス [14] の音声信号データセット（nonpara30）を使用し，これらの音声ファイルに RWCP データベース [15] の JR2 インパルス応答を畳み込むことで，1 ファイル当たり 20 s，残響長 470 ms の音声ファイルを 190 ファイル（男性 46 名 95 ファイル，女性 48 名 95 ファイル）作成した．畳み込みに使用したインパルス応答は，文献 [12] に記載のマイク間隔 5.66 cm 及び音源方位 60°&120° のものを使用した．スペクトログラムの作成については，窓長 512 ms 及びシフト長 128 ms でハミング窓を用いて STFT を適用した．

本実験では，上記の畳み込み音声信号を 2 話者選び，それらの周波数成分をランダムにシャッフルして，計 20 万ファイルの混合信号の学習データを作成し，学習用及び検証用に 2 等分した．これは，前段の FDICA が全ての周波数で完全に分離できたという状況を想定している．また，テストデータについては，学習データに含まれていない男女の話者（男性 3 名，女性 3 名）の混合信号を同様の手順で計 100 ファイル作成した．

DNN の学習では，バッチサイズを 128 とし，最適化アルゴリズムには Adam を用いた．各ハイパーパラメータは学習率を 0.001，$\beta = 0.9$ に設定した．式 (12) 及び (13) における短時間区間長 τ は 20 とし，短時間区間のストライド幅は 4 とした．この時の入力層の次元は $20 \times 4 = 80$ であり，同一周波数に対する DNN の適用回数は $T = 37$ 回であった．

4.2　実験結果

Fig. 5 (a) にパーミュテーション問題を解決する前のスペクトログラムおよび (b) に学習済み DNN に基づいて解決したスペクトログラムの一例を示す．このグラフから，パーミュテーション問題を起こしていた各スペクトログラムが完全に分離されている様子がわかる．

Fig. 6 (a) には，テストデータ 100 ファイルに対して，DNN の適用時に 1 回の推定結果で並び直した場合の周波数毎の不正解数および (b) に $T = 37$ 回の同一周波数に対するパーミュテーション推定結果から多数決を取った場合の周波数毎の不正解数をヒストグラムで示す．このグラフから，1 回の推定結果から基づくパーミュテーション解決では，周波数全体で間違った推定をする可能性があることがわかる．しかし，推定結果の多数決を取ることで，間違った推定は 0 Hz 付近及びナイキスト周波数の 8000 Hz 付近のみとなっている．これらの周波数では，人間の音声の成分が存在しないため，多くのデータで不正解となったが，音源分離において問題にはならないと考えられる．以上より，周波数毎に完全に分離されている混合信号のパーミュテーション問題に対して，本手法は高い解決性能を示す．

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 によって解決する手法を新たに提案した．提案手法では，局所的な時間周波数構造から正解のパーミュテーションを推定する DNN を学習している．また，同一周波数の推定結果に対して多数決を取ることで，頑健なパーミュテーション解決を実現している．

実験の結果，周波数ビン毎に完全に分離されている理想信号に対して，高い精度でパーミュテーション問題を解決できることを確認した．実際の FDICA では，完全に分離されていない周波数ビンも存在する上，中間周波数ビンでパーミュテーション解決を誤ると，分離精度が大きく劣化する可能性があるため，より頑健な DNN モデルが必要となることが考えられる．

謝辞　本研究の一部は JSPS 科研費 19K20306 及び NVIDIA GPU Grant 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 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] 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.

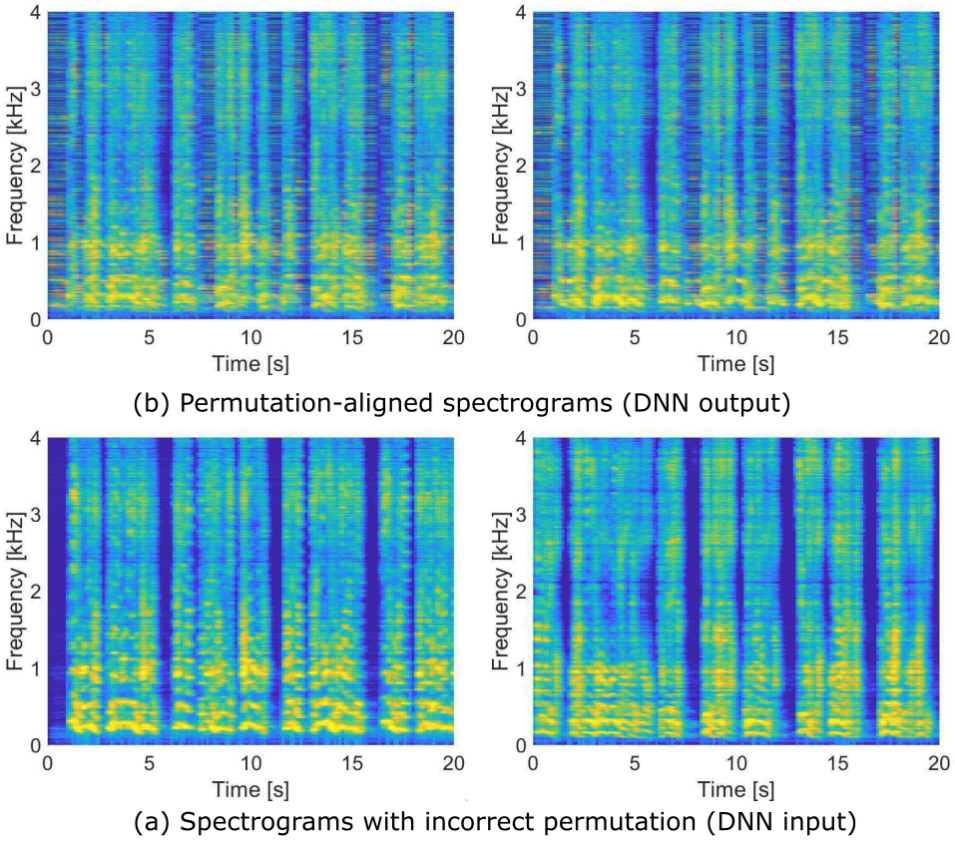

Fig. 5 Spectrograms of (a) input and (b) output data of DNN.

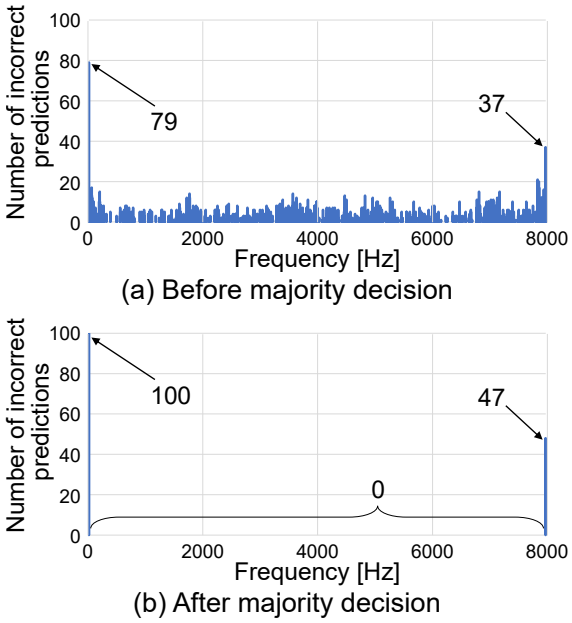

Fig. 6 Histgrams of incorrect predictions: (a) before majority decision and (b) after majority decision.

[4] 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.

[5] 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, Sep. 2004.

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

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

[10] 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.

[11] 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. ASLP*, vol. 27, no. 10, pp. 1601–1615, 2019.

[12] 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.

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

[14] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint*, 1908.06248, Aug. 2019.

[15] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. LREC*, pp. 965–968, 2000.