



香川高専

卒業研究論文

論文題目

深層パーミュテーション解決法の基礎的検討

提出年月日	令和 4 年 2 月 25 日	
学 科	電気情報工学科	
氏 名	蓮池 郁也	印
指導教員（主査）	北村 大地 講師	印
副 査	齋元 洋一 助教	印
学 科 長	辻 正敏 教授	印

香川高等専門学校

Basic Study for Deep Permutation Solver

Fumiya Hasuike

Department of Electrical and Computer Engineering

National Institute of Technology, Kagawa College

Abstract

Audio source separation is a technique for estimating specific audio sources from an observed mixture signal with multiple sources. This technique can be used for many applications, e.g., speech estimation of multiple speech mixture, reduction of background noise, and extraction of individual musical instrument signals in music. Frequency-domain independent component analysis (FDICA) is a typical audio source separation method that applies independent component analysis to each of frequencies. However, FDICA encounters the so-called permutation problem, which is a frequency-wise reordering problem of separated source components. Thus, FDICA requires the permutation solver as post processing. Recently, a new permutation solver based on deep neural networks (DNN) was proposed, but the existing method has a problem that the algorithm becomes complicated for the separation of three or more sources. In this thesis, I propose a simpler algorithm of deep permutation solver and investigates the validity of using DNN to solve the permutation problem by experiments. The proposed method directly predicts the permutation of the separated source components for all the frequencies. Since the permutation of the entire estimated sources is arbitrary, permutation invariant training is introduced. Experimental results show that the proposed deep permutation solver provides nearly 100% correct permutations for artificially produced permutation problems. In addition, for the actual speech and music signals, the proposed method achieved over 90% accuracy for a block-wise permutation problem.

Keywords: frequency-domain independent component analysis, permutation solver, deep neural networks

(和訳)

音源分離とは、複数の未知の音源が混ざった観測信号から、混ざる前の個々の音源を推定する技術である。この技術は、複数人の同時発話に対する各人の音声の推定、背景雑音の抑圧、音楽信号の各楽器音の抽出等に利用される。代表的な音源分離手法の1つである時間周波数領域独立成分分析 (frequency-domain independent component analysis: FDICA) は、周波数毎に独立成分分析を適用することで分離を行う。しかし FDICA には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付随するため、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。近年では、深層ニューラルネットワーク (deep neural networks: DNN) を用いたパーミュテーションの解決法（深層パーミュテーション解決法）が提案されたが、既存手法は3音源以上の音源分離においてアルゴリズムが極端に複雑化する課題がある。本論文では、より簡便なアルゴリズムによる深層パーミュテーション解決法を提案し、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実験的に調査する。提案手法は、全周波数成分について分離信号の正しいパーミュテーションを直接予測する。このとき、音源信号全体の予測順序は任意であるため、順序不变学習を用いる。実験結果から、提案する深層パーミュテーション解決法は人工的に作成したパーミュテーション問題に対して、100%に近い正答率を示した。また、実際の音声及び音楽信号に対してもブロック単位のパーミュテーション問題であれば、90%を超える正答率を示した。

目次

第 1 章	序論	1
1.1	本論文の背景	1
1.2	本論文の目的	3
1.3	本論文の構成	4
第 2 章	基礎理論と従来手法	5
2.1	まえがき	5
2.2	ICA の基本原理	5
2.2.1	信号源の混合モデルと分離方法	5
2.2.2	統計的独立性	6
2.2.3	ICA における任意性	7
2.3	STFT	8
2.4	周波数領域における BSS の定式化	10
2.5	FDICA	10
2.6	パーミュテーション問題とその解決	11
2.7	IVA と ILRMA	13
2.8	深層パーミュテーション解決法	14
2.9	本章のまとめ	15
第 3 章	提案手法	16
3.1	まえがき	16
3.2	動機	16
3.3	DNN の入出力	19
3.4	DNN の構造	23
3.5	DNN 学習時の損失関数	23
3.6	学習済の DNN のテストデータへの適用	25
3.7	本章のまとめ	26
第 4 章	実験	27
4.1	まえがき	27
4.2	実験条件	27

4.2.1	人工データを用いた基礎実験の条件	27
4.2.2	実際の音響信号を用いた実験の条件	31
4.3	実験結果	33
4.3.1	人工データに対する実験結果	33
4.3.2	音声及び音楽信号に対する実験結果	40
4.4	本章のまとめ	43
第 5 章 結言		44
謝辞		45
参考文献		45
付録 A	Birkhoff–von Neumann の定理	49
付録 B	人工データに対する予測結果	50

第1章

序論

1.1 本論文の背景

音源分離とは、観測したある混合音源から、混合前の信号を推定する技術である。この技術の具体的な応用例を Fig. 1.1 に示す。音源分離の例として音声信号に対する分離が挙げられる。一例ではあるが、音声信号に対する分離では、混合信号から雑音を除去して音声だけを抽出及び強調するタスクや、複数人が会話をっている状況下で個人毎に分離するような音声同士の分離タスク、楽器音の自動採譜タスクなどがある。近年では、スマートスピーカーのような音声認識技術を用いた製品が増えている中で、雑音や非目的話者の音声信号等の混合に起因した音声認識精度の低下を回避するためにも、目的話者のみのクリアな单一音声信号が入力として求められている。音声認識だけでなく、イヤホンのノイズキャンセリング機能や補聴器の音声強調機能のように、人間の聴覚機能をサポートする面でも音源分離の応用先は数多く存在する。

上記のように、音源分離技術は歴史的にみても非常に重要な技術として長年研究されており、これらのタスクを満足するには高精度な音源分離手法が求められる。この経緯から 1990 年代から今日まであらゆる音源分離手法が提案してきた。その音源分離手法の中でも、マイクロホンや音源の位置等の事前情報が無いという条件下で、複数の信号源が混合した混合音から、混合前の分離音を推定するような分離手法をブラインド音源分離 (blind source separation: BSS) [1] という。Fig. 1.2 は BSS の概要を示しており、未知の混合系 \mathbf{A} (マイクロホンや音源位置や部屋の形状及び材質などに依存して変化) から混合信号が生成される。これに対して混合系 \mathbf{A} の逆系である分離系 \mathbf{W} を推定し、観測信号 \mathbf{X} に適用することで混合前の音源を推定する。

特に、観測マイクロホン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ぶ。この条件下での音源分離には、音源信号間の統計的独立性の仮定に基づく手法が広く用いられている。独立成分分析 (independent component analysis: ICA) [2] は、優決定条件下的 BSS に広く適用されている代表的な手法である。音響信号の混合問題では一般的に残響の影響を受けて、瞬時混合ではなく時間畳み込み混合となることから、直接 ICA を時間領域の観測信号に適用しても BSS を達成することは不可能である。そこで、観測信号を時間周波数領域に変

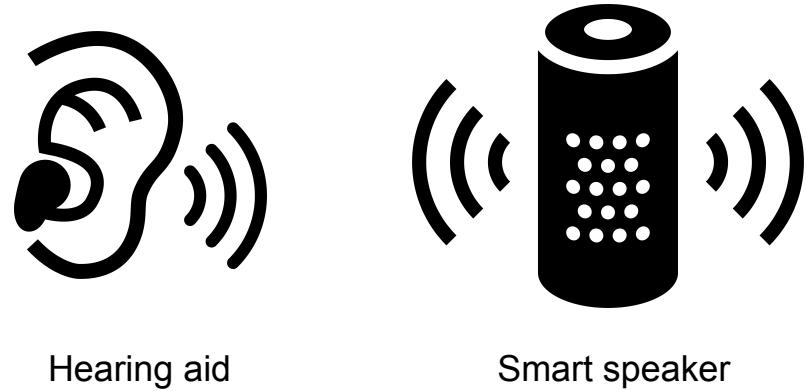


Fig. 1.1: Examples of application using speech source separation.



Fig. 1.2: Overview of BSS.

換することで周波数毎の瞬時混合として混合系をモデル化し、周波数毎に ICA を適用する時間周波数領域 ICA (frequency-domain ICA: FDICA) [3] が提案された。ここで、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。FDICAにおいて、周波数毎の分離信号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 [4, 5]、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 [6]、及びその両者を組み合わせた手法 [7] 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可能な限り回避しながら周波数毎の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) [8, 9] 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 (nonnegative matrix factorization: NMF) [10] 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) [11, 12] 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。

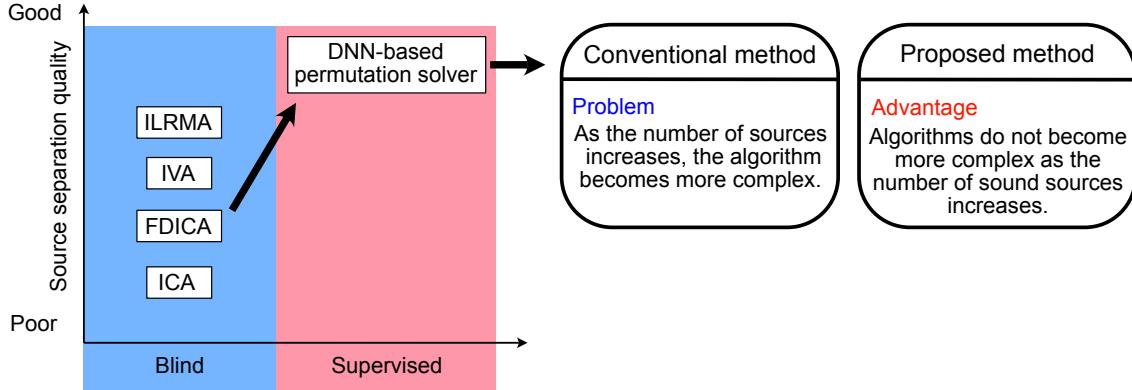


Fig. 1.3: Scope of this thesis.

1.2 本論文の目的

前述したブラインドな音源分離手法は、パーミュテーション問題を回避しつつ、高い精度で分離するモデルへと発展を遂げてきた。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い。特に複数音声の混合信号や、複数の調波楽器音の混合信号における頑健・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。一方で、文献 [13] 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。従って、FDICAにおいて各周波数での音源分離は高精度であり、パーミュテーション問題のみが課題として残っている。近年では、パーミュテーション問題を解決するために、深層ニューラルネットワーク (deep neural networks: DNN) を用いてサブバンドと呼ばれる局所帯域毎に、隣接した周波数のアクティベーションの相関を調べる手法 [14] が提案してきた。しかし、この手法は局所帯域毎に処理をしているため、複雑なアルゴリズム構成となっており、3 音源以上の音源分離を行うことは現実的には難しい。そこで、本論文では、3 音源以上でもアルゴリズムが複雑化しない、DNN を用いたデータ駆動型（教師あり）パーミュテーション解決法（以後、深層パーミュテーション解決法と呼ぶ）について提案し、その妥当性について実験的に調査する。同時に、ブロックパーミュテーション問題に対する有効性についても調査する。この提案手法と既存手法の位置関係の概念図を Fig. 1.3 に示す。本論文では、FDICA におけるパーミュテーション問題のみに焦点を当てており、分離信号成分の正しいパーミュテーションを予測する様に学習した DNN を用いてパーミュテーション問題を解決することを目的とする。ここでは、DNN 優決定条件下での複数音声の混合を模倣した人工的なデータと実際の音声及び音楽信号に対して、深層パーミュテーション解決法を適用することを考える。

1.3 本論文の構成

まず、2章では、本論文の解決すべき課題であるパーミュテーション問題の説明に必要となるICAの基本原理や音響信号の時間周波数領域への変換である短時間Fourier変換(short-time Fourier transform: STFT)に加え、パーミュテーション問題を可能な限り回避するBSSのIVA及びILRMA、そして既存の深層パーミュテーション解決法について詳しく説明する。これらは、いずれも提案手法の説明に必要となる知識である。3章では、本論文の提案手法である深層パーミュテーション解決法の新たなアルゴリズムの詳細について、DNNの構造からパーミュテーション解決の処理までを詳細に述べる。4章では、人工データと実際の音声及び音楽信号に対する音源分離実験を行い、提案深層パーミュテーション解決法の性能の検証を行う。最後に5章では、すべての章を総括した結言を述べる。

第 2 章

基礎理論と従来手法

2.1 まえがき

本章では、音源分離技術において必要となる手法の基礎理論とこれまでに提案してきた音源分離手法について述べる。まず 2.2 節では、BSS の基礎理論となる ICA について説明する。2.3 節では、音響信号処理でよく用いられる STFT について説明する。2.4 節では、時間周波数領域における音源信号及び BSS の定式化について説明する。2.5 節では、時間周波数領域で周波数毎に ICA を適用する FDICA について説明する。2.6 節では、本論文の主題として、FDICA に付随するパーミュテーション問題の説明と、既存のパーミュテーション解決法について説明する。2.7 節では、パーミュテーション問題を可能な限り回避する BSS の IVA 及び ILRMA について詳細を述べる。2.8 節では、既存の深層パーミュテーション解決法とその問題について説明する。2.9 節では、本章のまとめを述べる。

2.2 ICA の基本原理

本章では、BSS の基礎である ICA [2] について説明する。なお、本章の説明では簡単のために、音源数及びマイクロホン数がいずれも 2 の場合を例として説明するが、本章記載の基本原理は音源数及びマイクロホン数がいずれも 3 以上の場合についても、一般性を失うことなく同様に説明できる。但し、後述の通り、音源数とマイクロホン数は常に等しいという仮定が必要である。BSS の文脈では、このような「音源数がマイクロホン数以下」という条件を優決定条件と呼ぶ。

2.2.1 信号源の混合モデルと分離方法

今、2つの信号源 $s_1(l)$ 及び $s_2(l)$ があり、その混合信号を 2 つのマイクロホンで観測するという状況を考える。ここで、 $l = 1, 2, \dots, L$ は離散時間インデクスを示す。マイクロホンで観測された信号を $x_1(l)$ 及び $x_2(l)$ とすると、2 つの信号源の混合現象は次の連立方程式でモ

ル化できる。

$$\begin{cases} x_1(l) = a_{11}s_1(l) + a_{12}s_2(l) \\ x_2(l) = a_{21}s_1(l) + a_{22}s_2(l) \end{cases} \quad (2.1)$$

ここで、信号の伝搬を表す係数 a_{mn} は、時刻 l には依存せず常に一定であると仮定する。即ち、信号源の位置及びマイクロホンの位置が動かないことを仮定している。また、 $n = 1, 2, \dots, N$ 、及び $m = 1, 2, \dots, M$ はそれぞれ音源及びチャネルのインデックスを示す。伝搬係数 a_{mn} をまとめた行列を以下のように定義する。

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad (2.2)$$

この行列 \mathbf{A} は混合行列と呼ばれる。観測信号ベクトル $\mathbf{x}(l) = [x_1(l), x_2(l)]^T$ 、信号源ベクトル $\mathbf{s}(l) = [s_1(l), s_2(l)]^T$ 及び混合行列 \mathbf{A} を用いて、式 (2.1) 及び (2.1) の連立方程式は次式のように書き直せる。

$$\mathbf{x}(l) = \mathbf{A}\mathbf{s}(l) \quad (2.3)$$

ここで、 \cdot^T はベクトルや行列の転置を表す。分離信号を $\mathbf{y}(l) = [y_1(l), y_2(l)]^T$ 、分離行列を \mathbf{W} とそれぞれ定義すると、音源分離は以下のように表される。

$$\mathbf{y}(l) = \mathbf{W}\mathbf{x}(l) \quad (2.4)$$

このとき、混合行列 \mathbf{A} の逆行列が存在する (\mathbf{A} が正則) ならば、 $\mathbf{W} = \mathbf{A}^{-1}$ となるように \mathbf{W} を推定することで、信号源 $\mathbf{s}(l)$ を推定することができる。

$$\mathbf{y}(l) = \mathbf{W}\mathbf{x}(l) \quad (2.5)$$

$$= \mathbf{A}^{-1}\mathbf{x}(l) \quad (2.6)$$

$$= \mathbf{A}^{-1}\mathbf{A}\mathbf{s}(l) \quad (2.7)$$

$$= \mathbf{s}(l) \quad (2.8)$$

このように、混合行列 \mathbf{A} の逆行列である分離行列 \mathbf{W} を推定することで、音源分離を達成することができる。しかしながら、音源やマイクロホンの位置関係が未知である BSS においては、混合行列 \mathbf{A} もまた未知である。そこで、ICA では、信号源の混合モデル式 (2.3) の仮定の他に、信号そのものの統計的なモデル ($p(s_1)$ 及び $p(s_2)$ に対する仮定) を導入することで、分離行列 \mathbf{W} を推定する。

2.2.2 統計的独立性

ICA による信号源分離を理解する上で重要な概念として、統計的独立性がある。今、信号源 $s_1(l)$ 及び $s_2(l)$ を確率変数として扱い、それらの生成モデルを $p(s_1)$ 及び $p(s_2)$ と定義する。通常、各信号源 ($s_1(l)$ 及び $s_2(l)$) は互いに無関係であり、例えば $s_1(l)$ から $s_2(l)$ を予測

や説明することはできないはずである。そのため、 $s_1(l)$ と $s_2(l)$ は互いに統計的に独立とみなすことができ、次式が成立する。

$$p(s_1, s_2) = p(s_1)p(s_2) \quad (2.9)$$

同様に、理想的な分離フィルタが推定できれば、分離信号 $y_n(l)$ も統計的に独立であるため、次式が成立する。

$$p(y_1, y_2) = p(y_1)p(y_2) \quad (2.10)$$

ここで、 $p(y_1)$ 及び $p(y_2)$ はそれぞれ分離信号 $y_1(l)$ 及び $y_2(l)$ の生成モデルであり、 $p(y_1, y_2)$ は同時分布である。従って ICA による BSS は、式 (2.10) が成立するような分離フィルタ \mathbf{W} を推定する問題であると解釈できる。上記の問題を定式化すると、次式のように書き表せる。

$$\arg \min_{\mathbf{W}} \mathfrak{I}(\mathbf{W}) \quad (2.11)$$

$$\mathfrak{I}(\mathbf{W}) = \mathfrak{D}_{\text{KL}}[p(y_1, y_2) || p(y_1)p(y_2)] \quad (2.12)$$

ここで、 $\mathfrak{D}_{\text{KL}}[p(s) || q(s)]$ は Kullback–Leibler (KL) ダイバージェンスと呼ばれ、2つの分布間 ($p(s)$ 及び $q(s)$) の距離を測る関数として次式のように定義される。

$$\mathfrak{D}_{\text{KL}}[p(s) || q(s)] = \int p(s) \log \frac{p(s)}{q(s)} ds \quad (2.13)$$

また、分離行列 \mathbf{W} で観測信号を線形変換する前 (\mathbf{x}) と後 (\mathbf{y}) の確率変数を考えたとき、それぞれの同時分布 $p(\mathbf{y}) = p(y_1, y_2)$ 及び $p(\mathbf{x}) = p(x_1, x_2)$ の間には、次式が成立する。

$$p(\mathbf{y}) = \frac{1}{|\det \mathbf{W}|} p(\mathbf{x}) \quad (2.14)$$

式 (2.13) 及び (2.14) を用いて式 (2.12) を変形すると、最終的な最小化関数 $\mathfrak{I}(\mathbf{W})$ は以下のように書ける。

$$\begin{aligned} \mathfrak{I}(\mathbf{W}) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2) \log p(x_1, x_2) dx_1 dx_2 - \log |\det \mathbf{W}| \\ &\quad - \int_{-\infty}^{\infty} p(y_1) \log p(y_1) dy_1 - \int_{-\infty}^{\infty} p(y_2) \log p(y_2) dy_2 \end{aligned} \quad (2.15)$$

ICA では、式 (2.15) が最小化される分離行列 \mathbf{W} を求めることで、信号源を分離する。

2.2.3 ICA における任意性

前項より、 $y_1(l)$ と $y_2(l)$ の独立性を最大化する分離行列 \mathbf{W} を求める ICA の最適化問題が定式化される。しかしながら、分離信号の順序及びスケール（大きさ）の違いは、独立性の尺度である式 (2.12) に影響を与えないことは明らかである。従って、ICA によって推定される分離信号 $y_1(l)$ 及び $y_2(l)$ には、以下の任意性が存在する。

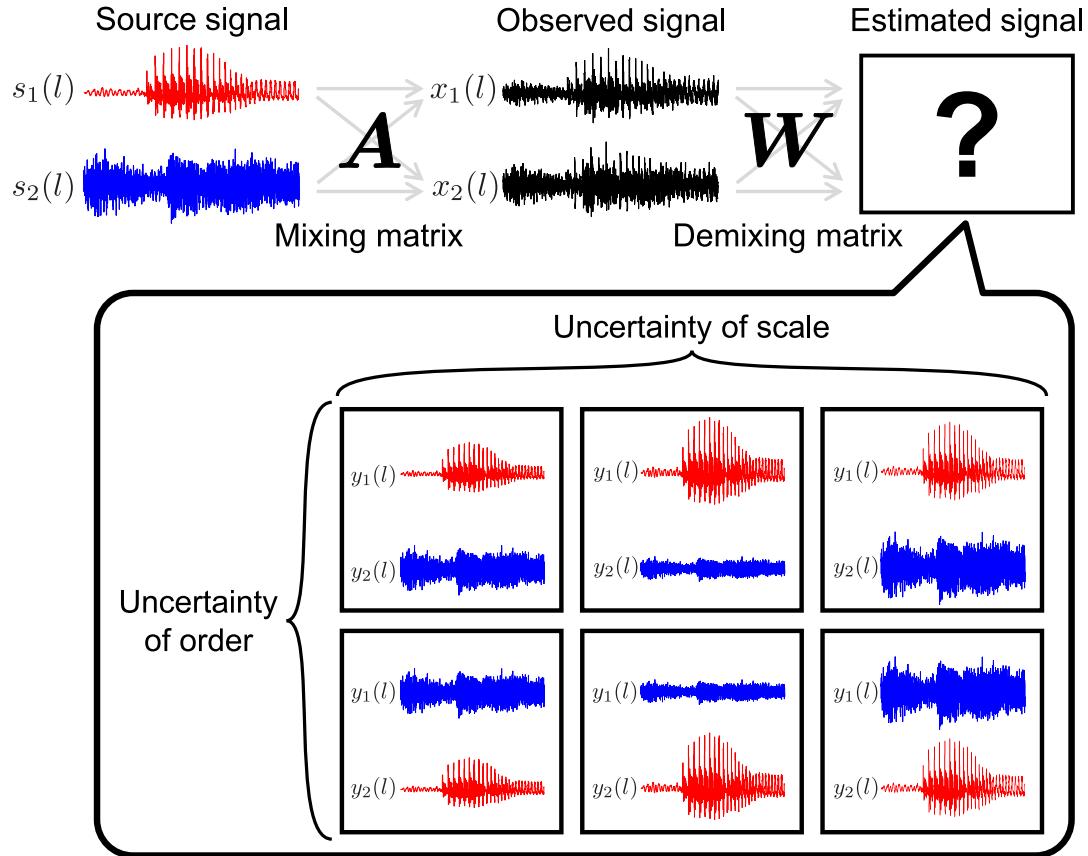


Fig. 2.1: Uncertainty in ICA. ICA cannot determine order and scales of estimated signals.

- (a) 分離信号の順序には任意性がある
- (b) 分離信号のスケールには任意性がある

これらの任意性は分離信号に対して Fig. 2.1 のように現れる。上記の任意性 (a) より、元々の信号源の順序が入れ替わる可能性がある。また、任意性 (b) より、分離信号のスケールが混合前の音源信号のスケールから変化してしまう可能性がある。なお、信号のスケールの任意性に関しては、プロジェクションバック (projection back: PB) 法 [15] と呼ばれる解析的な補正方法が提案されており、2.4 節で説明する FDICA においても大きな問題にはならない。一方、順序の任意性は FDICA におけるパーティション問題を引き起こす要因となる。

2.3 STFT

STFT は Fig. 2.2 に示すような時間的に変化するスペクトルを表現するための手法である。いま、音響信号の時間波形を次式で定義する。

$$\mathbf{x} = [x(1), x(2), \dots, x(l), \dots, x(L)]^T \in \mathbb{R}^L \quad (2.16)$$

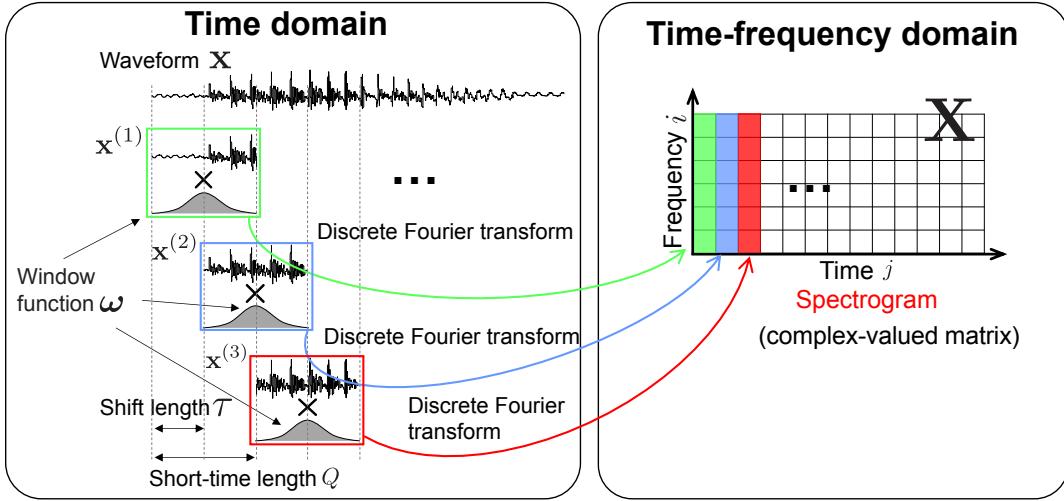


Fig. 2.2: Mechanism of STFT. Each of windowed short-time signals are transformed to frequency domain by discrete Fourier transform.

STFT の分析窓関数の長さ及びシフト長をそれぞれ Q 及び τ としたとき、時間領域の信号 \mathbf{x} の j 番目の短時間区間（時間フレーム）の信号は次式で表される。

$$\mathbf{x}^{(j)} = [\mathbf{x}((j-1)\tau+1), \mathbf{x}((j-1)\tau+2), \dots, \mathbf{x}((j-1)\tau+Q)]^T \quad (2.17)$$

$$= \left[\mathbf{x}^{(j)}(1), \mathbf{x}^{(j)}(2), \dots, \mathbf{x}^{(j)}(q), \dots, \mathbf{x}^{(j)}(Q) \right]^T \in \mathbb{R}^Q \quad (2.18)$$

ここで、 $j = 1, 2, \dots, J$ 及び $q = 1, 2, \dots, Q$ は、それぞれ時間フレーム及び時間フレーム内のサンプルを示す。また、セグメント数 J は次式によって与えられる。

$$J = \frac{L}{\tau} \quad (2.19)$$

ただし、時間領域の信号 \mathbf{x} は式 (2.18) が自然数となるように、信号の末尾に必要な分だけ零値が追加されているものとする。このとき、信号 \mathbf{x} の STFT を次式で表す。

$$\mathbf{X} = \text{STFT}_\omega(\mathbf{x}) \in \mathbb{C}^{I \times J} \quad (2.20)$$

ここで、 \mathbf{X} は（複素）スペクトログラムと呼ばれ、Fig. 2.2 に示すように時間と周波数の 2 次元の行列である。スペクトログラム \mathbf{X} の (i, j) 番目の要素は次式で表される。

$$x_{ij} = \sum_{q=1}^Q \omega(q) \mathbf{x}^{(j)}(q) \exp \left\{ \frac{-i2\pi(q-1)(i-1)}{F} \right\} \quad (2.21)$$

ここで F は $\lfloor \frac{F}{2} \rfloor + 1 = I$ を満たす整数 ($\lfloor \cdot \rfloor$ は床関数)、 $i = 1, 2, \dots, I$ は周波数ビンのインデクス、 i は虚数単位は短時間信号 $\mathbf{x}^{(j)}$ の両端の不連続性を解消するための解析窓関数をそれぞれ示している。このように STFT は、時間領域の信号を一定幅の短時間信号に分割して解析窓関数を乗じて離散フーリエ変換を適用し、スペクトログラムと呼ばれる複素時間周波数行

列に変換する処理である。音源分離等の多くの音響信号処理では、このスペクトログラムを信号処理の対象とする。

2.4 周波数領域における BSS の定式化

本節以降、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 とする。また、音源信号、観測信号、及び分離信号の時間周波数毎の成分をそれぞれ次式で表す。

$$\mathbf{s}_{ij} = [s_{ij1}, s_{ij2}, \dots, s_{ijn}, \dots, s_{ijN}]^T \in \mathbb{C}^N \quad (2.22)$$

$$\mathbf{x}_{ij} = [x_{ij1}, x_{ij2}, \dots, x_{ijm}, \dots, x_{ijM}]^T \in \mathbb{C}^M \quad (2.23)$$

$$\mathbf{z}_{ij} = [z_{ij1}, z_{ij2}, \dots, z_{ijn}, \dots, z_{ijN}]^T \in \mathbb{C}^N \quad (2.24)$$

式 (2.22)–(2.24) 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 番目の音源信号のスペクトログラム、 m 番目の観測信号のスペクトログラム、及び n 番目の分離信号のスペクトログラムをそれぞれ $\mathbf{S}_n \in \mathbb{C}^{I \times J}$ 、 $\mathbf{X}_m \in \mathbb{C}^{I \times J}$ 、及び $\mathbf{Z}_n \in \mathbb{C}^{I \times J}$ と定義する。これらの行列の (i, j) 番目の要素はそれぞれ s_{ijn} 、 x_{ijm} 、及び z_{ijn} に一致する。

2.5 FDICA

2.2 節で説明したように、ICA とは、観測信号が独立信号の線形結合として観測される場合に、各信号間の独立性を最も高めるような分離行列を推定することで BSS を実現する手法である。実際の音響信号の混合は収録環境の残響の影響を受けるため、各音源から各マイクロホンまでの空間伝達系のインパルス応答が畳み込まれて混合される。インパルス応答の畳み込みは残響長 R を用いて次式のように表される。

$$\tilde{\mathbf{x}}(l) = \sum_n \sum_{l'=0}^{R-1} \tilde{\mathbf{a}}_n(l') \tilde{s}_n(l - l') \quad (2.25)$$

ここで、 $\tilde{\mathbf{x}}(l) = [\tilde{x}_1(l), \tilde{x}_2(l), \dots, \tilde{x}_M(l)]^T$ 及び $\tilde{s}_n(l)$ はそれぞれ時間領域の観測信号及び (n 番目の) 音源信号であり、 $\tilde{\mathbf{a}}_n(l)$ は音源 n に対する畳み込み混合係数ベクトル (n 番目の音源から全マイクロホンまでのインパルス応答を時間 l 每にまとめたもの) である。式 (2.25) のように混合される複数の音源を分離するためには、分離行列ではなく逆畳み込みフィルタを推定することが必要となる。一般的に逆畳み込みフィルタの推定非常に困難な問題となってしまうことから、時間領域での ICA による BSS は容易ではない。この問題を解決するために、各信号の STFT による時間周波数表現を用いて、式 (2.25) の時間領域における畳み込み混合を、時間周波数領域での周波数毎の瞬時混合に変換し、時間周波数領域で周波数毎に ICA を行う FDICA が提案された [3]。

FDICA では、周波数毎の時不变な混合行列 $\mathbf{A}_i = [\mathbf{a}_{i1} \ \mathbf{a}_{i2} \ \cdots \ \mathbf{a}_{in} \ \cdots \ \mathbf{a}_{iN}] \in \mathbb{C}^{M \times N}$ を

定義し、混合信号が次式で表現できると仮定する。

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{s}_{ij} \quad (2.26)$$

この混合モデルは、観測信号収録時の残響長 R よりも STFT の短時間区間長 Q が十分長い場合に成立する。以後、決定的な系 ($M = N$) を仮定すると、混合行列 \mathbf{A}_i が正則であれば、周波数毎の分離行列 $\mathbf{W}_i = \mathbf{A}_i^{-1} = [\mathbf{w}_{i1} \ \mathbf{w}_{i2} \ \dots \ \mathbf{w}_{in} \ \dots \ \mathbf{w}_{iN}]^H$ を用いて、分離信号を次式で表せる。

$$\mathbf{z}_{ij} = \mathbf{W}_i \mathbf{x}_{ij} \quad (2.27)$$

ここで、 $.^H$ はベクトルや行列のエルミート転置を示す。分離行列の行ベクトルである $\mathbf{w}_{i,n} \in \mathbb{C}^M$ は、 i 番目の周波数ビンにおいて、観測信号から n 番目のみの音源が含まれる分離信号へ変換する分離フィルタである。このように FDICA では、観測信号 \mathbf{x}_{ij} の各周波数ビンに対しそれぞれ独立に（複素数）ICA を適用することで、周波数毎の分離行列 \mathbf{W}_i を全周波数にわたって推定し、BSS の達成を目指す。

2.6 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、2.2.3 項で述べた通り、分離された推定信号の周波数毎のスケール及び順番に関しては不定である。従って、FDICA の推定分離行列を $\hat{\mathbf{W}}_i$ とすると、次式のような不定性が残る。

$$\hat{\mathbf{W}}_i = \mathbf{D}_i \mathbf{P}_i \mathbf{W}_i \quad (2.28)$$

ここで、 $\mathbf{P}_i \in \{0, 1\}^{N \times N}$ は分離行列 \mathbf{W}_i の行ベクトル \mathbf{w}_{in} の順番を入れ替えうるパーミュテーション行列（置換行列）である。例えば、 $N = M = 2$ の場合は

$$\mathbf{P}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (2.29)$$

の 2 種類がパーミュテーション行列であり、 $N = M = 3$ の場合は

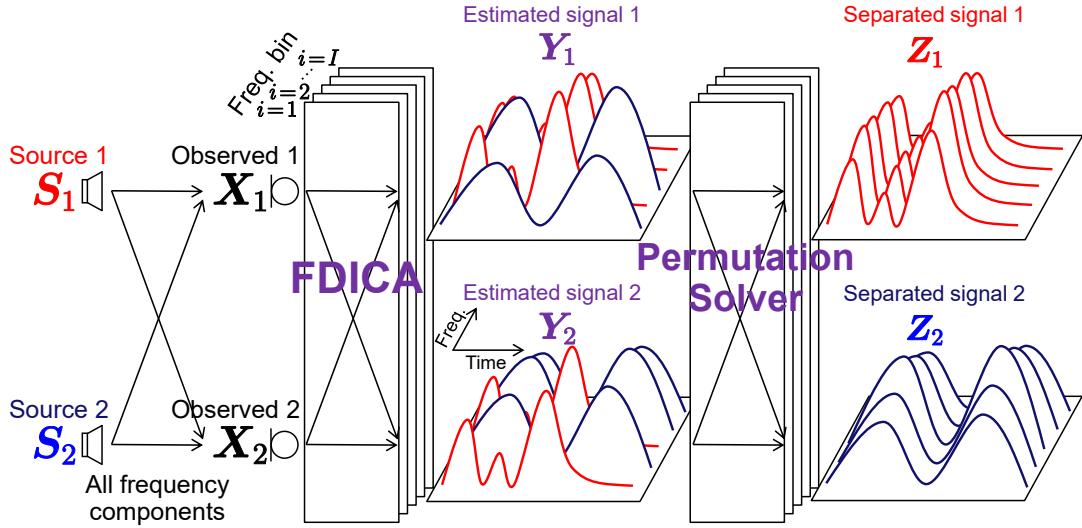
$$\mathbf{P}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \text{ or } \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad (2.30)$$

の 6 種類がパーミュテーション行列である。一方、 $\mathbf{D}_i \in \mathbb{R}^{N \times N}$ は、 \mathbf{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。従って、FDICA で推定される分離信号

$$\mathbf{y}_{ij} = \hat{\mathbf{W}}_i \mathbf{x}_{ij} \quad (2.31)$$

$$= [y_{ij1}, y_{ij2}, \dots, y_{ijn}, \dots, y_{ijN}]^T \in \mathbb{C}^N \quad (2.32)$$

は、推定音源の順番やスケールが周波数毎にばらばらになっている状態である。このうち、 \mathbf{D}_i によって生じるスケールの任意性は、時間領域での ICA の場合と同様に PB 法 [15] で解析的

Fig. 2.3: Permutation problem in FDICA, where $N = M = 2$.

に復元可能である。一方で、 \mathbf{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 I 個の全周波数 bin に関して復元することは、組み合わせ爆発が生じるため容易ではない。具体的には、 I 個の周波数 bin のそれぞれで N 個の音源の順番は $N!$ 種類あるため、全周波数のパーミュテーションは $(N!)^I$ 通り存在することになり、その内の正解（全周波数で同一の音源パーミュテーションとなるもの）は $N!$ 個である。この問題は、一般的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. 2.3 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 \mathbf{y}_{ij} の n 番目のスペクトログラムを $\mathbf{Y}_n \in \mathbb{C}^{I \times J}$ と定義している。FDICA 直後の \mathbf{Y}_n に注目すると、周波数毎での音源分離は達成できている。しかし、時間周波数構造全体としては、異なる音源の分離成分が 1 つの時間周波数内に混在していることが分かる。これがパーミュテーション問題であり、ICA の分離信号の順番に関する不定性に起因して発生している。そのため、FDICA にはポスト処理として、分離された音源の順番を全周波数 bin にわたって正しく並べ直す必要がある。このパーミュテーション問題を解決する処理は次式で表される。

$$\mathbf{z}_{ij} = \mathbf{P}_i^{-1} \mathbf{D}_i^{-1} \mathbf{y}_{ij} \quad (2.33)$$

スケールの不定性を補正する \mathbf{D}_i^{-1} はプロジェクションバック法によって解析的に求められる。従って、パーミュテーション問題の解決とは、全周波数 bin にわたって \mathbf{P}_i^{-1} を求める問題として解釈できる。

このパーミュテーション問題を解決するために、これまでにも数々のパーミュテーション解決法が提案してきた。代表的な既存手法の 1 つに、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 [4, 5] がある。これは、Fig. 2.3 に示す赤色と青色の分離信号のように、分離信号のパーミュテーションが正しければ、隣接した周波数アクティベーション間の相関が高くなりやすいという仮定の下で並べ替える手法である。

このとき、離れた周波数においても、同じ音源のアクティベーション間の相関が高くなるように並び替えられている。他にも、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 [6] 及び音源到来方位と周波数毎の時系列強度の両者を組み合わせたパーミュテーション解決法 [7] 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、とくに音源数が増加した際ににおける頑健・高精度なパーミュテーション問題の解決はいまだ実現できていない。

2.7 IVA と ILRMA

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可能な限り回避しつつ分離信号を推定する手法が登場している。例えば、IVA [8, 9] は、同一音源の周波数成分の共起を仮定しており、FDICA では周波数毎に独立性を最大化していたのに対し、IVA では全周波数成分をまとめてベクトル変数とし、ベクトル間の独立性を最大化するようなモデルとなっている。そのため、実際に複数の周波数ビンで同時に共起する成分为同一音源としてまとめられるような分離行列が推定され、パーミュテーション問題を可能な限り回避することが期待できる。IVA の「同一音源であれば全周波数が共起する」という仮定は、音源信号の時間周波数構造に関するモデルである。実際に、音声信号はこのような時間周波数構造が比較的適合するため、IVA を用いることである程度パーミュテーション問題を回避できる。さらに、IVA の音源信号の時間周波数構造に関するモデル（以後、音源モデルと呼ぶ）をより詳細なモデルに発展させた BSS として、ILRMA [11, 12] が提案されている。ILRMA は、IVA で提案された音源モデルに NMF [10] を用いている。NMF は時間周波数構造を低ランク近似できることから、「同一音源であれば時間周波数構造は低ランク行列になる」という仮定を考えている。このような音源モデルは音声信号だけでなく音楽信号にもよく適合することから、ILRMA の登場によって多くの場合において IVA よりも高品質な BSS を達成することができるようになった。

しかし、声質の近い複数音声の混合や、音源数が $N \geq 4$ となる過酷な条件においては、IVA や ILRMA を用いてもしばしば分離に失敗してしまう。これは、各音源信号の時間周波数成分がダイナミックに変動することから、IVA や ILRMA が仮定する音源モデルが同一音源の時間周波数成分を正しく捉えられないことに起因していると思われる。例えば、IVA や ILRMA において、まとまった周波数帯域でパーミュテーションが入れ替わる問題（ブロックパーミュテーション問題）[16] が報告されている。Fig. 2.4 にブロックパーミュテーション問題の様子を示す。Fig. 2.4 では、4 kHz 以上の周波数帯がまとまって入れ替わった状態で分離信号が推定されてしまっている。このような事実からも、依然としてパーミュテーション問題の解決は不十分であり、更なる高精度なパーミュテーション解決法の模索が重要であることが分かる。Fig. 2.4 のような明らかなブロックパーミュテーションであれば、ユーザアノテーションにより修正するインタラクティブな BSS アルゴリズム [17] も適用可能であるが、多くの帯域にブ

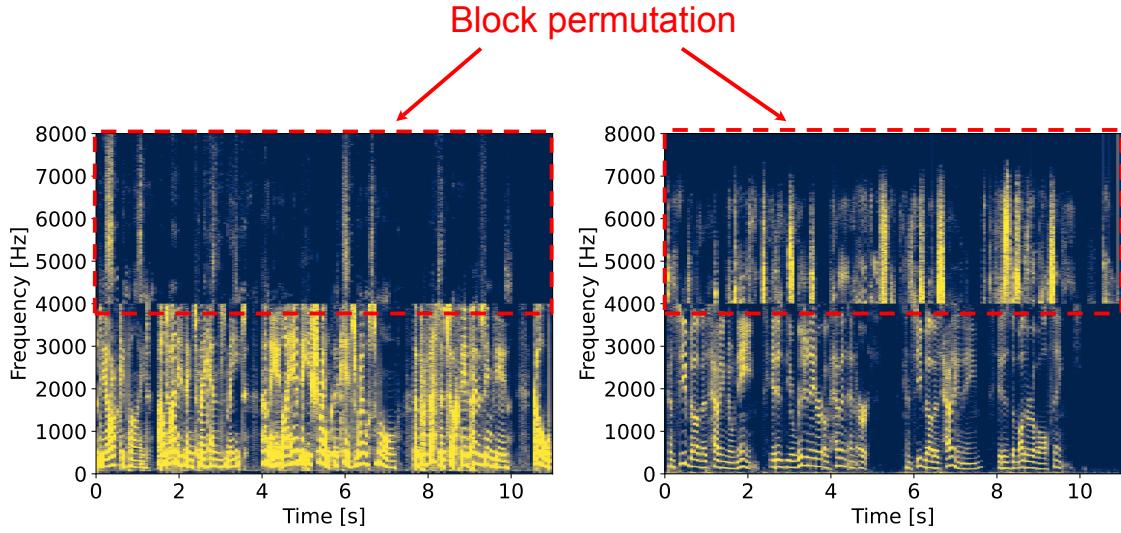


Fig. 2.4: Example of block permutation problem.

ロックパーミュテーション問題が発生する場合もあり、ユーザアノテーションの利用も難しい状況が存在する。

2.8 深層パーミュテーション解決法

前節で述べた通り、IVA や ILRMA のように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方法は頑健性や汎化性という観点で課題が残る。観測信号中に混合している各音源信号の時間周波数構造に合致した適切な音源モデルが仮定できれば高性能となる反面、合致しなければロックパーミュテーション問題を引き起こしてしまう。

この問題を解決するために、様々な種類の音源のデータからパーミュテーション問題を解決する最適なモデルを学習するアプローチ（深層パーミュテーション解決法）が近年提案された[14]。この手法では、学習データのパーミュテーション問題を解く DNN を構築することで、あらゆる種類の観測信号に対しても高精度にパーミュテーション問題を解くことを目指している。但し、全周波数ビンを一度に取り扱いパーミュテーション問題を解くことは DNN を用いてもなお困難であったため、前段で一定幅の周波数帯域（サブバンド）内のパーミュテーション問題の解決を様々なサブバンドに適用し、後段でサブバンド間のパーミュテーション問題をスティッ칭 [18] により解決するという複雑な二段階処理のアルゴリズムとなっている。さらに、前段のサブバンド内のパーミュテーション解決さえも困難であったことから、「参照周波数ビンとその他の周波数ビンの推定信号成分が同一音源か否か」という 2 クラス分類 DNN を学習しており、これに起因して音源数が $N \geq 3$ の場合は後段のサブバンド間のスティッ칭が非常に複雑・煩雑なアルゴリズムとなってしまう問題をはらんでいる。そのため、文献[14]の深層パーミュテーション解決法は $N = 2$ の場合を想定しており、一般的な

BSS への応用は難しい。

しかしながら、学習データを活用した深層ペーミュテーション解決法というアプローチは、前述の通り多様な音源信号に対して適用できる可能性があるという観点で深い意義がある。本論文においても、次章の動機で述べる通り、深層ペーミュテーション解決法の可能性を基礎実験的に調査し、その有用性について検証する。

2.9 本章のまとめ

本章では、提案手法において必要となる基礎理論及び各種従来手法について説明した。2.2 節では、ICA の基本原理と分離信号における順序とスケールの任意性について説明した。2.3 節では、音響信号処理でよく用いられる手法である STFT について説明した。2.4 節では、各信号の成分を時間周波数毎に定式化を行い、2.5 節以降で用いる。2.5 節では、時間周波数領域での周波数毎に ICA を適用することで音源分離を行う FDICA について説明した。2.6 節では、FDICA に伴い生じるペーミュテーション問題について説明した。2.7 節では、ペーミュテーション問題を可能な限り回避するような手法である、BSS の IVA と ILRMA について説明した。そして、2.8 節では既存の深層ペーミュテーション解決法の概要と問題点について述べた。次章以降では、本論文で提案する新しい深層ペーミュテーション解決法の動機とアルゴリズムについて詳しく述べる。

第3章

提案手法

3.1 まえがき

前章では、音響信号の BSS において重要な FDICA のパーミュテーション問題について詳しく述べた。また、音源モデルに基づきパーミュテーション問題を回避する手法や、近年提案された深層パーミュテーション解決法について説明した。さらに、既存の深層パーミュテーション解決法では、音源数 N の増加に伴ってアルゴリズムが極端に複雑になってしまう課題について述べた。本章では、音源数 N が増加した場合でもアルゴリズムが複雑化することのない深層パーミュテーション解決法を新たに提案する。まず 3.2 節では、BSS において深層学習を用いてパーミュテーション問題の解決を目指す動機について述べる。3.3 節及び 3.4 節では、本論文で提案する深層パーミュテーション解決法の DNN モデルの入出力及びネットワーク構造をそれぞれ説明する。3.5 節及び 3.6 節では、誤差逆伝播に用いる損失関数の取り方とパーミュテーション行列を正確に推定するモデルを学習するための入力データ及び正解データ（ラベル）の取得方法をそれぞれ説明する。3.7 節で本章のまとめを述べる。

3.2 動機

文献 [13] では、IVA や ILRMA に基づく BSS の STFT における最適な短時間区間長（窓長） Q について実験的に調査している。Fig. 3.1(b) は、文献 [13] の実験結果の図を引用したものである。詳しい実験条件等は文献 [13] を参照されたい。縦軸は信号対歪み比（source-to-distortion ratio: SDR）[20] の改善量であり、これは即ち音源分離の性能を表している。この結果より、IVA 及び ILRMA では、残響時間が 470 ms という比較的残響の強い条件では、IVA も ILRMA も高精度な音源分離に失敗していることが分かる。一方で、FDICA に対して、音源信号 s_{ij} を用いる理想的なパーミュテーション解決法（ideal permutation solver: IPS）を適用した結果（すなわち FDICA の達成しうる限界性能）では 10 dB 以上の SDR の改善を達成している。この事実は、高残響下での音声信号の混合という難しい観測条件であっても、 \hat{W}_i の推定自体（すなわち周波数ビン毎の BSS）は FDICA でも高精度に実現できてい

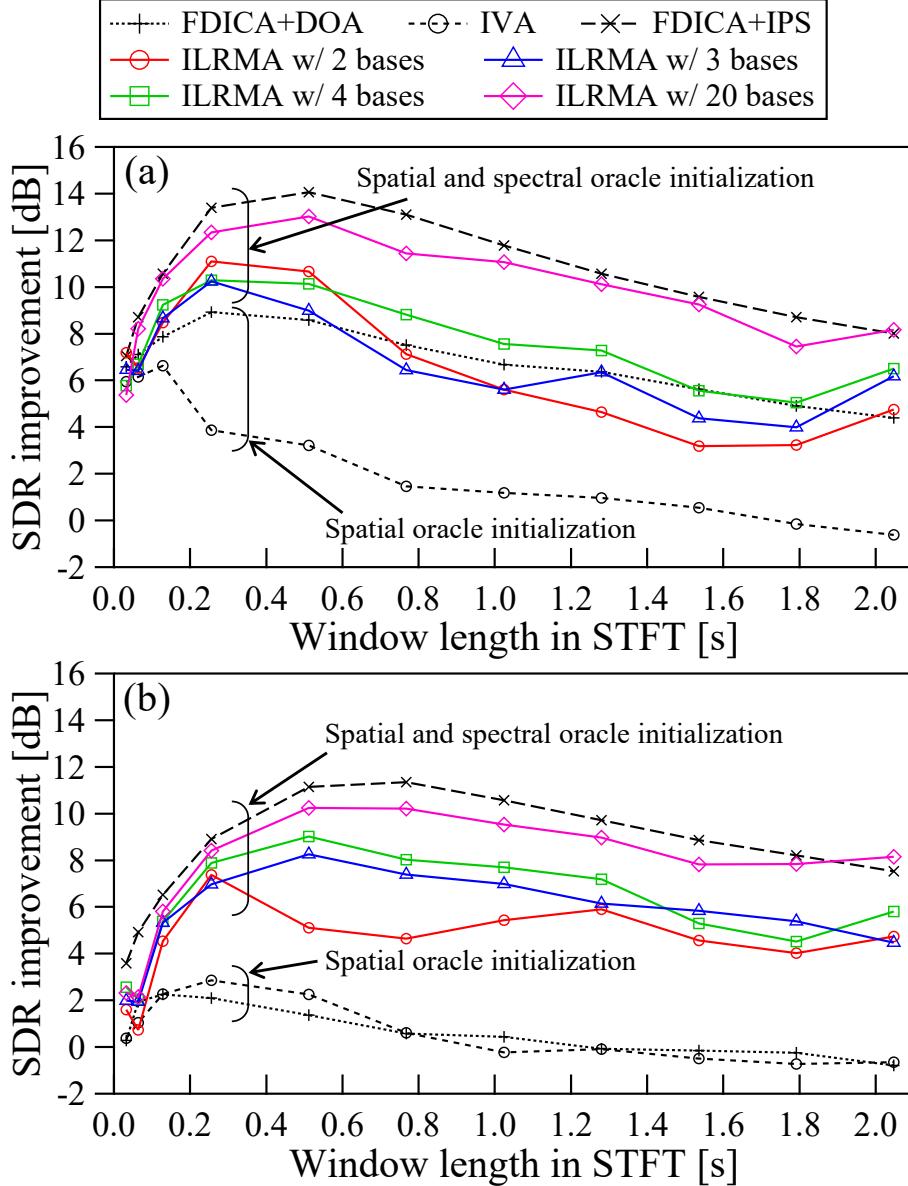


Fig. 3.1: Average source separation results for speech signals using random initialization: (a) E2A ($T_{60} = 300$ ms) and (b) JR2 ($T_{60} = 470$ ms) impulse responses. For details of this figure, see [13].

ることを示している。すなわち、残る課題は推定信号 \mathbf{y}_{ij} を正しい順番に並び変えるパーミュテーション問題の解決 (\mathbf{P}_i^{-1} の推定) のみであることを示唆している。また、2.8 節で述べた通り、従来の深層パーミュテーション解決法では、サブバンド内のパーミュテーション問題を解決する際に、参照周波数ビンに対して他の周波数ビンの推定信号成分が同一音源の成分か否かの 2 クラス分類問題を DNN で予測している。音源数が $N = 2$ であれば、この「同一音源の成分か否か」の 2 クラス分類はすなわち「どちらの音源の成分か」に一致するが、音源

数が $N \geq 3$ となった場合は、「同一音源の成分ではない」と DNN が判断した場合にその成分がどの音源の成分かが確定しない。従って、この場合に各推定成分がどの音源に対応するかを確定させるためには、先の 2 クラス分類 DNN モデルを音源数 N 個の中から 2 つ選ぶ組み合わせ数 (${}_N C_2$) 分適用せねばならず、さらに後段のサブバンド間のパーミュテーション問題の解決（全サブバンドのスティッ칭）の処理を考えると、そのアルゴリズムは非常に複雑・煩雑になってしまう。

そこで、本論文では、簡潔なアルゴリズムでパーミュテーション問題を正確に解くことに焦点を当て、新しい深層パーミュテーション解決法を提案する。以後、本論文では、提案するパーミュテーション問題の解決法が実現可能かどうかを判断するための基礎的な調査として、FDICA を適応した後の分離信号を模倣した人工データと実際の音響信号を用いてパーミュテーション問題の解決性能を実験的に調査する。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、本論文は基礎的な実験に終始するため、音源数及びチャネル数が $N = M = 2$ の状況のみを取り扱う。 $N \geq 3$ 以上の条件での調査については今後の課題となる。

本論文で提案する深層パーミュテーション解決法を適用する処理の概要は以下の通りである。

- (a) パーミュテーション問題が未解決の状態である推定信号 \mathbf{Y}_1 及び \mathbf{Y}_2 に対し、両信号のパワー比に基づく正規化 [5] を施す
- (b) 正規化された両信号のスペクトログラムから、ある時間フレーム j とその前後 $j \pm \beta$ の時間フレームの部分的なスペクトログラムを抽出し、時間フレーム j を中心とした局所時間振幅スペクトログラムを両信号で構成する
- (c) 両信号の局所時間振幅スペクトログラムをベクトル化し、DNN に入力する。
- (d) DNN は入力ベクトル中の \mathbf{Y}_1 及び \mathbf{Y}_2 の正規化局所時間振幅スペクトログラムの各周波数ビンの成分がそれぞれどの音源信号に属するかを分類問題として予測し、周波数毎及び音源毎の確率値をまとめたベクトルを出力する
- (e) (b)–(d) の処理を全時間フレームに対して適用し、時間フレーム毎の確率値ベクトルを取得する
- (f) 全時間フレームの確率値ベクトルを用いて時間方向に多数決処理を適用し、全時間フレーム共通の（1 本の）確率値ベクトルを得る
- (g) 確率値ベクトルから周波数毎のパーミュテーション行列 \mathbf{P}_i の推定値 $\hat{\mathbf{P}}_i$ を構成する
- (h) 式 (2.33) よりパーミュテーション問題が解決された分離信号を得る

上記の処理の詳細や DNN の学習方法については、次節以降で詳しく述べる。

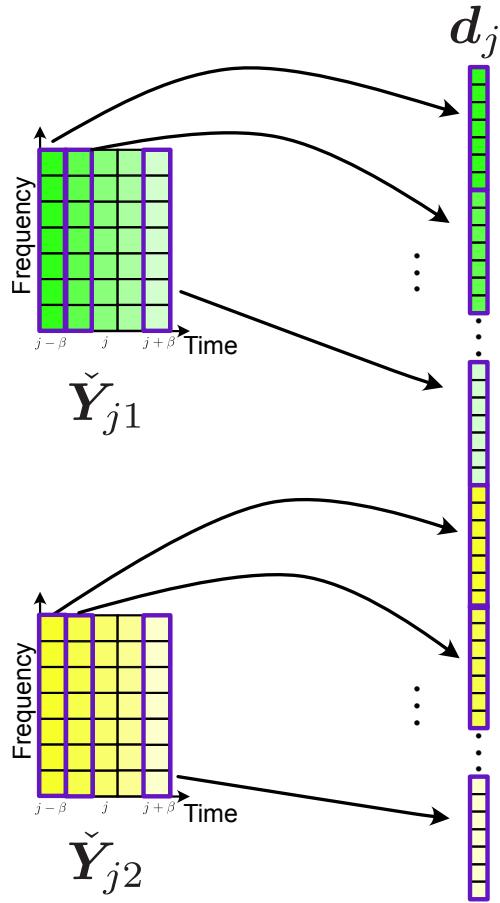


Fig. 3.2: Input vector of DNN.

3.3 DNN の入出力

提案する深層パーミュテーション解決法で用いられる DNN は複数の全結合層からなる多層パーセプトロン (multi-layer perceptron: MLP) を想定している。MLP の入出力はあらかじめ決められた次元数のベクトルでなければならない。今、観測信号 $(\mathbf{X}_1, \mathbf{X}_2)$ に FDICA を適用した場合を考える。FDICA からは、パーミュテーション問題が発生した状態の推定信号 $(\mathbf{Y}_1, \mathbf{Y}_2)$ が得られる。ここで、同一音源に属する成分の相関を強調するため、推定信号 $(\mathbf{Y}_1, \mathbf{Y}_2)$ をパワースペクトログラム ($|\mathbf{Y}_1|^2, |\mathbf{Y}_2|^2$) の比率に変換する正規化 [5] を施す。この処理は次式で表される。

$$\overline{\mathbf{Y}}_1 = \frac{|\mathbf{Y}_1|^2}{|\mathbf{Y}_1|^2 + |\mathbf{Y}_2|^2} \in [0, 1]^{I \times J} \quad (3.1)$$

$$\overline{\mathbf{Y}}_2 = \frac{|\mathbf{Y}_2|^2}{|\mathbf{Y}_1|^2 + |\mathbf{Y}_2|^2} \in [0, 1]^{I \times J} \quad (3.2)$$

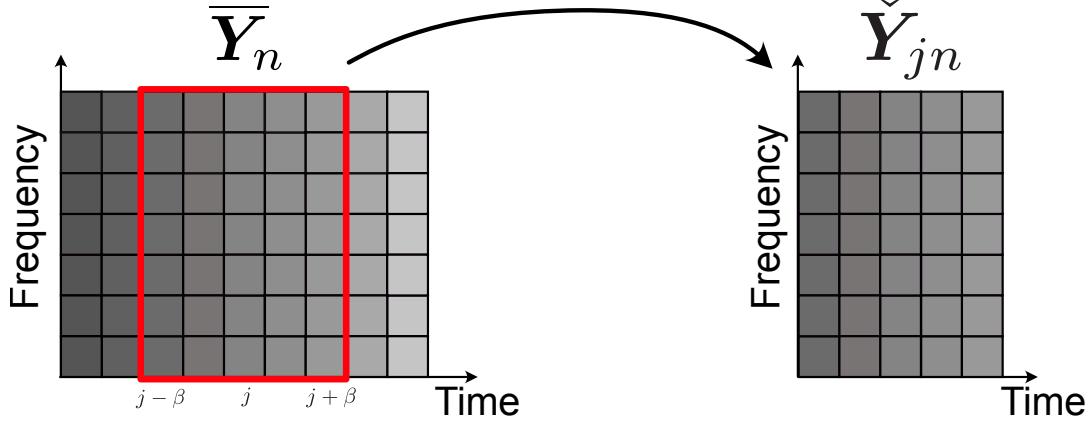


Fig. 3.3: Extraction of local-time-frame amplitude spectrogram.

ここで、行列に対する絶対値記号は要素毎の絶対値、行列やベクトルに対するドット付き指數乗は要素毎の指數乗、及び行列間のベクトルは要素毎の商を示している。このような正規化は、文献 [5] で詳しく解析されているように同一音源に属する成分の相関を強調させる利点があるだけでなく、推定信号の値が区間 $[0, 1]$ の範囲に限定されることから、DNN の学習を安定させる効果も期待できる。次に、推定信号の正規化振幅スペクトログラム (\bar{Y}_1, \bar{Y}_2) から、Fig. 3.3 に示すように、時間フレーム j を中心とする局所時間振幅スペクトログラムを抽出する。この処理は次式で表される。

$$\check{Y}_{j1} = [\bar{y}_{(j-\beta)1}, \bar{y}_{(j-\beta+1)1}, \dots, \bar{y}_{(j-1)1}, \bar{y}_{j1}, \bar{y}_{(j+1)1}, \dots, \bar{y}_{(j+\beta)1}] \in [0, 1]^{I \times (2\beta+1)} \quad (3.3)$$

$$\check{Y}_{j2} = [\bar{y}_{(j-\beta)2}, \bar{y}_{(j-\beta+1)2}, \dots, \bar{y}_{(j-1)2}, \bar{y}_{j2}, \bar{y}_{(j+1)2}, \dots, \bar{y}_{(j+\beta)2}] \in [0, 1]^{I \times (2\beta+1)} \quad (3.4)$$

ここで、 $\bar{y}_{jn} \in [0, 1]^I$ は正規化振幅スペクトログラム \bar{Y}_n の j 列目の列ベクトル（時間フレーム j の正規化振幅スペクトログラム）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パラメータである。提案手法では、DNN の入力ベクトルは、式 (3.3) 及び (3.4) で得られる両信号の正規化局所時間振幅スペクトログラム $(\check{Y}_{j1}, \check{Y}_{j2})$ を Fig. 3.2 のように一次元に整形（ベクトル化）したベクトルである。入力された行列をベクトル化する処理を $\text{vec}(\cdot)$ と表記すると、DNN の入力ベクトルは次式となる。

$$\mathbf{d}_j = \begin{bmatrix} \text{vec}(\check{Y}_{j1}) \\ \text{vec}(\check{Y}_{j2}) \end{bmatrix} \in [0, 1]^{2I(2\beta+1)} \quad (3.5)$$

DNN による予測は次式で表される。

$$\hat{\mathbf{l}}_j = \text{DNN}(\mathbf{d}_j) \in [0, 1]^{2I} \quad (3.6)$$

ここで、 $\hat{\mathbf{l}}_j = [\hat{l}_{11j}, \hat{l}_{21j}, \dots, \hat{l}_{I1j}, \hat{l}_{12j}, \hat{l}_{22j}, \dots, \hat{l}_{I2j}]^T$ は出力である予測ベクトルを表す。入力されたベクトルを行列化する処理を $\text{mat}(\cdot)$ と表記すると、予測ベクトルは次式で再成型される。

$$\hat{\mathbf{L}}_j = \text{mat}(\hat{\mathbf{l}}_j) \in [0, 1]^{I \times 2} \quad (3.7)$$

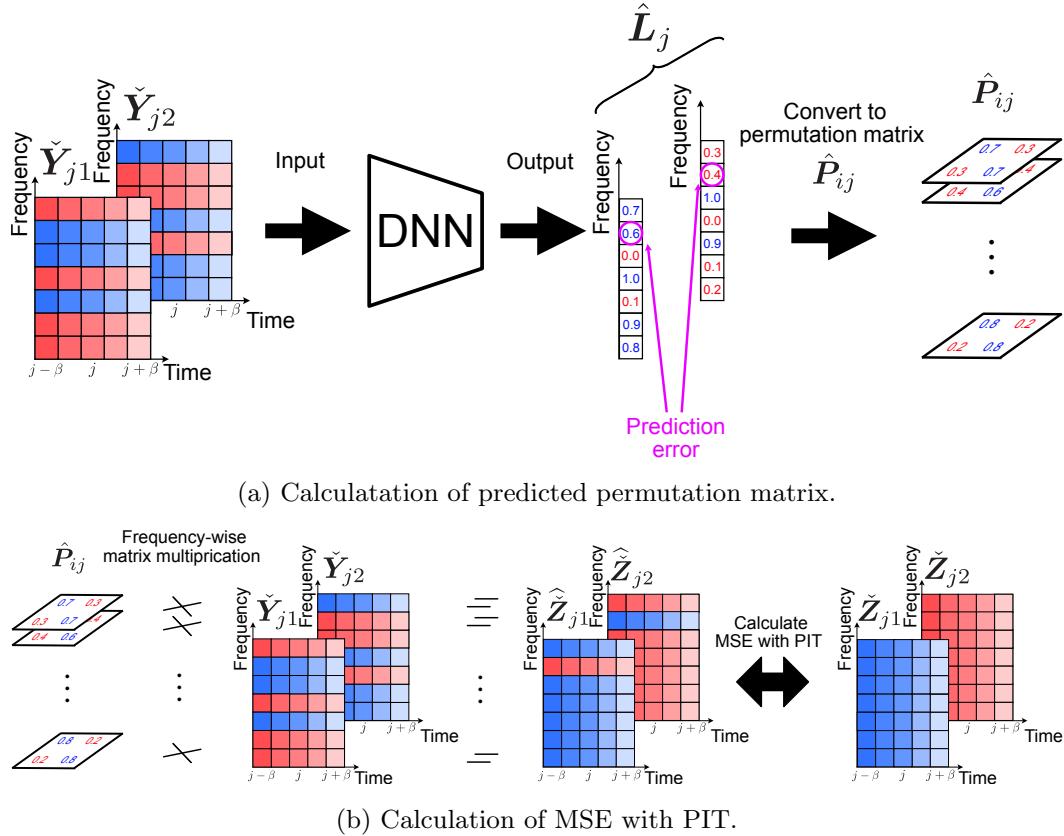


Fig. 3.4: Process of calculating predicted permutation matrix and loss function value.

再成型された行列 \hat{L}_j は Fig. 3.4(a) に示すように、2つのパーミュテーション問題が生じている入力信号 ($\check{Y}_{j1}, \check{Y}_{j2}$) の各周波数成分のそれぞれが「1番目の音源の成分である確率 l_{i1} 」と「2番目の音源の成分である確率 l_{i2} 」を d_j から予測したものと定義し、提案手法ではこの定義に基づいて正確な予測ができる DNN を学習する。ここで、 (l_{i1}, l_{i2}) は離散確率値であるため $l_{i1} + l_{i2} = 1$ を満たし、それらの予測値である $(\hat{l}_{i1j}, \hat{l}_{i2j})$ もまた $\hat{l}_{i1j} + \hat{l}_{i2j} = 1$ を満たすように DNN の中で制約する必要がある。この制約は次節で述べる通り、softmax 関数を用いて実現できる。また、詳細は後述するが、パーミュテーション問題の解は時間方向には変化しない（式 (2.28) における P_i は時間フレーム j によらない時不变行列である）ため、様々な局所時間振幅スペクトログラムの入力 d_j の予測結果 \hat{L}_j を j に関して多数決処理することで、より精度の高い予測である予測結果 \hat{L} （この結果は j によらない）を生成できる。

重要なこととして、確率値 (l_{i1}, l_{i2}) は式 (2.27) で述べたパーミュテーション行列それ自身と本質的に等価である。従って、DNN の予測結果である $(\hat{l}_{i1}, \hat{l}_{i2})$ から推定パーミュテーション行列を次式で構成できる。

$$\hat{P}_i = \begin{bmatrix} \hat{l}_{i1} & \hat{l}_{i2} \\ \hat{l}_{i2} & \hat{l}_{i1} \end{bmatrix} \quad (3.8)$$

ここで、 \hat{l}_{i1} 及び \hat{l}_{i2} は \hat{L} の要素である。正解のパーミュテーション行列は順列を並び替える

行列であるため、 $N = 2$ の場合は式 (2.29) のいずれかとなる。推定パーミュテーション行列 $\hat{\mathbf{P}}_i$ は式 (3.8) であるため、予測が不完全であれば \mathbf{I} 又は $\mathbf{1} - \mathbf{I}$ にはならない可能性があるが、それでも $\hat{l}_{i1} + \hat{l}_{i2} = 1$ を満たすため、二重確率行列 (doubly stochastic matrix: DSM) であることがわかる。また、Birkhoff–von Neumann の定理（付録 A 参照）を考慮すると、パーミュテーション問題の発生している入力データから DSM を予測する提案手法の DNN は、考えうる全てのパーミュテーション行列に対する凸結合係数を推定していることになる。即ち、考えうるパーミュテーション行列の中でどの行列が正解かという確信度を予測していると解釈することもできる。

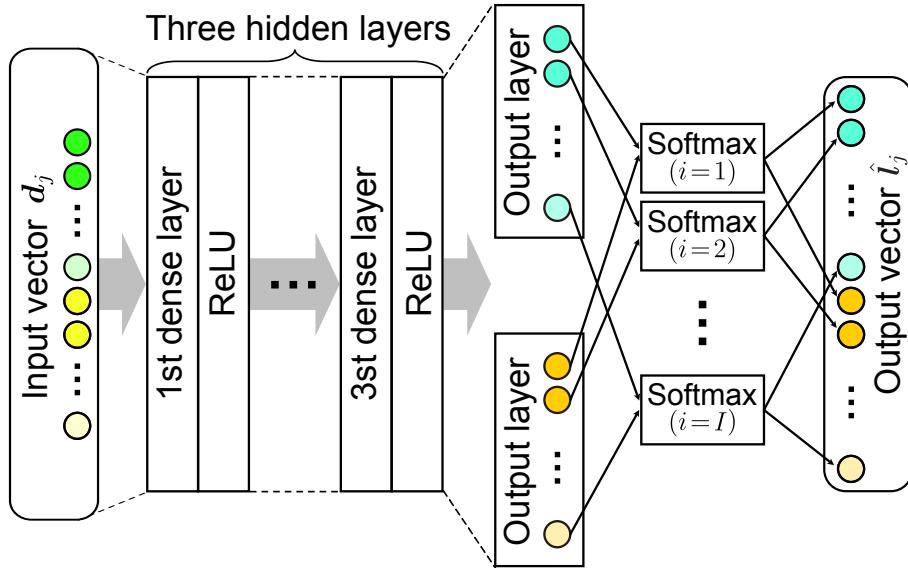


Fig. 3.5: DNN architecture.

3.4 DNN の構造

Fig. 3.5 に提案深層パーミュテーション解決法で用いる DNN の構造を示す。この DNN は、入力層、隠れ層 3 層、及び出力層の計 5 層の全結合層 (dense layer) からなる MLP となっており、隠れ層の 1 層目から 3 層目には非線形関数として rectified linear unit (ReLU) [21] 関数を用いている。また、3 層目の隠れ層から出力層に変換する際には、Fig. 3.5 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^{*1}。その後、2 つの I 次元の同一インデックスの要素に対して 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ックスの要素の和が 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し、これによって予測ベクトルを確率値としてみなすことが可能となる。

3.5 DNN 学習時の損失関数

DNN の学習は、何らかの損失関数を定義しその値を最小化するパラメータを誤差逆伝播により推定する処理となる。提案手法の DNN は 3.3 節で述べた通り、入力データから周波数毎の正しい音源パーミュテーションを予測するモデルである。これは（音源数が $N = 2$ であれば） (l_{i1}, l_{i2}) の 2 クラス分類器であるため、softmax 関数を用いて各クラスへの確率値を出力している。通常、多クラス分類器の損失関数には、カテゴリカル分布^{*2}の負対数尤度関数であ

^{*1} すなわち、 $2I$ 次元への全結合層による変換と同様であるが、明示的に分岐させて定義している。

^{*2} 多項分布における試行回数を 1 回とした際の分布である。

るカテゴリカル交差エントロピー (categorical cross entropy: CCE) を用いることで, DNN の学習を最尤推定の枠組みで行うことができる。しかしながら, 提案手法の深層パーミュテーション解決法の本来の目的は, 全周波数ビンにおいてパーミュテーション行列を正確に予測することではなく, 分離信号 (Z_1, Z_2) を正確に予測することである。例えば, 推定信号 (\hat{Y}_1, \hat{Y}_2) のどちらにもエネルギーがほとんど無いような周波数ビンは, 実際は誤った分離信号の順序となっていても得られる分離信号 (Z_1, Z_2) の音源分離精度には影響しない。もし CCE で DNN の損失関数を定義すると, このようなエネルギーが少ない (音源分離にとって重要ではない) 周波数ビンのパーミュテーション予測精度と, 大きなエネルギーを有する (音源分離にとって重要な) 周波数ビンの予測精度が等しい重要度で扱われることになるため, 音源分離性能向上の妨げとなる可能性がある。

そこで提案手法では, 下記で説明する通り, DNN で予測された音源パーミュテーションに基づいて推定信号 (\hat{Y}_1, \hat{Y}_2) を並び替えた予測分離信号 (\hat{Z}_1, \hat{Z}_2) と正解の分離信号 (Z_1, Z_2) の間の平均二乗誤差 (mean squared error: MSE) を示す。

Fig. 3.4(a) に損失関数の計算の処理の流れを示す。まず, 予測結果に対応する行列 $\hat{V} \in \mathbb{R}^{R \times C}$ とラベルに対応する行列 $V \in \mathbb{R}^{R \times C}$ の間の MSE を次式で定義する。

$$\text{MSE}(\hat{V}, V) = \frac{1}{RC} \|\hat{V} - V\|_{\text{Fr}}^2 \quad (3.9)$$

$$= \frac{1}{RC} \sum_{r,c} (\hat{v}_{rc} - v_{rc})^2 \quad (3.10)$$

ここで, \hat{v}_{rc} 及び v_{rc} はそれぞれ行列 \hat{V} 及び V の要素, $r = 1, 2, \dots, R$ 及び $c = 1, 2, \dots, C$ はそれぞれ行列 \hat{V} 及び V の行と列のインデックス, $\|\cdot\|_{\text{Fr}}$ は Frobenius ノルムである。次に, Fig. 3.4(a) に示すように, DNN の入力である正規化局所時間振幅スペクトログラム ($\check{Y}_{j1}, \check{Y}_{j2}$) に対する予測結果 L_j と式 (3.8) を用いて, (j を中心とする局所時間フレームの) 推定局所時間パーミュテーション行列 \hat{P}_{ij} を構成する。また, Fig. 3.4(b) に示すように, 式 (2.28) で音源パーミュテーションを並び替えた予測分離信号 ($\hat{\check{Z}}_{j1}, \hat{\check{Z}}_{j2}$) を求める。さらに, この予測分離信号に対する正解ラベル (Fig. 3.3 と同様の手順で, 分離信号 (Z_1, Z_2) から j を中心とする局所時間フレームの局所時間振幅スペクトログラムを抽出した行列) を ($\check{Z}_{j1}, \check{Z}_{j2}$) と定義する。これらの信号と式 (3.10) を用いて, 前述の誤差関数 \mathcal{L} は, $(\hat{\check{Z}}_{j1}, \hat{\check{Z}}_{j2})$ 及び $(\check{Z}_{j1}, \check{Z}_{j2})$ 間の MSE として次式で表せる。

$$\mathcal{L} = \text{MSE}(\hat{\check{Z}}_{j1}, \check{Z}_{j1}) + \text{MSE}(\hat{\check{Z}}_{j2}, \check{Z}_{j2}) \quad (3.11)$$

但し, パーミュテーション問題の解決は全周波数で推定音源成分を正しく並び替えることだけが目標であり, 並び替えた後の分離信号そのものの順序は予測の対象としない。すなわち, 深層パーミュテーション解決法を適用した結果が, (Z_1, Z_2) 及び (Z_2, Z_1) のどちらの順序で出力されようとも構わない。式 (3.11) で損失関数を定義した場合, 分離信号は必ず (Z_1, Z_2) という順序で予測することを DNN に強いているため, この問題を解消するために順序不变学習 (permutation invariant training: PIT) [19] を導入する。具体的には, 損失関数を次式で

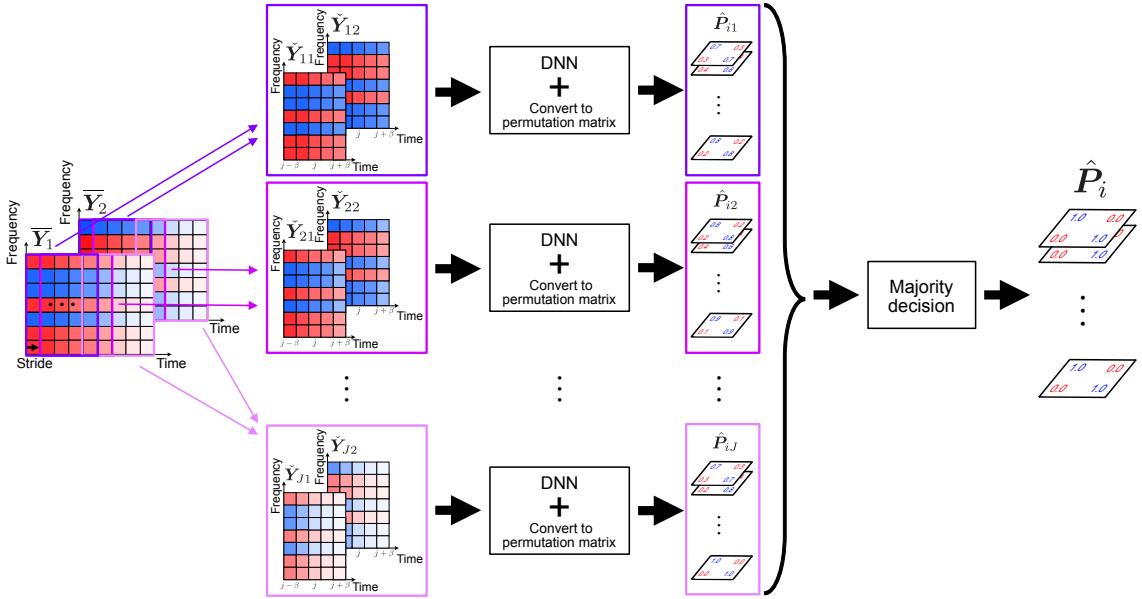


Fig. 3.6: DNN predictions for all local-time-frame amplitude spectrograms and their majority decision.

定義する。

$$\mathcal{L} = \min \left(\text{MSE}(\hat{\tilde{Z}}_{j1}, \check{Z}_{j1}) + \text{MSE}(\hat{\tilde{Z}}_{j2}, \check{Z}_{j2}), \text{MSE}(\hat{\tilde{Z}}_{j1}, \check{Z}_{j2}) + \text{MSE}(\hat{\tilde{Z}}_{j2}, \check{Z}_{j1}) \right) \quad (3.12)$$

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。この関数の誤差逆伝播は自動微分により実装される。このように、PIT を導入することで、周波数ビン間のパーミュテーション問題さえ解決されれば良く分離信号そのものの出力の順序には依存しないような学習が可能となる。

3.6 学習済の DNN のテストデータへの適用

DNN 学習後は、提案手法である深層パーミュテーション解決法を FDICA 等の推定信号 (\bar{Y}_1, \bar{Y}_2) に適用することができる。このテストデータへの適用時においては、より高精度にパーミュテーション問題を解決するために、次に示す 2 つの処理を施す。

- (a) FDICA 等で実現される周波数ビン毎の BSS が完全に達成されているならば、推定すべきパーミュテーション行列 P_i は 0 及び 1 の要素を持つバイナリ行列であるため、推定局所時間パーミュテーション行列 \hat{P}_{ij} もバイナリ行列に変換する
- (b) FDICA 等の時不变な分離行列 W_i を推定する BSS により生じるパーミュテーション問題は、時間フレーム方向には一定である (P_i は j に非依存) ため、推定局所時間パーミュテーション行列 \hat{P}_{ij} を時間方向に多数決処理し、時不变な行列 \hat{P}_i に変換する

上記 (a) については、次式でバイナリ行列への変換処理を実現する。

$$\hat{\mathbf{P}}_{ij} \leftarrow \text{round}(\hat{\mathbf{P}}_{ij}) \in \{0, 1\}^{N \times N} \quad (3.13)$$

ここで、 $\text{round}(\cdot)$ は入力された行列の各要素に関して四捨五入を適用する処理であり、また \leftarrow は変数の更新を表す。但し、式 (3.13) によるバイナリ行列への変換は、前段の周波数ビン毎の BSS が完全に達成されていることを仮定している。実際には FDICA でも周波数ビン毎の BSS には誤差が生じるため、式 (3.13) を適用すべきか否かは前段の BSS の性能に依存して決める必要がある。本論文では、次章の実験条件で述べる通り、前段の BSS が完全であることを仮定しているため、式 (3.13) の処理を適用している。

一方、上記 (b) については、純粋にパーミュテーション問題の解決精度の向上に寄与する処理である。DNN に入力する局所時間振幅スペクトログラム $(\check{\mathbf{Y}}_{1j}, \check{\mathbf{Y}}_{2j})$ は推定信号 $(\mathbf{Y}_1, \mathbf{Y}_2)$ の各時間フレームにおいて抽出できるため、Fig. 3.6 に示すように $(\check{\mathbf{Y}}_{1j}, \check{\mathbf{Y}}_{2j})$ の抽出範囲をストライドさせ、その全てである $((\check{\mathbf{Y}}_{1j}, \check{\mathbf{Y}}_{2j}))_{j=1}^J$ を個々に DNN に入力し、全ての予測結果 $((\hat{\mathbf{Z}}_{1j}, \hat{\mathbf{Z}}_{2j}))_{j=1}^J$ を得ることができる。これらの予測結果を推定パーミュテーション行列 $(\hat{\mathbf{P}}_{ij})_{j=1}^J$ に変換し、次式の多数決処理を適用する。

$$\hat{\mathbf{P}}_i = \text{round} \left(\frac{1}{J} \sum_{j=1}^J \hat{\mathbf{P}}_{ij} \right) \quad (3.14)$$

なお、式 (3.13) のバイナリ行列への変換を適用しない場合においても、式 (3.14) を計算することで時間方向の平均化ができるため、式 (3.14) は上記 (a) の適用の有無にかかわらず計算することが望ましい。

3.7 本章のまとめ

本章では、FDICA のポスト処理として DNN に基づくパーミュテーション解決法について提案した。3.2 節では、FDICA において理想的なパーミュテーション解決法を適用した場合、高精度で音源分離が可能となることを説明した。3.3 節では、DNN の入力に局所時間振幅スペクトログラムを用いることと、同一音源に属する成分の相関を強調させるため正規化を行うことを説明した。3.4 節では、隠れ層 3 層の全結合層からなる DNN の構造について説明した。3.5 節では、DNN の予測に従ってパーミュテーション行列を作成した後、推定信号を並び替えた予測分離信号と正解の分離信号との間で損失を取得することを説明した。3.6 節では、テストデータに対して時間方向に多数決処理を行うことで、パーミュテーション問題の解決精度を向上させることを説明した。

第4章

実験

4.1 まえがき

前章で提案した DNN に基づくパーミュテーション解決法の有効性を確認するために、人工的に作成したデータと実際の音声及び音楽信号を用意し、提案パーミュテーション解決法を適用してその性能を評価した。4.2 節では、本実験における条件を詳細に示し、4.3 節では提案手法のパーミュテーション解決性能を示している。4.4 節で本章のまとめを述べる。

4.2 実験条件

本実験では、提案する深層パーミュテーション解決法において、どの程度各周波数成分の正しい並び替えができるかを実験的に確認した。本実験では、まず最初に基礎実験として、音響信号ではなく人工的に作成した 2 次元の行列を用いた性能評価を実施した。人工的に作成した行列を 2 つ用意してパーミュテーション問題を模擬することで推定信号を生成し、これらを提案手法に入力してどの程度パーミュテーション問題が解決されるかを評価した。この基礎実験における詳細な実験条件については 4.2.1 項に示す。次に、実際の音響信号（音声及び音楽信号）の振幅スペクトrogramに対する性能評価も実施した。基礎実験の場合と同様にパーミュテーション問題を模擬し、提案手法に入力して性能を評価した。実際の音響信号を用いた実験の詳細な実験条件については、4.2.2 項に示す。

4.2.1 人工データを用いた基礎実験の条件

基礎実験ではまず、パーミュテーション問題の生じていない（完全に解決された状態の）分離信号 (Z_1, Z_2) として、Figs. 4.1–4.3 に示す 3 種類の 2 次元行列のペアを用いた。これらはいずれもサイズが $I = J = 100$ であり、それぞれ下記の構造を持っている。

- 全成分が 0 の Z_1 と全成分が 1 の Z_2 (Fig. 4.1)
- 25 列毎に 0 と 1 が入れ替わる Z_1 及び Z_2 (Fig. 4.2)

- 1列毎に 0 と 1 が入れ替わる Z_1 及び Z_2 (Fig. 4.3)

次に, 分離信号 (Z_1, Z_2) の各行ベクトルを音源間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を模擬し, パーミュテーション問題が生じている推定信号 (Y_1, Y_2) を生成した. このとき, FDICA で一般的に生じるような (Fig. 2.3 のような) 1行単位でランダムに入れ替わるパーミュテーション問題だけでなく, Fig. 2.4 のような複数行のブロック単位でまとめて入れ替わるブロックパーミュテーション問題を模擬した推定信号 (Y_1, Y_2) も生成した. ブロックパーミュテーション問題を模擬する際の各ブロックの行数は, Fig. 4.4 に示すように γ 行と定義し, これを $\gamma = 1, 2, 4, 8$ の 4 種類として実験した ($\gamma = 1$ は通常の 1行毎のパーミュテーション問題に対応). 従って, 3種類の分離信号のペア (Z_1, Z_2) と 4種類のブロック単位の合計 12 種類の実験条件を用意した.

本基礎実験では, 前述の 12 種類の各実験条件のそれぞれに対して専用の深層パーミュテーション解決法 (DNN) を学習した. 従って, 各 DNN を学習する際に用いる学習データは, 各実験条件に従う推定信号 (Y_1, Y_2) (入力) 及び分離信号 (Z_1, Z_2) (ラベル) であり, 入力はパーミュテーション問題を模擬するランダムな入れ替えを 300 パターン (重複無し) 生成することで学習データを構築した. この 300 パターンの推定信号 (Y_1, Y_2) に対するラベルはパーミュテーション問題が解決された信号であるため, 常に同じ分離信号 (Z_1, Z_2) となる. また, 性能評価に用いる検証データとテストデータは同一とし, 学習データの 300 パターンには含まれていないランダムな入れ替えでパーミュテーション問題を模擬した 1 パターンの推定信号 (Y_1, Y_2) を用いた. DNN の最適化法には Adam [23] を用いる. Adam の重みの最適化アルゴリズムを次式に示す.

$$\mathbf{g}^{(t)} = \nabla \mathcal{L}(\mathbf{w}^{(t)}) \quad (4.1)$$

$$\mathbf{m}_t = \rho_1 \mathbf{m}_{t-1} + (1 - \rho_1) \mathbf{g}^{(t)} \quad (4.2)$$

$$\mathbf{v}_t = \rho_2 \mathbf{v}_{t-1} + (1 - \rho_2) (\mathbf{g}^{(t)})^2 \quad (4.3)$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \rho_1^t} \quad (4.4)$$

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \rho_2^t} \quad (4.5)$$

$$\Delta \mathbf{w}^{(t)} = -\frac{\eta}{\sqrt{\hat{\mathbf{v}}_t + \varepsilon}} \hat{\mathbf{m}}_t \quad (4.6)$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta \mathbf{w}^{(t)} \quad (4.7)$$

ここで, $\mathcal{L}(\mathbf{w})$ 及び \mathbf{w} はそれぞれ損失関数及び DNN の最適化変数をまとめたベクトルであり, \mathbf{m}_t 及び \mathbf{v}_t はいずれも過去の勾配変化を表すモーメンタムと呼ばれる量である. また, 上付き文字の t は最適化 (変数更新) の反復回数を表す. \mathbf{m}_t 及び \mathbf{v}_t が考慮されていることにより, 変数更新における振動を抑えながら高速かつ安定な最適化が可能となる. 式 (4.2)–(4.6) 中のハイパーパラメータはそれぞれ標準的な設定値である $\varepsilon = 1.0 \times 10^{-8}$, $\rho_1 = 0.9$, $\rho_2 = 0.999$, 及び学習率 $\eta = 0.001$ に設定した. その他の学習パラメータについては, バッチサイズを 8, エポック数を 1000 として誤差逆伝播による学習を行った.

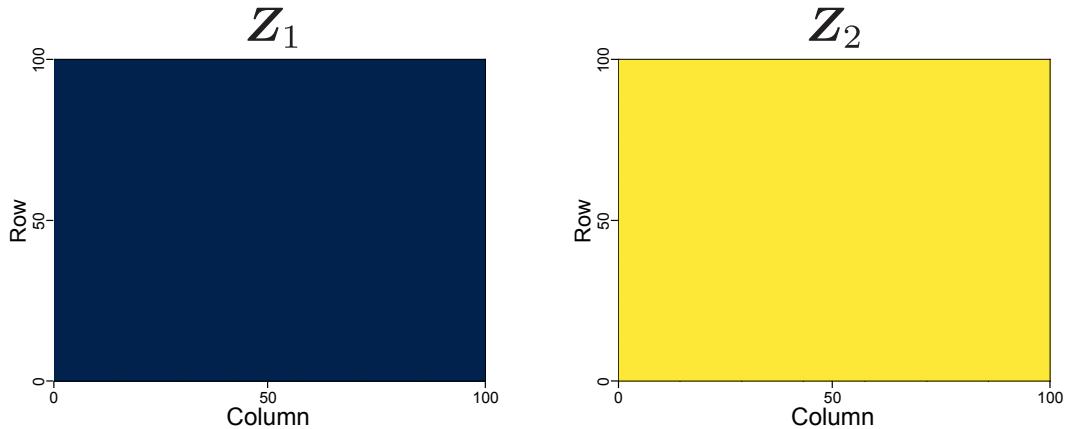


Fig. 4.1: Artificial source matrices (Z_1, Z_2) with only zero or one elements.

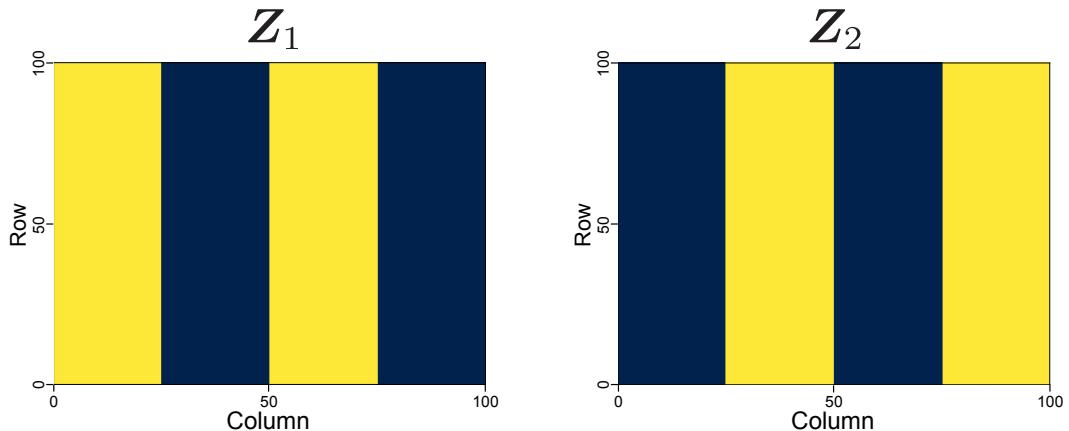


Fig. 4.2: Artificial source matrices (Z_1, Z_2) with zero and one elements swapping every 25 columns.

深層パーミュテーション解決法の性能を評価するための客観評価尺度には、正しく並び替えを行うことができた行の割合、即ち検証データに対する正答率を用いる。

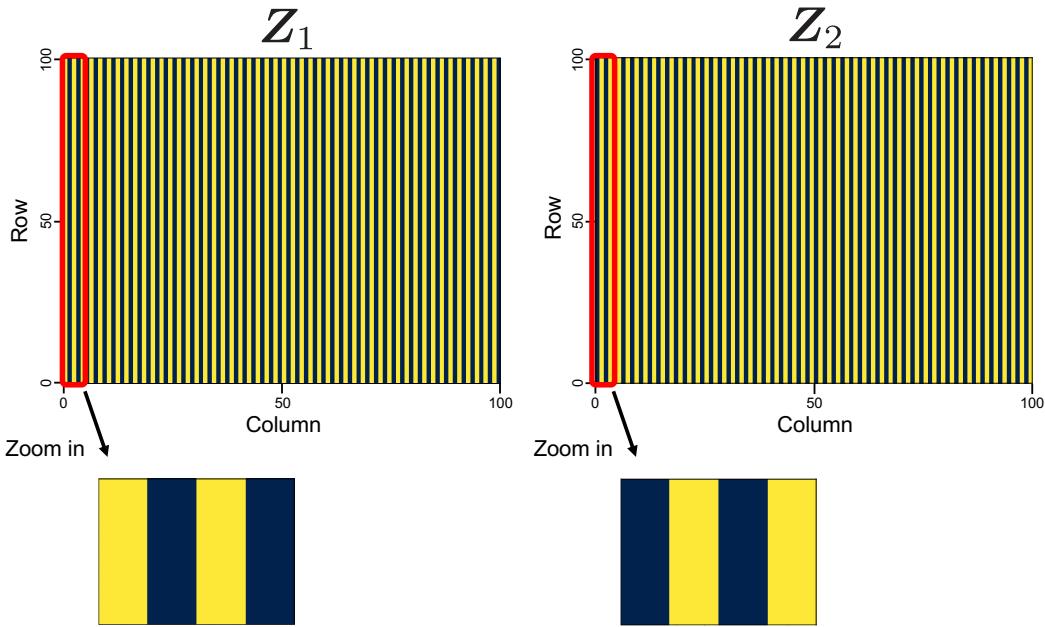


Fig. 4.3: Artificial source matrices (Z_1, Z_2) with zero and one elements swapping every columns.

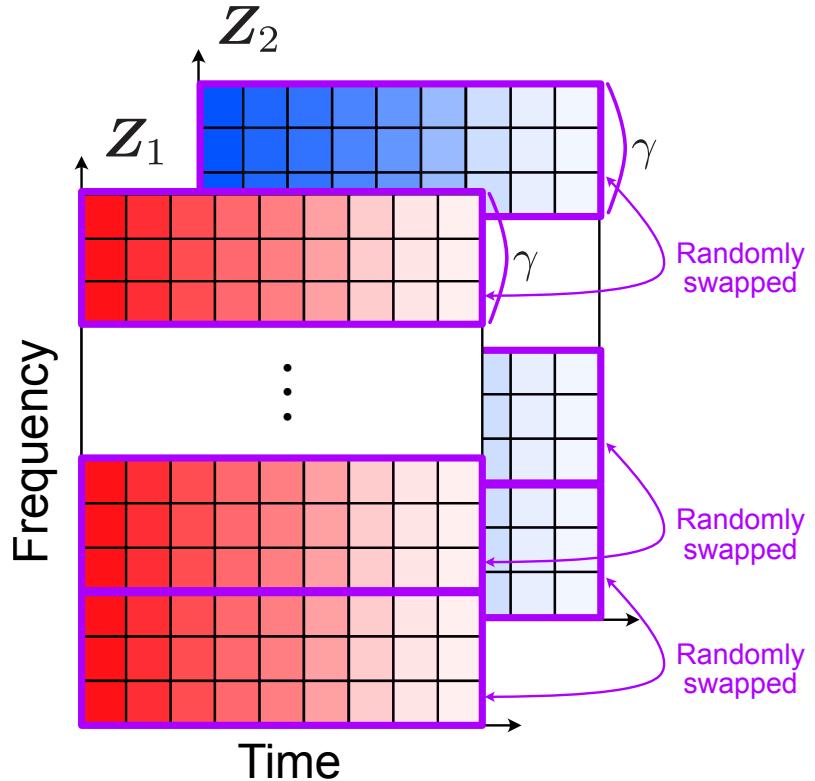


Fig. 4.4: Simulation of block permutation problems, where γ is row size of each block.

Table 4.1: Speech sources obtained from SiSEC2011

Signal type	Data name	Length [s]
Speech	dev3_female4_src_2	10.0
Speech	dev2_male4_src_2	10.0

Table 4.2: Music instrument sources obtained from SiSEC2011

Signal type	Data name	Length [s]
Piano	dev2_nodrums_liverec_250ms_src_3	11.0
Drums	dev2_wdrums_liverec_250ms_src_3	11.0

4.2.2 実際の音響信号を用いた実験の条件

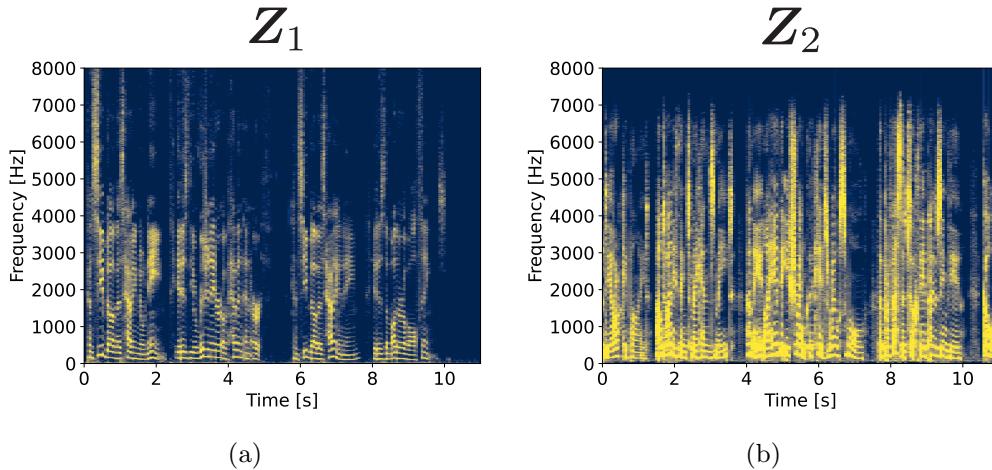


Fig. 4.5: Spectrograms of speech sources: (a) female and (b) male.

実際の音響信号を用いた実験では、パーティション問題の生じていない（完全に解決された状態の）分離信号 (Z_1, Z_2) として、Table 4.1 に示す男女の音声信号及び Table 4.2 に示すドラムとピアノの音楽信号を用いた。これらの信号のスペクトログラムはそれぞれ Figs. 4.5 及び 4.6 にそれぞれ示している。両信号のサンプリング周波数は 16 kHz であり、STFT における分析窓関数長（短時間信号長）を $Q = 2048$ 点 (128 ms), シフト長を $\tau = 1024$ 点 (64 ms) と設定したため、スペクトログラムのサイズは周波数ビン数が両信号ともに $I = 1025$, 時間フレーム数が $J = 158$ (音声信号) 及び $J = 173$ (音楽信号) となった。これらの信号を用いた実験では、IVA や ILRMA で生じる可能性があるブロックパーティション問題を模

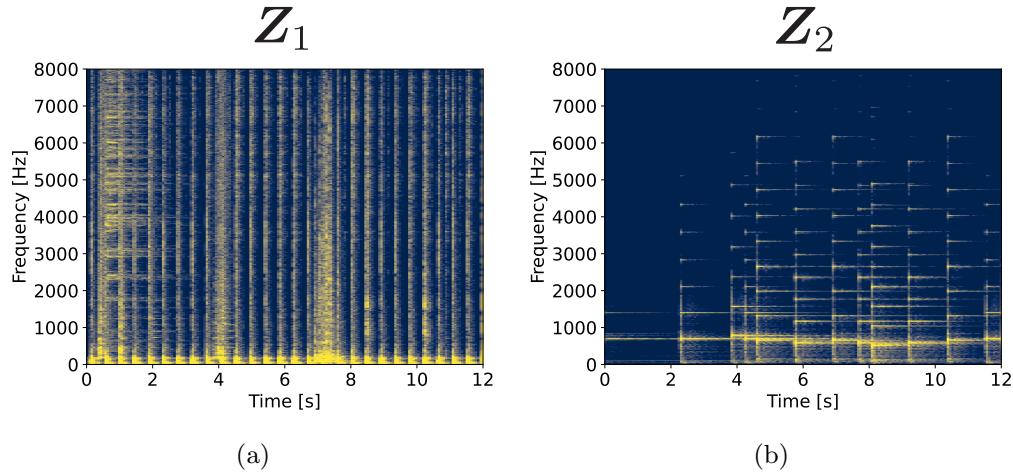


Fig. 4.6: Spectrograms of musical instrument sources: (a) drums and (b) piano.

擬した。具体的には、各ブロックの行数を $\gamma = 16$ 行とし、ブロック単位でまとめて入れ替わるブロックパーティーション問題を模擬した推定信号 $(\mathbf{Y}_1, \mathbf{Y}_2)$ を生成した。学習データ、検証データ、及びテストデータの用意や DNN の最適化の条件等については、4.2.1 項と同様である。

本実験では、検証データに対する正答率に加え、SDR の改善量を用いて提案手法のパーミュテーション問題の解決性能を評価する。SDR 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。今、音源分離の目的音源信号を $s(l)$ 、目的音以外の音源（干渉音源）信号を $n(l)$ とすると、これらが混合した信号 $x(l)$ は次式となる。

$$x(l) = s(l) + n(l) \quad (4.8)$$

このとき、混合信号 $x(l)$ に音源分離を適用し得られる目的音源の推定信号 $\hat{s}(l)$ は次式で表される。

$$\hat{s}(l) = s_{\text{target}}(l) + e_{\text{interf}}(l) + e_{\text{artif}}(l) \quad (4.9)$$

ここで, $s_{target}(l)$, $e_{interf}(l)$, 及び $e_{artif}(l)$ はそれぞれ推定信号中の目的音源成分, 残留した干渉音源成分, 及び音源分離処理によって生じた歪み成分である. このとき, SDR は次のように算出できる.

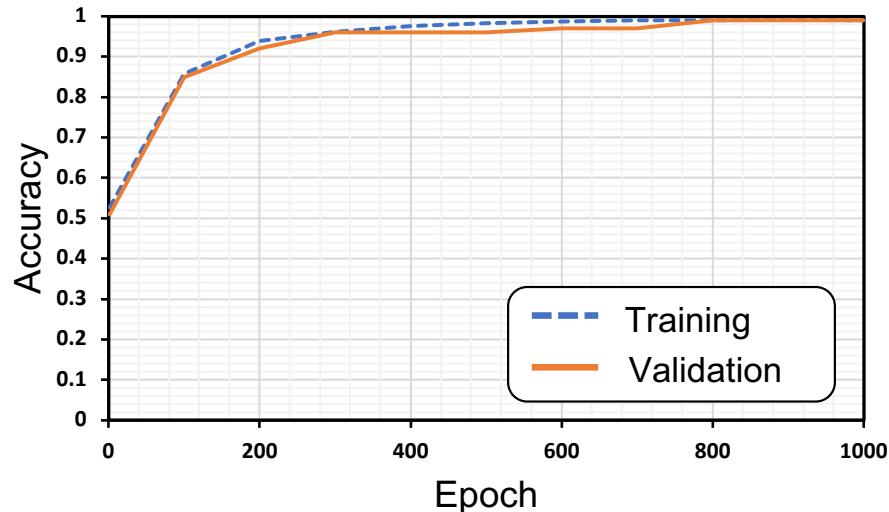
$$\text{SDR} = 10 \log_{10} \sum_{l=1}^L \frac{|\mathbf{s}_{\text{target}}(l)|^2}{|\mathbf{e}_{\text{interf}}(l) + \mathbf{e}_{\text{artif}}(l)|^2} \quad [\text{dB}] \quad (4.10)$$

4.3 実験結果

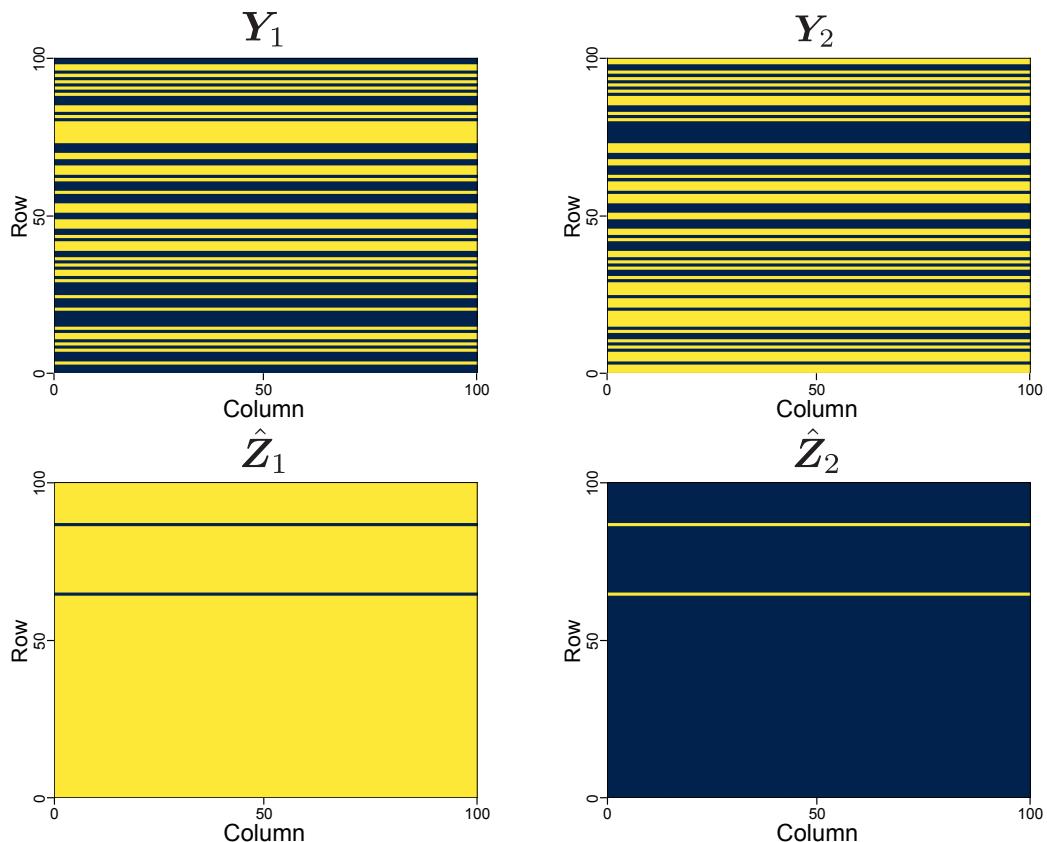
4.3.1 人工データに対する実験結果

Figs. 4.7–4.9 にそれぞれ、Figs. 4.1–4.3 の人工データ行列に対して 1 行毎にランダムに入れ替える場合 ($\gamma = 1$) の実験結果を示している。この結果では、学習時の学習データ及び検証データに対する正答率（各図における (a)）と検証データの入力及び予測結果（各図における (b)）をそれぞれ示している。この結果を見ると、Figs. 4.1 及び 4.2 に示すような単純な構造を持つ行列に対しては、Figs. 4.7 (a) 及び 4.8 (a) に示すようにいずれも検証データに対する正答率が 100% に近い値となっている。予測結果である Figs. 4.7 (b) 及び 4.8 (b) の推定分離信号 (\hat{Z}_1, \hat{Z}_2) は少しの間違いを含んでいるものの、高精度で正しい並び替えができることが分かる。しかしながら、Fig. 4.3 に示す 1 列毎に 0 と 1 の値が入れ替わる行列に対しては、Fig. 4.9 (a) に示すように検証データに対する正答率が 54% 程度となった。予測結果である Fig. 4.9 (b) の推定分離信号 (\hat{Z}_1, \hat{Z}_2) を見ても正しい並び替えができていないことが分かる。

次に、Figs. 4.10–4.12 にそれぞれ Figs. 4.1–4.3 の人工データ行列に対して 2 行毎にランダムに入れ替える場合 ($\gamma = 2$) の実験結果を示している。Figs. 4.7–4.9 と同様に、学習時の学習データ及び検証データに対する正答率（各図における (a)）と検証データの入力及び予測結果（各図における (b)）をそれぞれ示している。これらの結果では、どの実験結果においても検証データに対する正答率が 90% を超えており、推定分離信号 (\hat{Z}_1, \hat{Z}_2) も高精度で正しい並び替えが達成できていることが確認できる。さらにブロックパーミュテーション問題におけるブロックサイズを大きくした $\gamma = 4$ 及び $\gamma = 8$ の結果についても付録 B に掲載している。いずれも検証データに対して高い正答率を達成している。これらの結果から分かるとして、提案手法の深層パーミュテーション解決法はブロックサイズが $\gamma = 2$ 以上のブロックパーミュテーション問題であれば、複雑な構造を持つ音源信号に対しても高精度に解決することができる可能性が高い。一方で、周波数ビン毎に推定信号の順序がランダムに入れ替える一般的なパーミュテーション問題について解決は精度の低下が見られるため、FDICA の推定信号に対してそのまま提案手法を適用することは効果的ではない可能性がある。

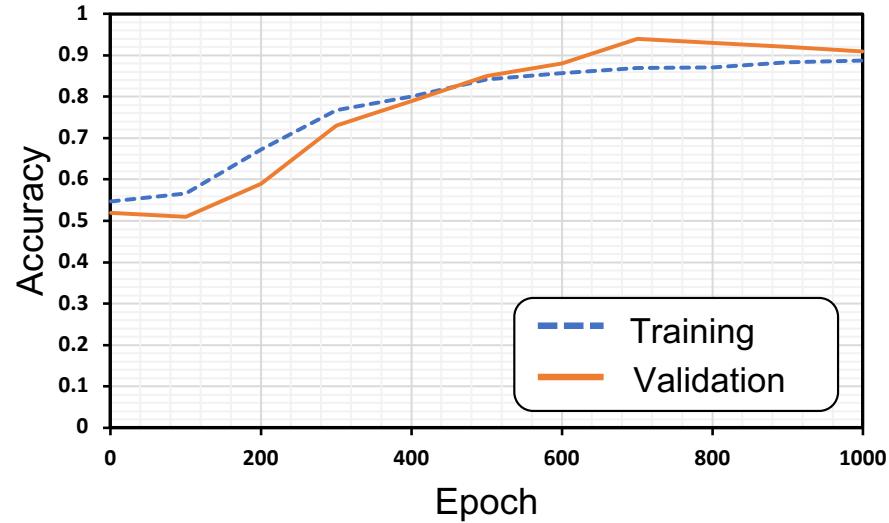


(a) Accuracy for training and validation data.

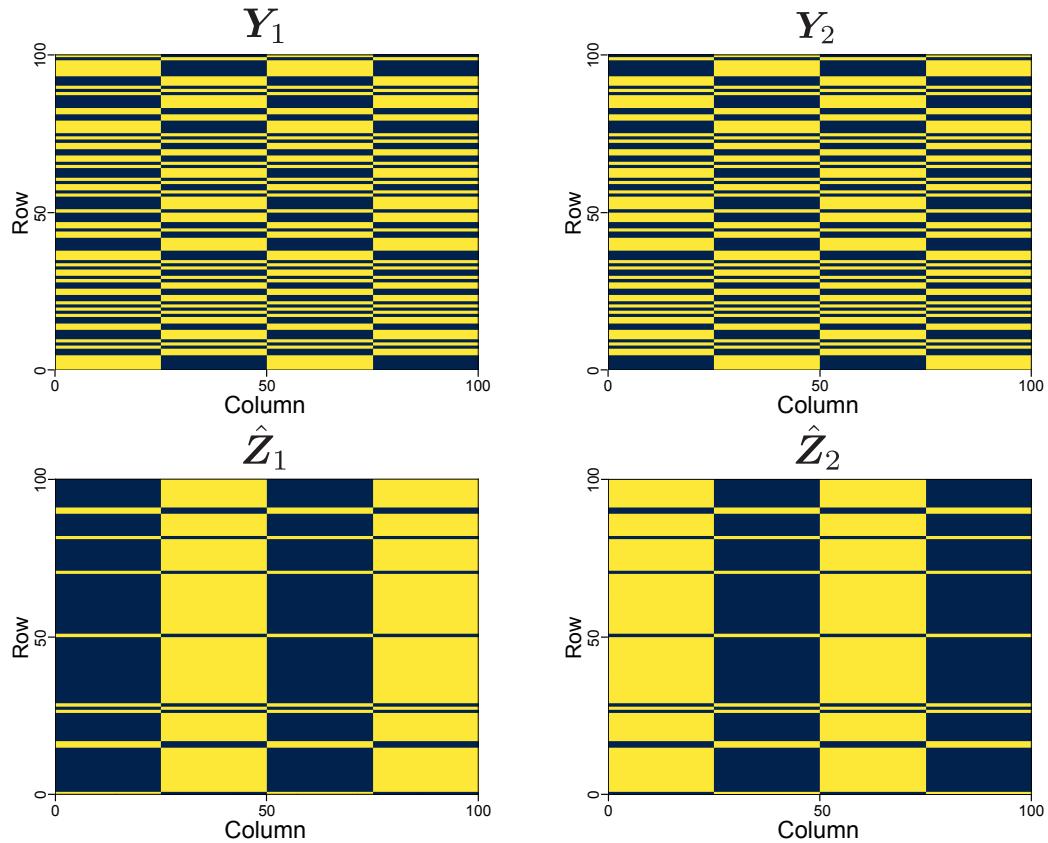


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. 4.7: Experimental results with $\gamma = 1$ using artificial source matrices of Fig. 4.1.

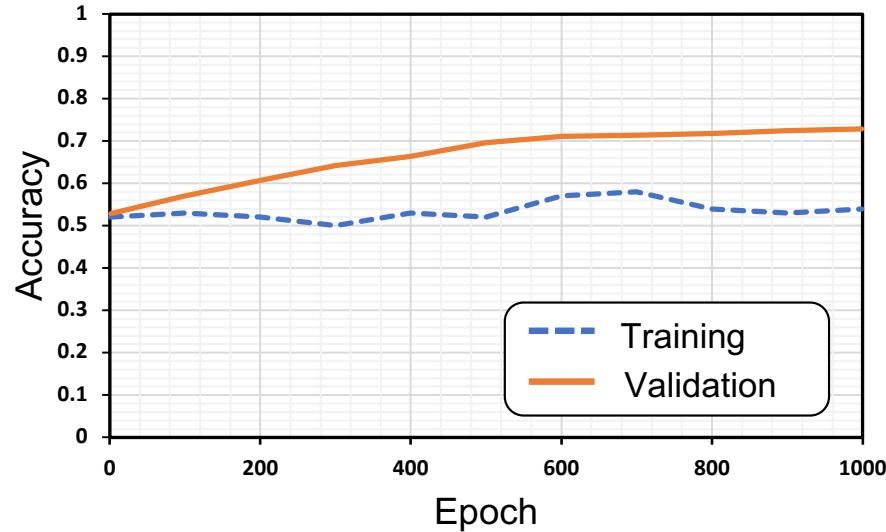


(a) Accuracy for training and validation data.

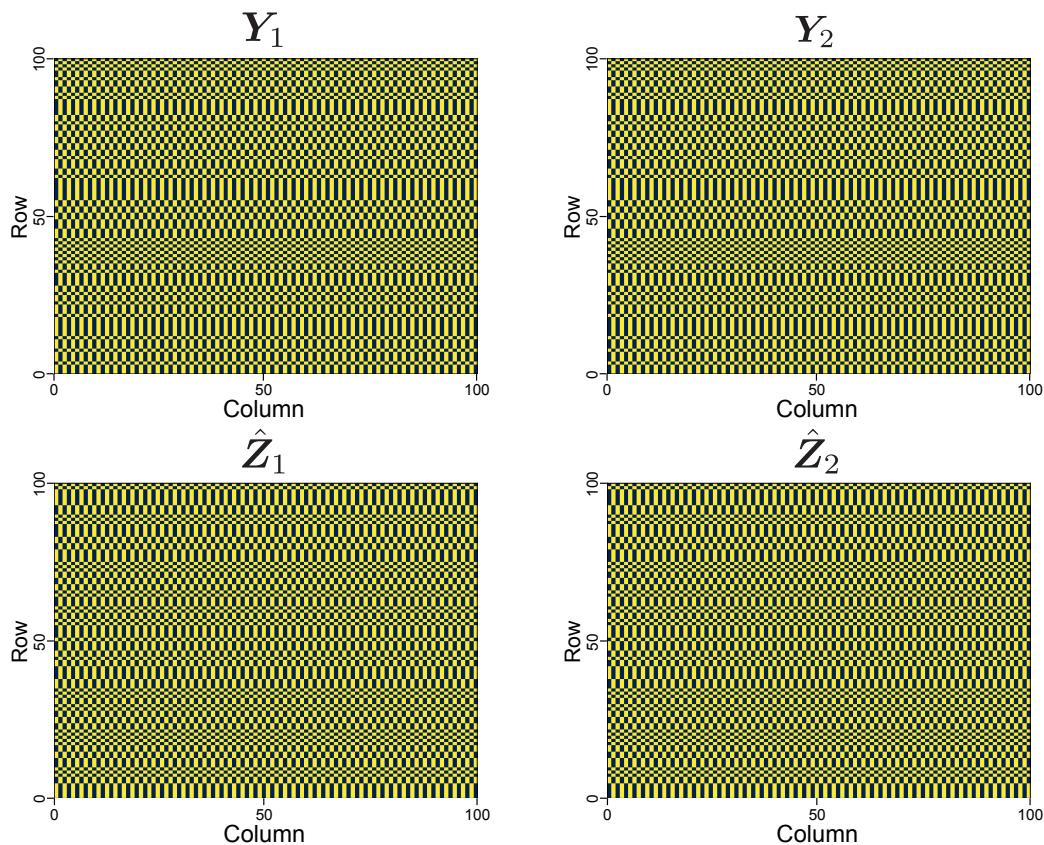


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. 4.8: Experimental results with $\gamma = 1$ using artificial source matrices of Fig. 4.2.

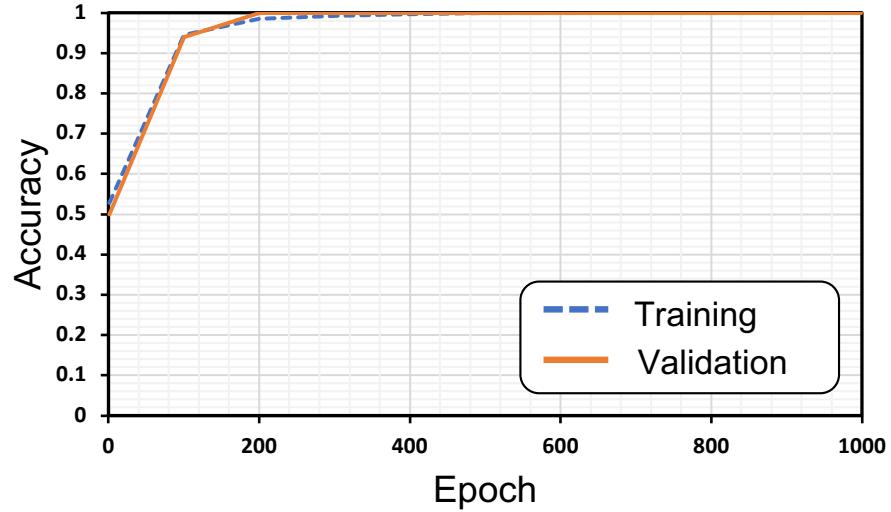


(a) Accuracy for training and validation data.

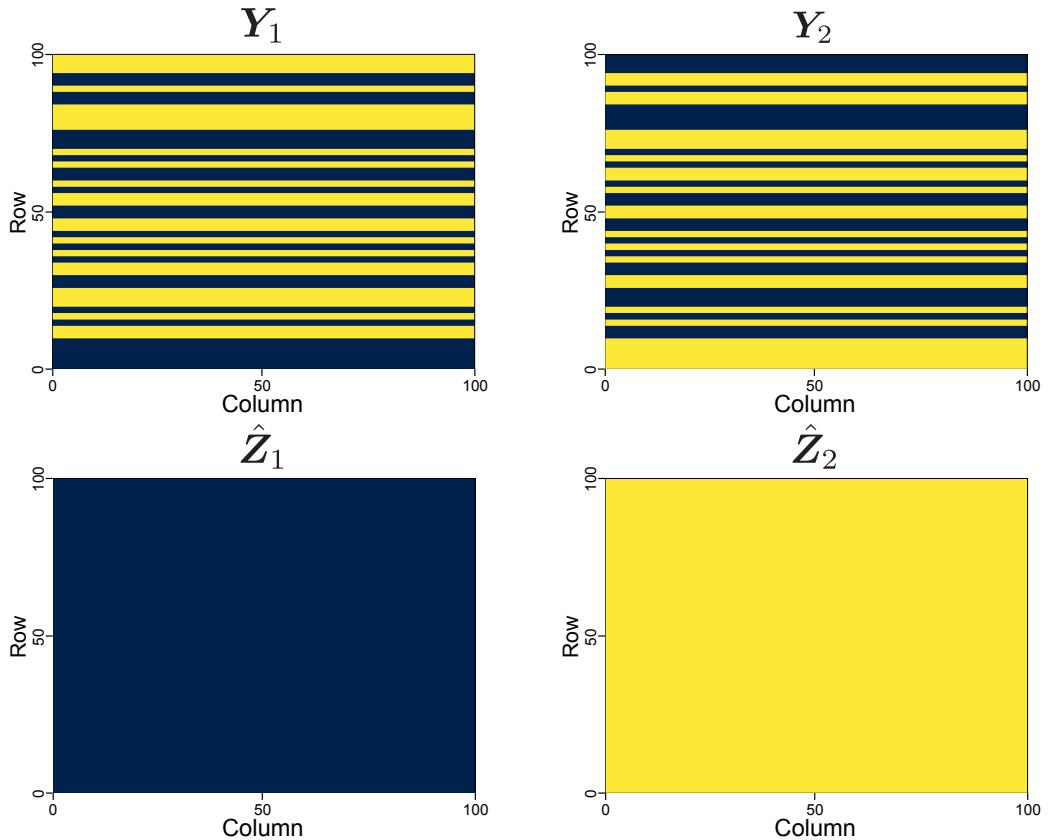


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. 4.9: Experimental results with $\gamma = 1$ using artificial source matrices of Fig. 4.3.

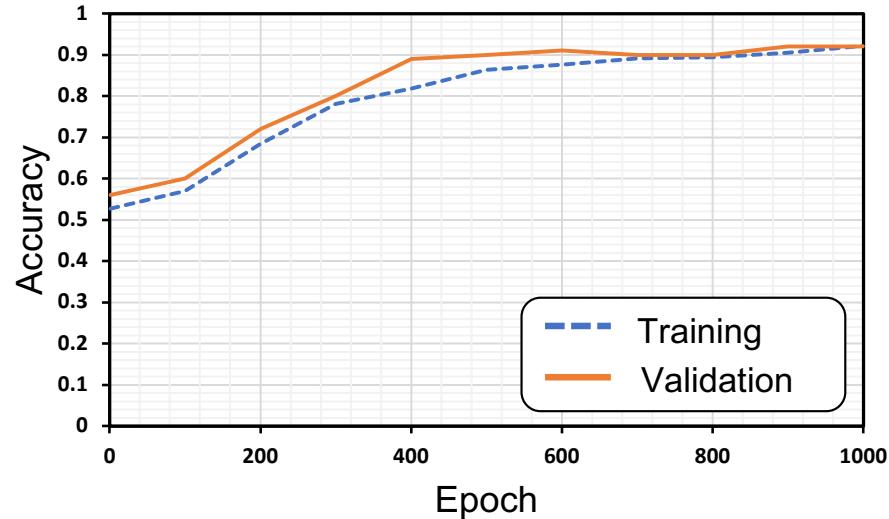


(a) Accuracy for training and validation data.

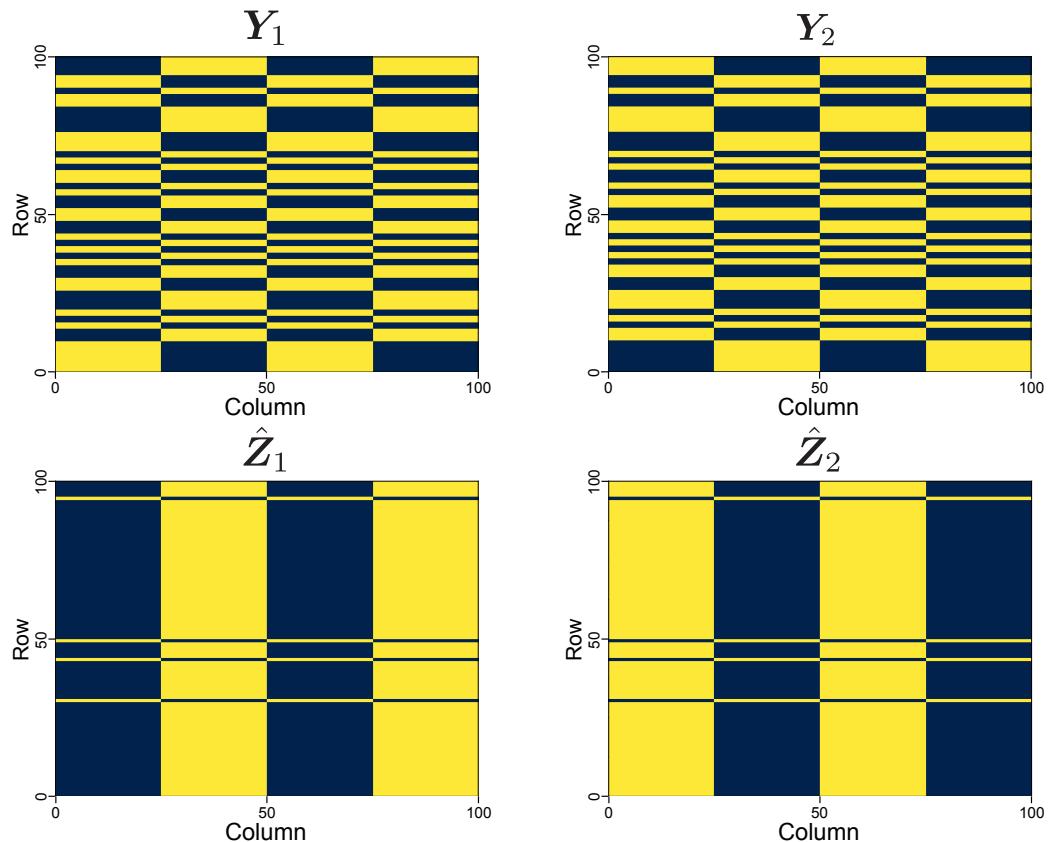


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. 4.10: Experimental results with $\gamma = 2$ using artificial source matrices of Fig. 4.1.

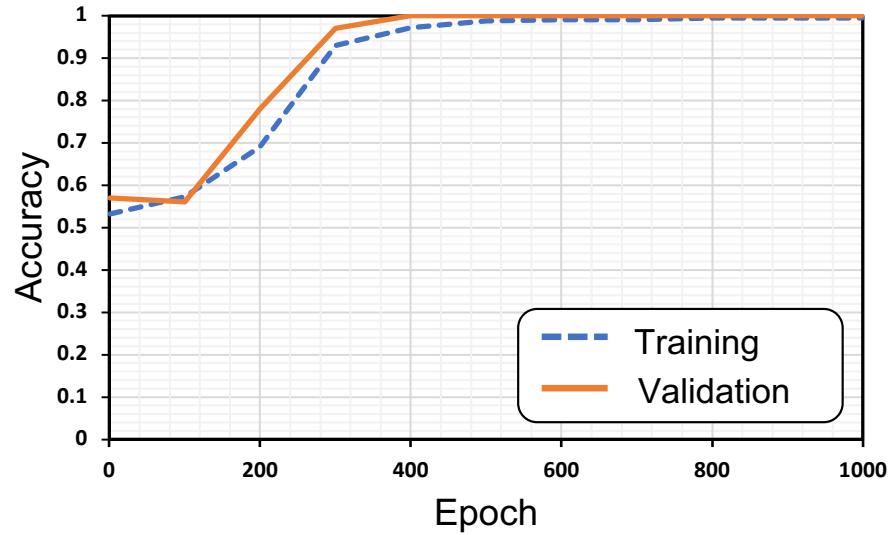


(a) Accuracy for training and validation data.

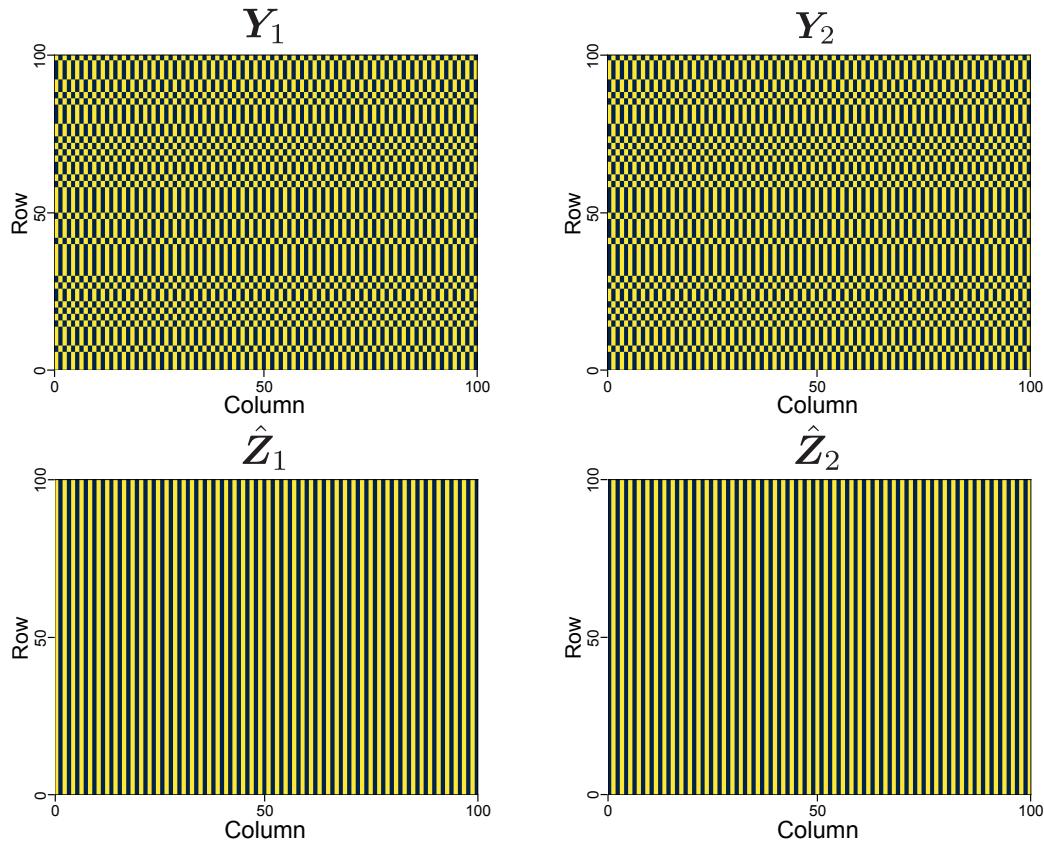


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. 4.11: Experimental results with $\gamma = 2$ using artificial source matrices of Fig. 4.2.



(a) Accuracy for training and validation data.

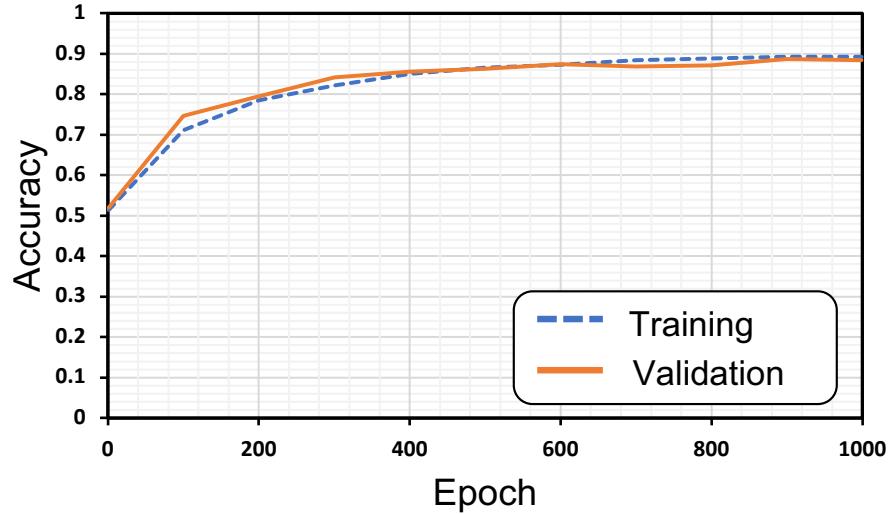


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

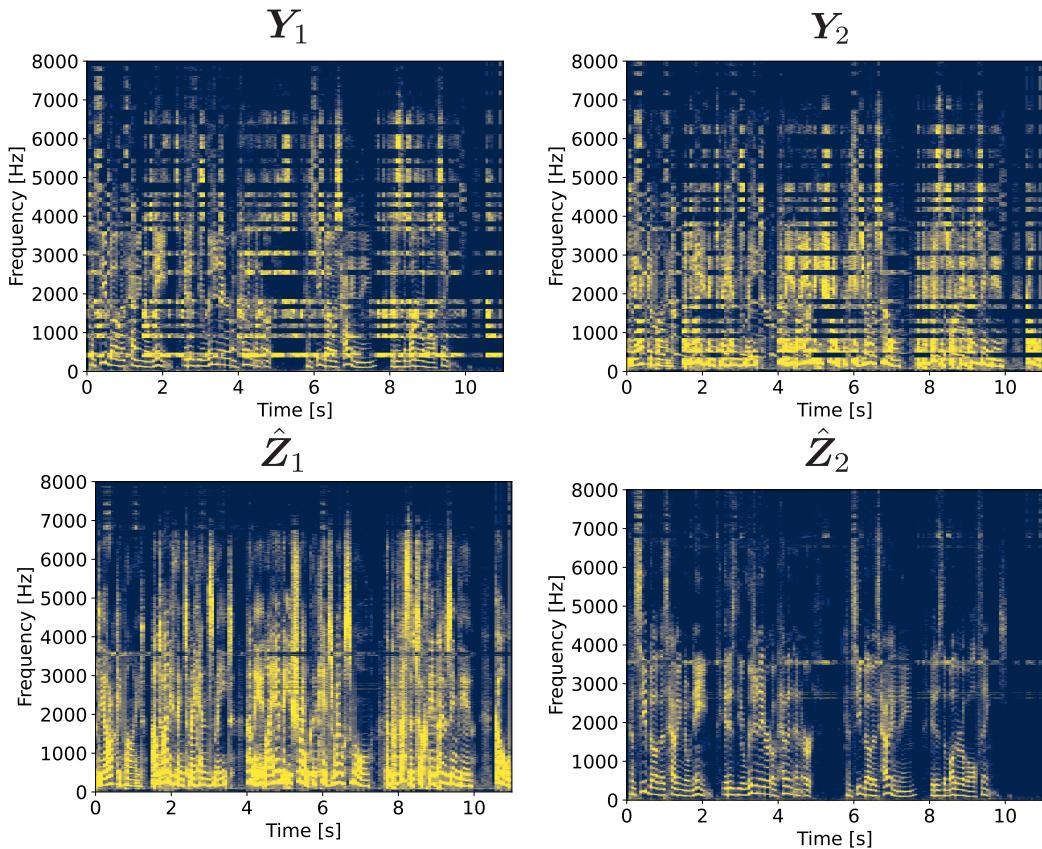
Fig. 4.12: Experimental results with $\gamma = 2$ using artificial source matrices of Fig. 4.3.

4.3.2 音声及び音楽信号に対する実験結果

Figs. 4.13 及び 4.14 にはそれぞれ、Figs. 4.5 及び 4.6 の音声及び音楽信号に対して 16 行毎にランダムに入れ替える場合 ($\gamma = 16$) の実験結果を示している。前項の結果と同様に、学習時の学習データ及び検証データに対する正答率（各図における (a)）と検証データの入力及び予測結果（各図における (b)）をそれぞれ示している。音声及び音楽信号のどちらに対しても、検証データに対する正答率が 90% を超えており、推定分離信号 (\hat{Z}_1, \hat{Z}_2) も概ね正確な並び替えができていることが分かる。SDR の改善量は、Fig. 4.13 の \hat{Z}_1 が 26.7 dB, \hat{Z}_2 が 31.0 dB であった。また、Fig. 4.14 の \hat{Z}_1 が 22.6 dB, \hat{Z}_2 が 27.6 dB であった。Figs. 4.13 及び 4.14 を比較すると、音楽信号の実験結果の方が、音声信号の実験結果よりも検証データに対する正答率が高いことが分かる。これは、本実験で使用した 2 種類の楽器音（ドラムとピアノ）が、男女の音声信号に比べて明確に異なる時間周波数構造を持っているためと推測できる。ドラムの音は、Fig. 4.6 の Z_1 のスペクトログラムに示すように、全周波数成分に対して大きなパワーの成分を持っている。一方で、ピアノの音は Fig. 4.6 の Z_2 のスペクトログラムに示すように、基本周波数とその整数倍という調波構造を持っており、ドラムとは大きく異なる時間周波数構造となっていることが分かる。このような時間周波数構造の違いにより、提案手法の DNN は音声信号よりも高精度に推定分離信号を予測することができたと考えられる。以上の実験より、ある程度のサイズを持つ周波数帯域で生じるブロックパーティション問題に対しては、実際の音声及び音楽信号でも、提案する深層パーティション解決法が高精度に正しい分離信号成分の並び替えを予測できることが確認された。

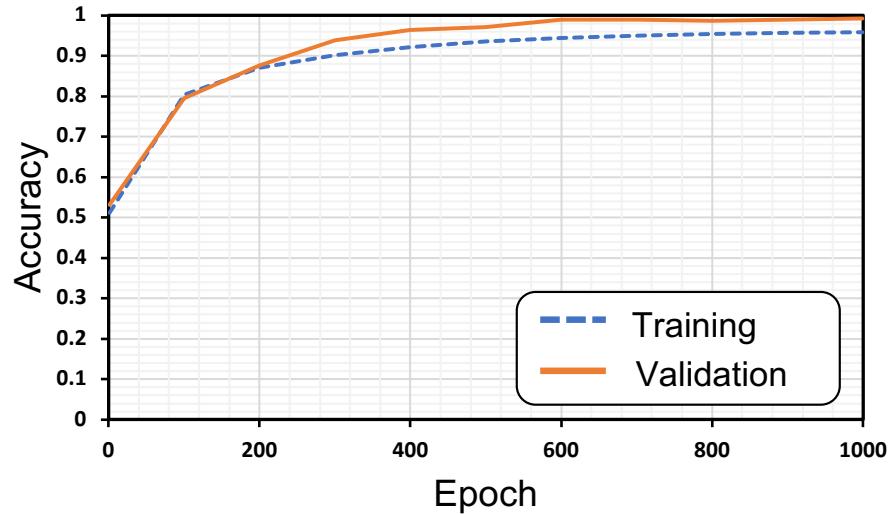


(a) Accuracy for training and validation data.

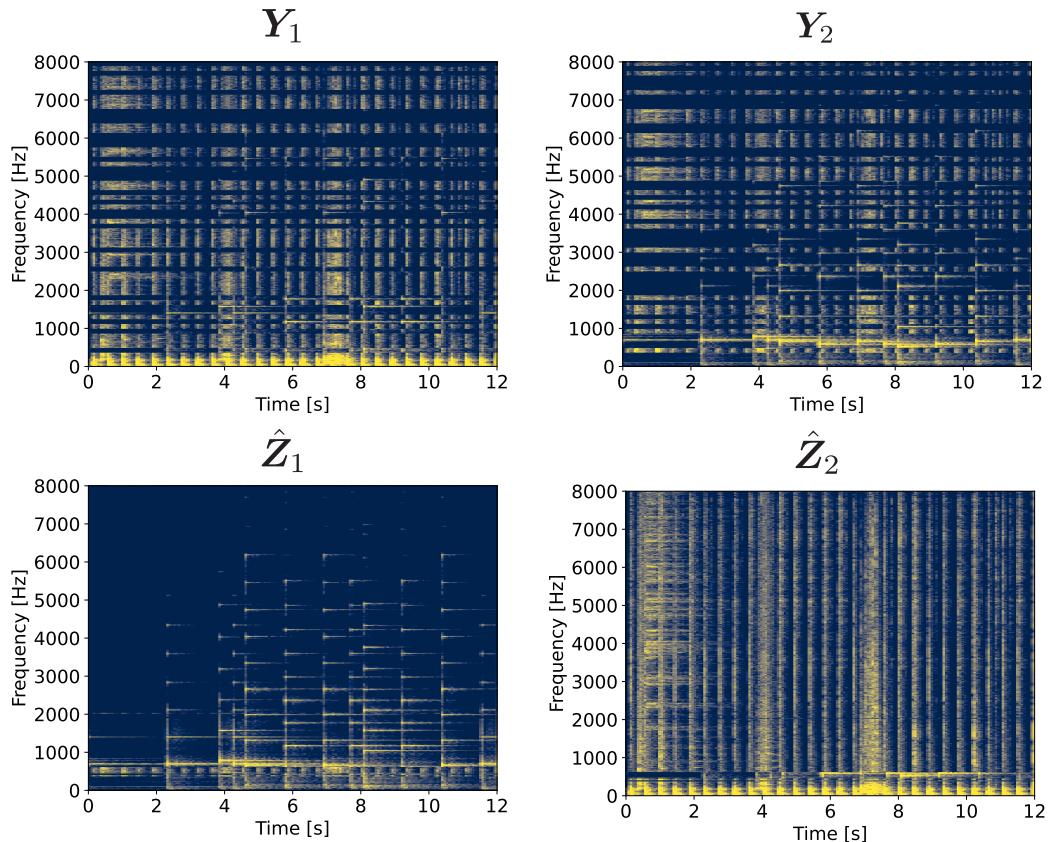


(b) Input spectrograms with permutation problem (upper) and permutation-aligned spectrograms using predicted results (bottom).

Fig. 4.13: Experimental results with $\gamma = 16$ using speech source spectrograms of Fig. 4.5.



(a) Accuracy for training and validation data.



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

Fig. 4.14: Experimental results with $\gamma = 16$ using musical instrument source spectrograms of Fig. 4.6.

4.4 本章のまとめ

本章では、提案手法の有効性を確認するため、人工的に作成したデータと実際の音声及び音楽信号を用意し実験を行った。実験の結果より、人工データを用いたブロック単位でのパーミュテーション問題に対しては、どのような行列であっても 100% に近い確率で解決できることを示した。実際の音声及び音楽信号に対しても、ブロック単位でランダムに入れ替えが行われている場合は 90% を超える正答率になることを示した。SDR の改善量は、音声信号に対して 28 dB 程度、音楽信号に対しては 25 dB 程度であった。次章では、本論文における総括とした結論を述べる。

第 5 章

結言

本論文では、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。DNN の入力には、正規化した分離信号から局所時間振幅スペクトログラム成分を抽出した値を用いた。DNN の出力には、softmax 関数を使用し確率値を出力する。この確率値は、各音源の成分である確率を意味する。DNN の出力である確率値を用いて、推定パーミュテーション行列を作成し分離信号の並び替えを行った。損失関数には MSE を用い、推定分離信号と完全分離信号のスペクトログラム間で損失を求め誤差逆伝搬を行った。テストデータに対しては DNN の入力となる局所時間振幅スペクトログラムをストライド幅に従ってずらしていくことで、時間方向に対して多数決処理を行った。実験結果より、ブロック単位でのパーミュテーション問題に対しては提案手法を用いて正しく並び替えができる음을示した。

最後に今後の展望を述べる。本論文では、深層パーミュテーション解決手法の可能性に注目しており、基礎的な実験を行ってきた。本実験では、学習データと検証データの音源に同じスペクトログラムを用いて実験を行っており、未だに音源の時間周波数構造に対する汎化性能は獲得できていない。この問題を解決するためには、学習データに多数の音声及び音楽信号を用意し、大量のデータを DNN に学習させる必要がある。また、更なる精度向上のために、DNN の構造として MLP を用いるのではなく、双向再帰型 DNN を使用することを検討している。さらには、2 音源の実験の拡張版として 3 音源以上に対する実験も行う必要がある。

謝辞

本論文は、香川高等専門学校電気情報工学科北村研究室にて行われた研究に基づくものです。まず、本研究を進めるにあたり、ご多忙のところ熱心にご指導くださいました指導教員の北村大地講師に心より感謝申し上げます。北村大地講師には、論文執筆や研究に関する議論など、細部にわたるまで丁寧にご指導いただきました。DNNの研究で用いるサーバの増設等にも取り組んでいただき、日々の研究を効率良く行うことができました。心よりありがとうございます。本論の副査である雛元洋一助教には、論文の構成や記述に関して有益な助言を頂き、大変お世話になりました。ここに厚く御礼申し上げます。北村研究室の先輩である専攻科2年の岩瀬佑太氏、大藪宗一郎氏、梶谷奈未氏、渡辺瑠伊氏には、音源分離に関する基礎概念のご説明をはじめ、研究の進め方に関して数々のご支援をいただきました。特に、北村研究室の先輩である専攻科2年の渡辺瑠伊氏には、DNNに関するアドバイスやサーバ管理に関する知見をはじめ、数々のご支援とご助言をいただきました。心より感謝申し上げます。また、北村研究室同期の川口翔也氏、細谷泰稚氏、村田佳斗氏、溝渕悠朔氏には、日頃のディスカッションのほか、1年に亘る研究室生活を様々な面で支えていただきました。ここに感謝申し上げます。

最後になりますが、現在に至るまで私の学生生活を金銭的に支え、暖かく見守って下さった両親には感謝の念に堪えません。これまで本当にありがとうございました。

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 Trans. Signal and Information Processing*, vol. 8, no. e12, pp. 1–14, 2019.
- [2] P. Comon, “Independent component analysis, a new concept?,” *Signal Processing*, 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. 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.
- [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 Trans. Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 666–678, 2006.
- [7] 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. Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
- [8] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.
- [9] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 189–192, 2011.
- [10] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

- [11] 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, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.
- [12] 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.
- [13] D. Kitamura, N. Ono, and H. Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," *Proc. European Signal Processing Conference*, pp. 1210–1214, 2017.
- [14] S. Yamaji and D. Kitamura, "DNN-based permutation solver for frequency-domain independent component analysis in two-source mixture case," *Proc. APSIPA Annual Summit and Conference*, pp. 781–787, 2020.
- [15] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proc. International Conference on Independent Component Analysis and Blind Signal Separation*, pp. 722–727, 2001.
- [16] Y. Liang, S.M. Naqvi, and J. Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electron. Lett*, pp.460–462, 2012.
- [17] 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.
- [18] T. Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Umbach, "Graph-PIT: Generalized permutation invariant training for continuous separation of arbitrary numbers of speakers," *INTERSPEECH*, pp. 3490–3494, 2021.
- [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. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 241-245, 2017.
- [20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proc. International Conference on Machine Learning*, 2010.
- [22] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011): -Audio source separation," *Proc. International Conference on Latent Variable Analysis and Signal Sep-*

- aration*, pp. 414–422, 2012.
- [23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv*, pp. 1412–6980, 2014.

付録 A

Birkhoff–von Neumann の定理

サイズ N の正方行列 \mathbf{D} が二重確率行列であるとき, \mathbf{D} はサイズ N の全てのパーミュテーション行列 $\{\mathbf{P}_i\}_{i=1}^{N!}$ の凸結合で表せる. 即ち, 凸結合の係数 $\sigma_i \geq 0$ を用いて次式が成立する.

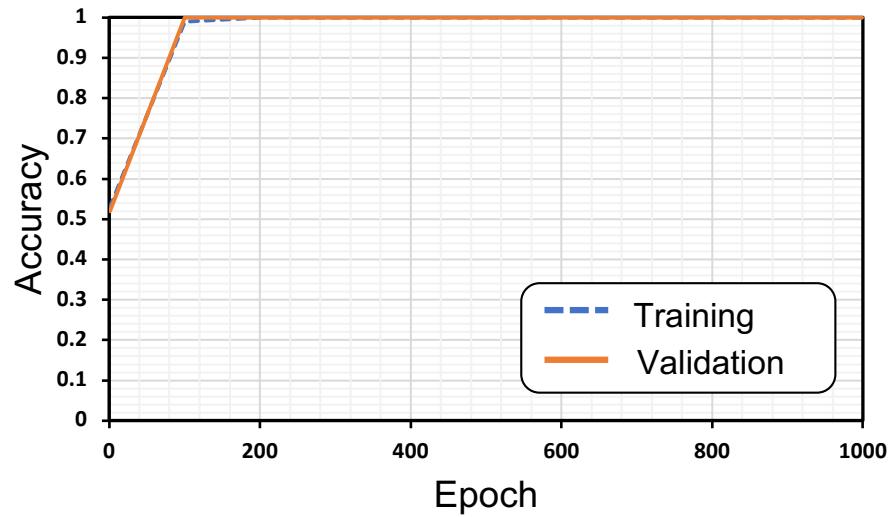
$$\mathbf{D} = \sum_{i=1}^{N!} \sigma_i \mathbf{P}_i \quad (\text{A.1})$$

但し, σ_i は凸結合係数であるため, $\sum_{i=1}^{N!} \sigma_i = 1$ を満たす.

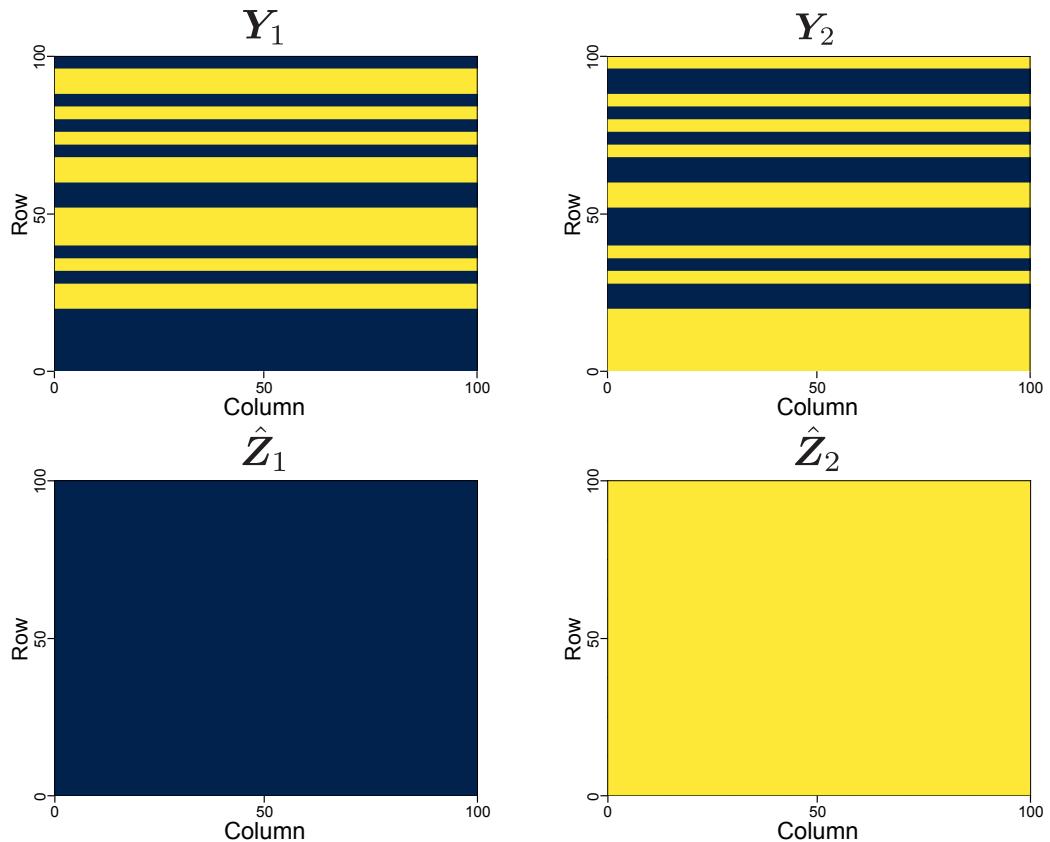
付録 B

人工データに対する予測結果

4章で掲載した人工データに対する実験結果の他にも、4行、8行毎に各周波数成分をシャッフルした場合の実験も行った。以下に実験結果を掲載する。

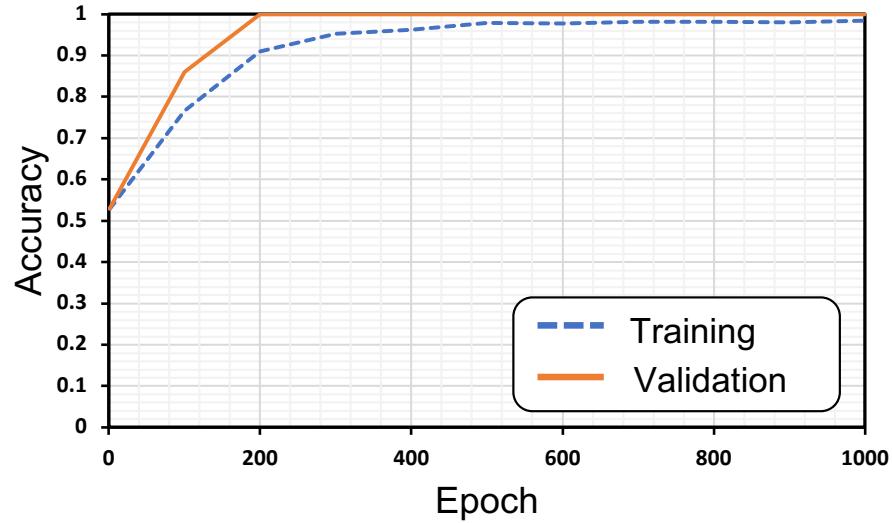


(a) Accuracy for training and validation data.

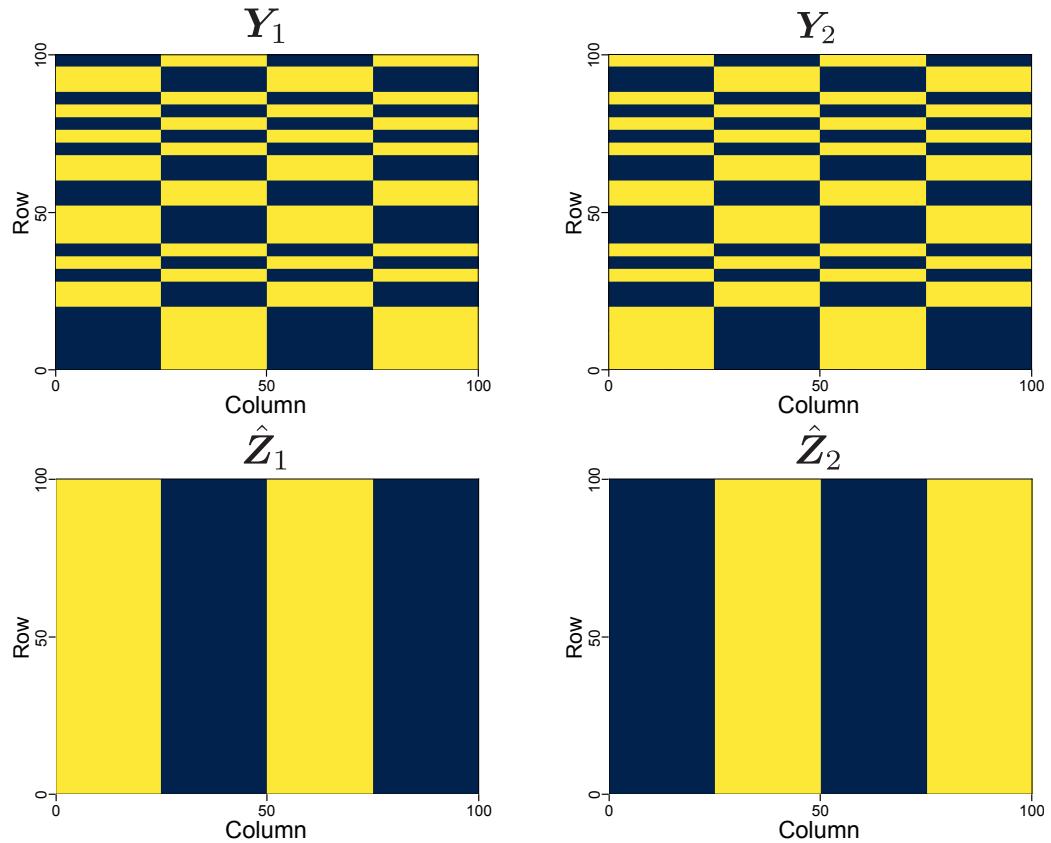


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. B.1: Experimental results with $\gamma = 4$ using artificial source matrices of Fig. 4.1.

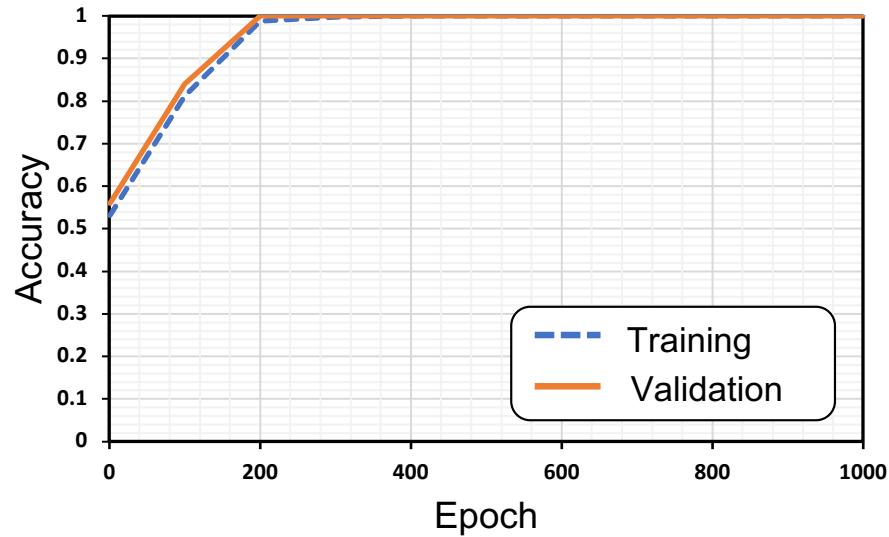


(a) Accuracy for training and validation data.

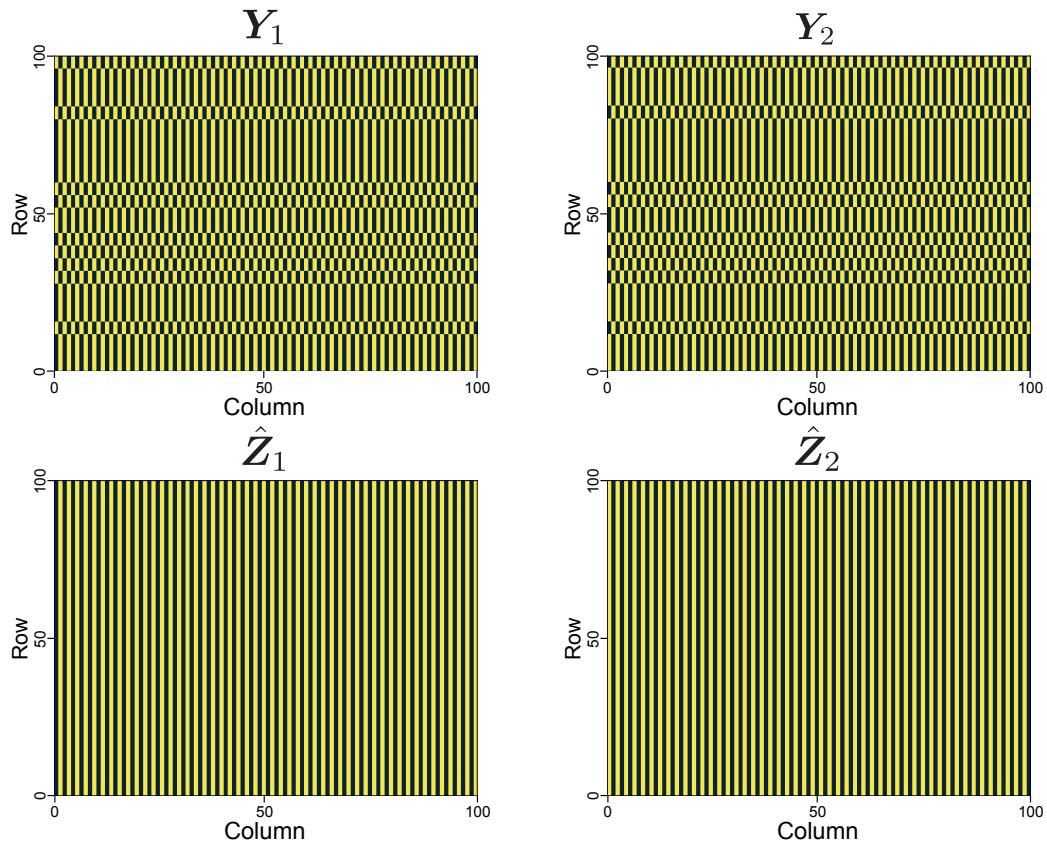


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. B.2: Experimental results with $\gamma = 4$ using artificial source matrices of Fig. 4.2.

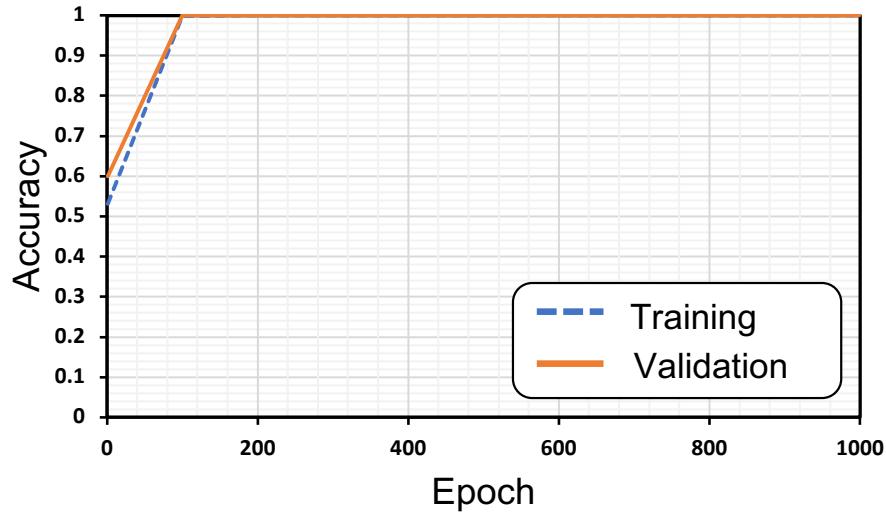


(a) Accuracy for training and validation data.

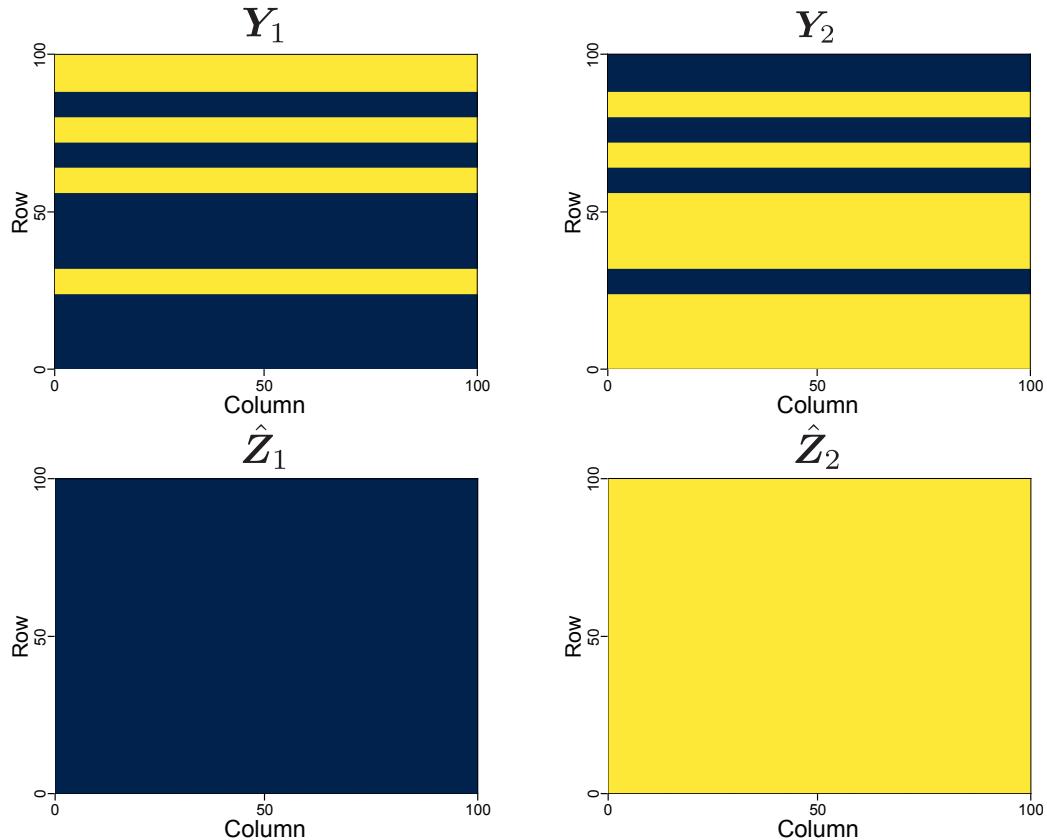


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. B.3: Experimental results with $\gamma = 4$ using artificial source matrices of Fig. 4.3.

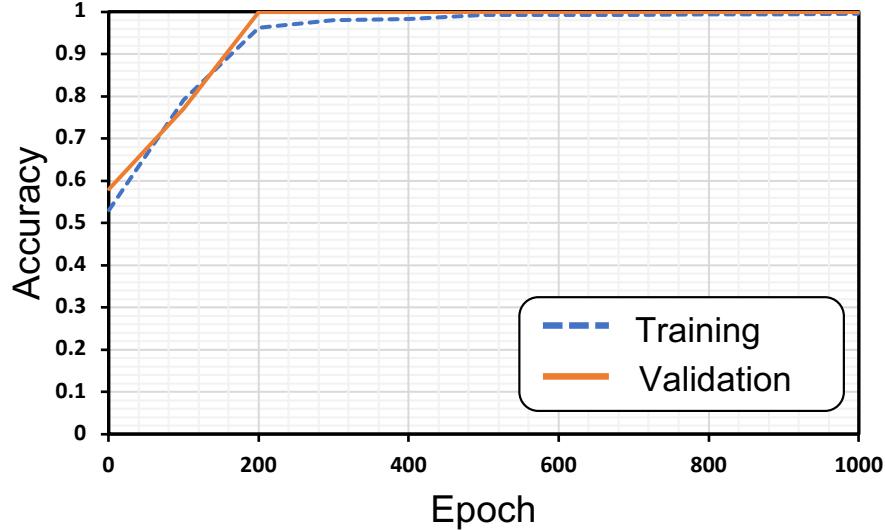


(a) Accuracy for training and validation data.

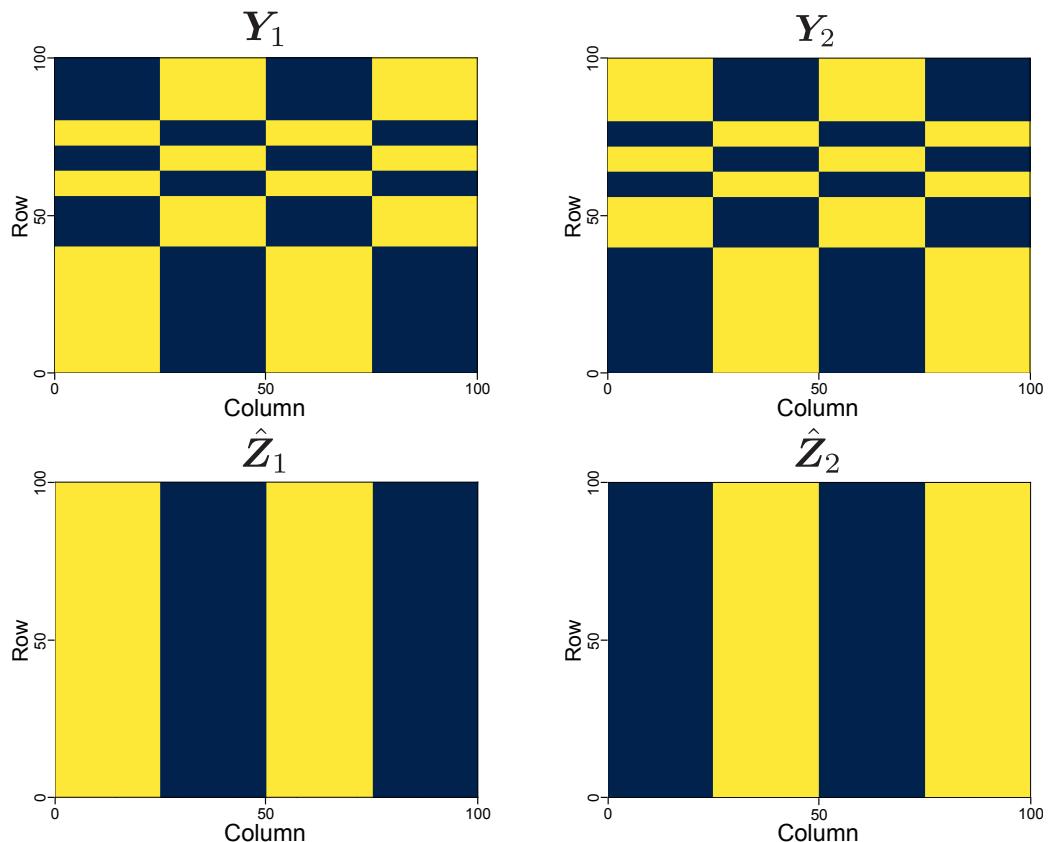


(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. B.4: Experimental results with $\gamma = 8$ using artificial source matrices of Fig. 4.1.



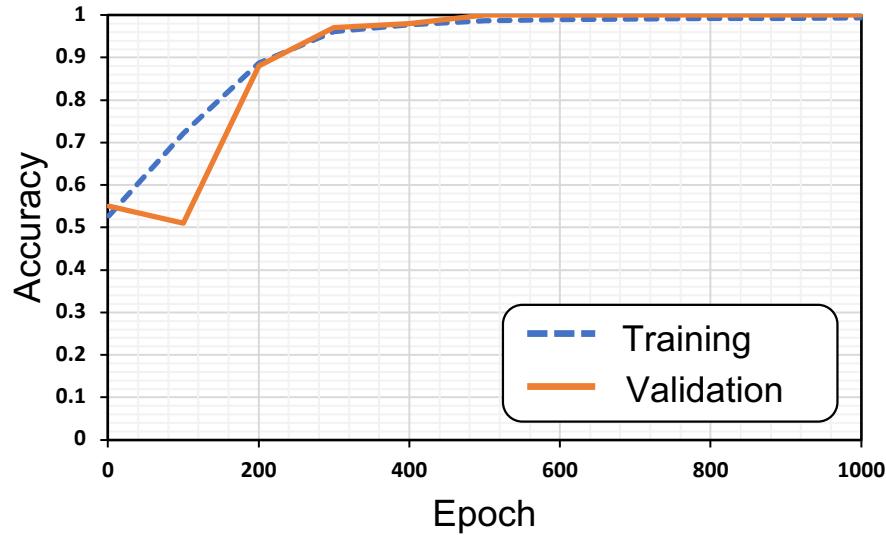
(a) Accuracy for training and validation data.



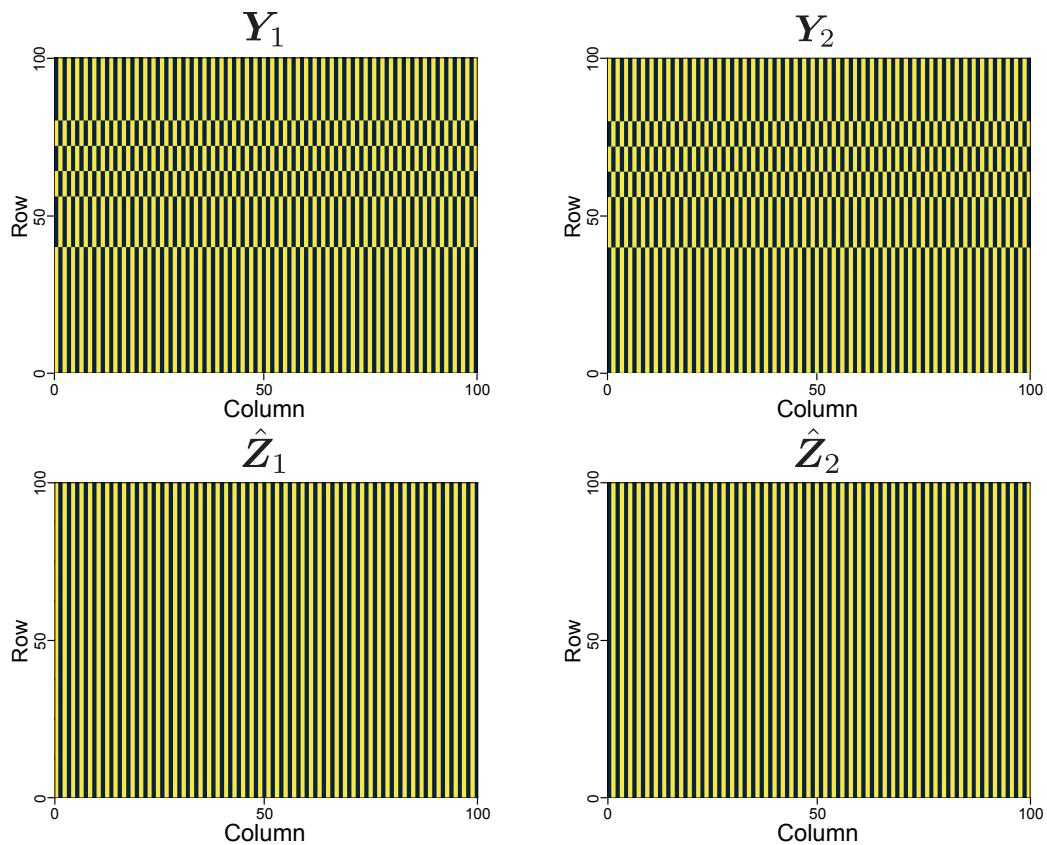
(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. B.5: Experimental results with $\gamma = 8$ using artificial source matrices of Fig. 4.2.

56 付録 B 人工データに対する予測結果



(a) Accuracy for training and validation data.



(b) Input matrices with permutation problem (upper) and permutation-aligned matrices using predicted results (bottom).

Fig. B.6: Experimental results with $\gamma = 8$ using artificial source matrices of Fig. 4.3.