



# 卒業研究論文

## 論文題目

深層学習に基づく多チャネル音源分離のための  
パーティション解決の基礎的実験

提出年月日	令和 X 年 X 月 X 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	雫元 洋一 助教	印
学 科 長	辻 正敏 教授	印

香川高等専門学校

# **Basic experiments on permutation resolution for multi-channel source separation based on deep learning.**

Fumiya Hasuike

Department of Electrical and Computer Engineering  
National Institute of Technology, Kagawa College

## **Abstract**

In this thesis, we deal with audio source separation, which is a technique to separate audio sources from an observed signal. This technology can be used to separate speech from multiple people speaking at the same time, to separate speech from background noise, or to separate musical instruments in a music signal. One of the popular source separation methods is frequency-domain independent component analysis (FDICA). In FDICA, the separation is performed with applying independent component analysis to each frequency. However, the order of the estimated signal in each frequency is not aligned among all frequencies, resulting in the so-called permutation problem. In recent years, deep neural networks (DNN) have been proposed to solve the permutation problem, but the problem is that source separation of more than three sources is not realistic from the viewpoint of computational complexity and Algorithmic Complexity. In this paper, we propose a new algorithm for the DNN-based permutation problem and the block-wise permutation problem. The proposed DNN learns the characteristics of the complex time-frequency structure of the separated signals in advance and predicts whether the permutation mismatch occurs or not for all frequency bins. The performance of the proposed method in solving the permutation problem is evaluated by the percentage of correct answers in all frequency bins. The experimental results show that the proposed DNN permutation solution has a correct answer rate close to 100% for artificially created pseudo block-wise permutation problems. In addition, a block-by-block permutation problem on actual audio data showed a correct answer rate of over 80%.

**Keywords:** **FDICA**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、複数人が同時に発話した内容をそれぞれの音声に分けたい場合や、背景雑音と音声を分離したいとき、さらには音楽信号における楽器音ごとの分離などに利用される。代表的な音源分離手法の1つとして時間周波数領域独立成分分析（frequency-domain independent component analysis: FDICA）がある。これは、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分離信号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: DNNsDNN）を用いたパーミュテーション問題の解決法が提案されてきたが、3音源以上の音源分離は計算量の観点及びアルゴリズムの複雑性から現実的ではないことが課題として挙げられる。本論文では、パーミュテーション問題、またブロック単位でのパーミュテーション問題に対して、**新たなアルゴリズムを提案するDNNに基づく新しいアプローチを提案し、パーミュテーション問題に対するDNNに基づく解法の妥当性について実験的に調査する**。提案手法のDNNは、分離信号の複雑な時間周波数構造の特徴を事前に学習し、全周波数ビンについてパーミュテーション不整合が生じているか否かを予測する。提案手法のパーミュテーション問題の解決性能は、全周波数の正答率で評価する。実験結果から、提案するDNNパーミュテーション解決法は人工的に作成した擬似的なブロック単位での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、実際の音声データに対してもブロック単位でのパーミュテーション問題として実験を行うと、80%を超える正答率を示した。

# 目次

<b>第 1 章</b>	<b>序論</b>	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
<b>第 2 章</b>	<b>基礎理論と従来手法</b>	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	6
2.2.2	統計的独立性	7
2.2.3	ICA における任意性	8
2.3	STFT	9
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	15
2.9	本章のまとめ	16
<b>第 3 章</b>	<b>提案手法</b>	17
3.1	まえがき	17
3.2	動機	17
3.3	DNN の入出力	20
3.4	DNN の構造	25
3.5	DNN 学習時の損失関数	25
3.6	学習済の DNN のテストデータへの適用	27
3.7	本章のまとめ	29
<b>第 4 章</b>	<b>実験</b>	30
4.1	まえがき	30
4.2	実験条件	30

4.2.1	人工データを用いた実験の条件	30
4.2.2	実際の音響信号を用いた実験の条件	33
4.3	人工データに対する実験結果	33
4.4	本章のまとめ	40
第 5 章 結言		41
謝辞		42
参考文献		42

# 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**近年ニーズが高まっており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されてきた。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、**混合系  $\mathbf{A}$  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**元の**音源数以上となる**優決定条件下での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、優決定条件下の**信号源分離問題BSS** に広く適用されている代表的な**音源分離手法**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。

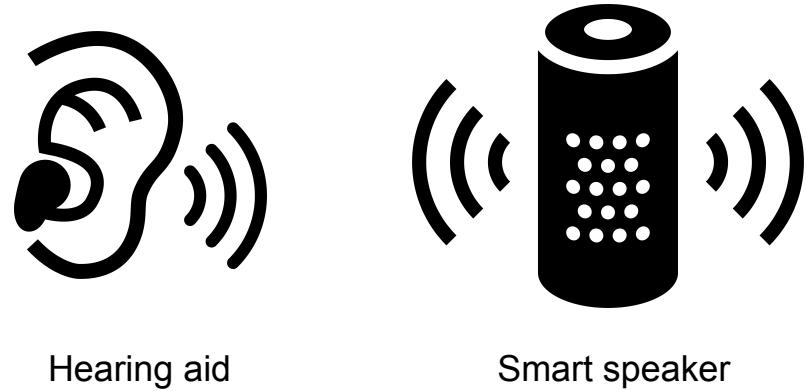


Fig. 1.1: Examples of application using speech source separation.



Fig. 1.2: Overview of BSS.

そこで、観測信号を時間周波数領域に変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は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『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] [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。

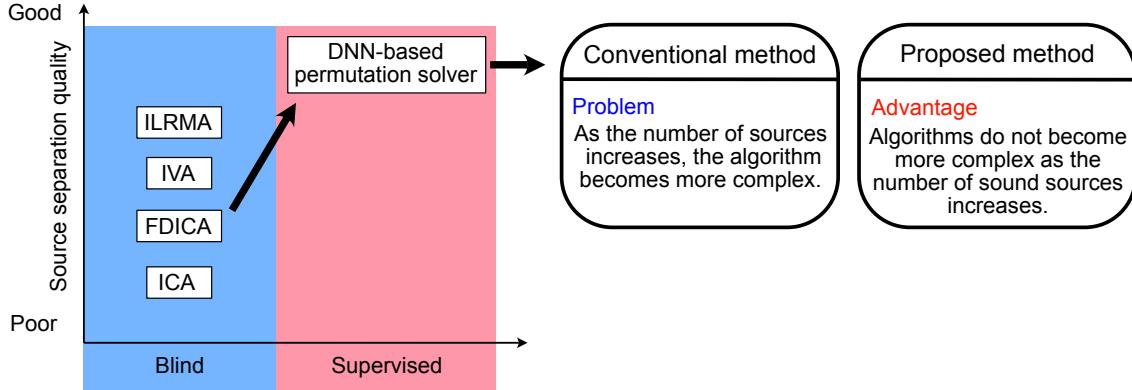


Fig. 1.3: Scope of this thesis.

## 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 音源以上にも対応できる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について言及する提案し、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しても言及する有効性についても調査する。この提案手法の既存手法に対する立ち位置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、Fig. ?? に示すように、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、パーミュテーションの正誤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データに対して、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。

### 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を扱う際に必要となるICAの基本原理やSTFTに加え、パーミュテーション問題を回避するような手法であるIVAやILRMA、DNNに基づく既存の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、提案手法を取り扱う際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DNNに基づく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章では音声の混合信号を模倣した人工データと実際の音声データに対する音源分離実験を行い、提案手法におけるパーミュテーション解決性能の検証を行う。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、提案手法の基礎理論となる音源分離手法の 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、短時間フーリエ変換 (short-time Fourier transform: 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 の定式化を導入し、2.5 節以降は、定式化したものを用いて説明する。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、**本提案手法の解決すべき課題であるパーミュテーション問題**と呼ばれる FDICA に伴う問題の説明と、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回避するような音源分離手法である IVA 及び ILRMA について詳細を述べる。2.8 節では、**本提案手法と既存の DNN に基づく手法の違いを理解するために、既存の DNN を用いたパーミュテーション解決法**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

本項では、BSSの基礎であるICAについて説明する。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デル化できる。

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

ここで、信号の伝搬を表す係数  $a_{mn}$  は、時刻  $\textcolor{blue}{l}$  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  $\mathbf{x}(l) = \underline{(x_1(l), x_2(l))^T} [\mathbf{x}_1(l), \mathbf{x}_2(l)]^T$  信号源ベクトル  $\mathbf{s}(l) = \underline{(s_1(l), s_2(l))^T} [\mathbf{s}_1(l), \mathbf{s}_2(l)]^T$  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  $\mathbf{y}(l) = \underline{(y_1(l), y_2(l))^T} [\mathbf{y}_1(l), \mathbf{y}_2(l)]^T$ 、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  $\mathbf{W}$  を選択推定することで、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、分離フィルタ行列  $\mathbf{W}$  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、**例えれば  $s_1(l)$  から  $s_2(l)$  を推定予測や説明**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 (2.9 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。

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

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

ここで、 $\mathfrak{D}_{KL} [p(s) || q(s)]$  はカルバッケライブラ・ダイバージェンス (Kullback–Leibler divergence: KL divergence) 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。

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

また、分離フィルタ行列  $\mathbf{W}$  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  $p(\mathbf{y}) = p(y_1, y_2 | \mathbf{x}_1, \mathbf{x}_2)$  と  $p(\mathbf{x}) = p(x_1, x_2 | \mathbf{y}_1, \mathbf{y}_2)$  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  $\mathfrak{J}(\mathbf{W})$  は以下のように書ける。

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

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  $\mathbf{W}$  について最小化することで、信号源を分離する。

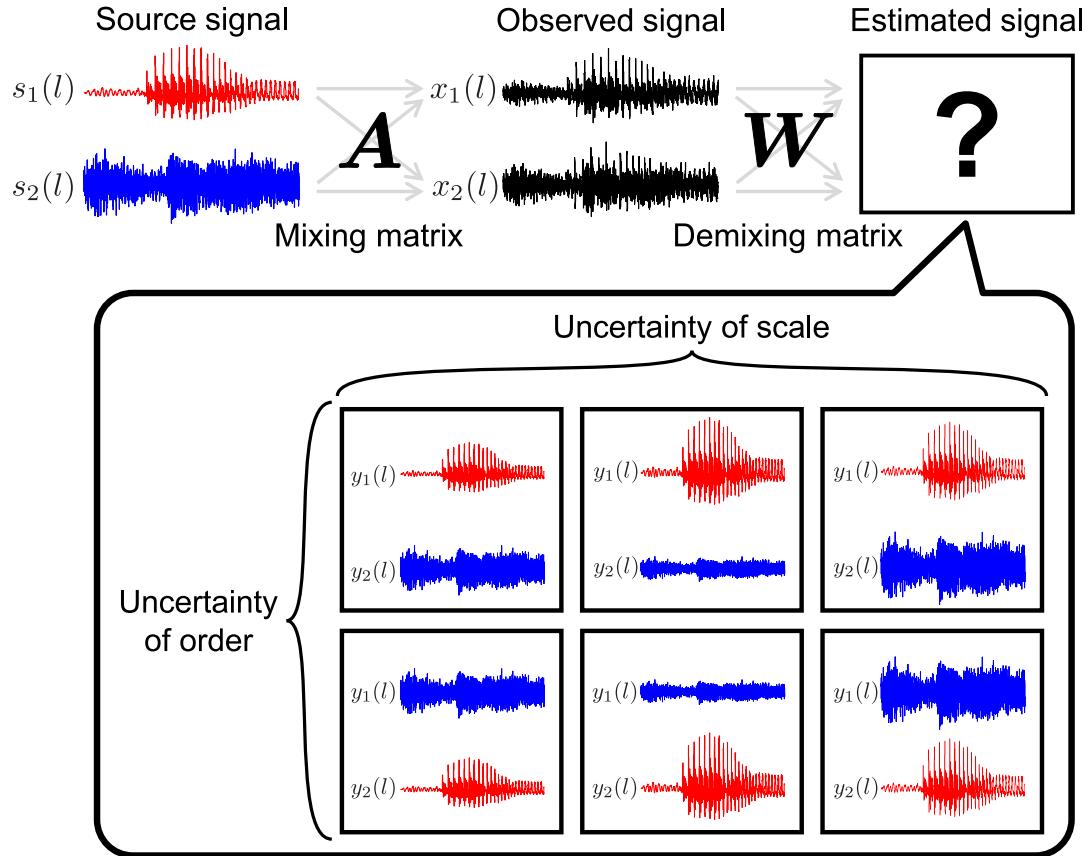


Fig. 2.1: Uncertainty in ICA. ICA cannot determine order and scales of estimated signals.

### 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  $W$  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。

- (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 1 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 2 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] と呼ばれる補正方法が提案されている。

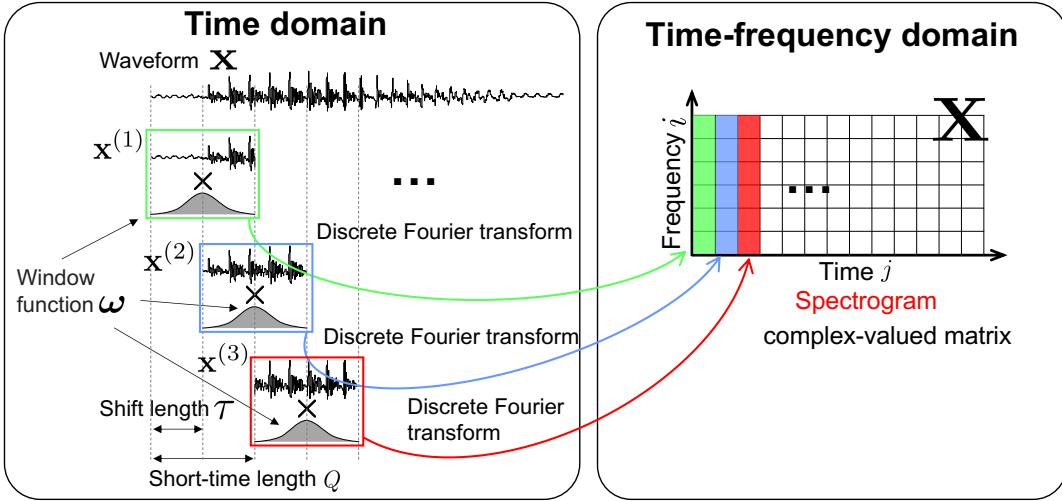


Fig. 2.2: Mechanism of STFT. Each of windowed short-time signals are transformed to frequency domain by discrete Fourier transform.

## 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。

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

STFT の分析窓関数の長さ及びシフト長をそれぞれ  $Q$  及び  $\tau$  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。

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

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

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。

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

また、各時間フレームの信号のSTFTは次式のようにして求められる。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。

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

スペクトログラム  $Z$  のここで、 $\mathbf{X}$  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。

$$\mathbf{x}_{ij} = \sum_{q=1}^Q \omega(q) \mathbf{x}^{(j)}(q) \exp \left\{ \frac{-\imath 2\pi(q-1)(i-1)}{F} \right\} \quad (2.21)$$

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) を、 $i = 1, 2, \dots, I$  は周波数ビンのインデクスを、 $\imath$  は虚数単位を、 $\boldsymbol{\omega} = [\omega(1), \omega(2), \dots, \omega(Q)]^T \in \mathbb{R}^Q$  は分析窓関数短時間信号  $\mathbf{x}^{(j)}$  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、時間領域の信号は一定幅短時間ごとに分析窓関数を乗じて離散フーリエ変換を行うことで、短時間信号に分割して分析窓関数を乗じて離散フーリエ変換を適用し、横軸が時間、縦軸が周波数のスペクトログラムと呼ばれる複素行列複素時間周波数行列  $Z$  を表すことができる。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

今一度本節以降、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、各観測音源信号をSTFTすることで得られる、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 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。

$$\mathbf{s}_{ij} = [s_{ij,1}, s_{ij,2}, \dots, s_{ij,n}, \dots, s_{ij,N}]^T \in \mathbb{C}^N \quad (2.22)$$

$$\mathbf{x}_{ij} = [x_{ij,1}, x_{ij,2}, \dots, x_{ij,m}, \dots, x_{ij,M}]^T \in \mathbb{C}^M \quad (2.23)$$

$$\mathbf{z}_{ij} = [z_{ij,1}, z_{ij,2}, \dots, z_{ij,n}, \dots, z_{ij,N}]^T \in \mathbb{C}^N \quad (2.24)$$

と表す。

また、複素スペクトログラム行列  $\mathbf{S}_n \in \mathbb{C}^{I \times J}$ ,  $\mathbf{X}_m \in \mathbb{C}^{I \times J}$ ,  $\mathbf{Z}_n \in \mathbb{C}^{I \times J}$  の成分をそれぞれ  $s_{i,j,n}$ ,  $x_{i,j,m}$  及び  $z_{i,j,n}$  と表す。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に線形分離行列な分離行列を推定することで BSS を

実現する手法である。実際に観測される音声信号には残響の影響を受けており、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、線形時不变な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。

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

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（音源  $n$  からマイクロフォン  $m$  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まとめたもの時間  $l$  毎にまとめたもの）である。これを式 (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は容易ではないことから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 は困難である。この問題を解決するために、式 (2.25) 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、STFT によって周波数領域上での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  $\mathbf{A}_i = (\mathbf{a}_{i,1} \ \mathbf{a}_{i,2} \ \cdots \ \mathbf{a}_{i,n} \ \cdots \ \mathbf{a}_{i,N}) \in \mathbb{C}^{M \times N}$   $\mathbf{A}_i = [\mathbf{a}_{i1} \ \mathbf{a}_{i2} \ \cdots \ \mathbf{a}_{in} \ \cdots \ \mathbf{a}_{iN}] \in \mathbb{C}^{M \times N}$  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、STFT の窓長が室内残響よりも長い場合にのみ成立観測信号収録時の残響長よりも 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、分離行列周波数毎の分離行列  $\mathbf{W}_i = \mathbf{A}_i^{-1} = (\mathbf{w}_{i,1} \ \mathbf{w}_{i,2} \ \cdots \ \mathbf{w}_{i,n} \ \cdots \ \mathbf{w}_{i,N})^H \mathbf{W}_i = \mathbf{A}_i^{-1} = [\mathbf{w}_{i1} \ \mathbf{w}_{i2} \ \cdots \ \mathbf{w}_{in} \ \cdots \ \mathbf{w}_{iN}]^H$  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、周波数  $i$  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 ICA (複素数の) 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することで音源分離を行うし、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、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 2.2.3 項で述べた通り、分離音源分離された推定信号の周波数毎のスケ

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  $\mathbf{w}_{i/n}$  の順番を入れ変えうるパーミュテーション行列（置換行列）である。 $\mathbf{D}_i \in \mathbb{R}^{N \times N}$  は、 $\mathbf{w}_{i/n}$  のスケールを変化させる可能性のある対角行列である。即ち、FDICA で推定される分離信号

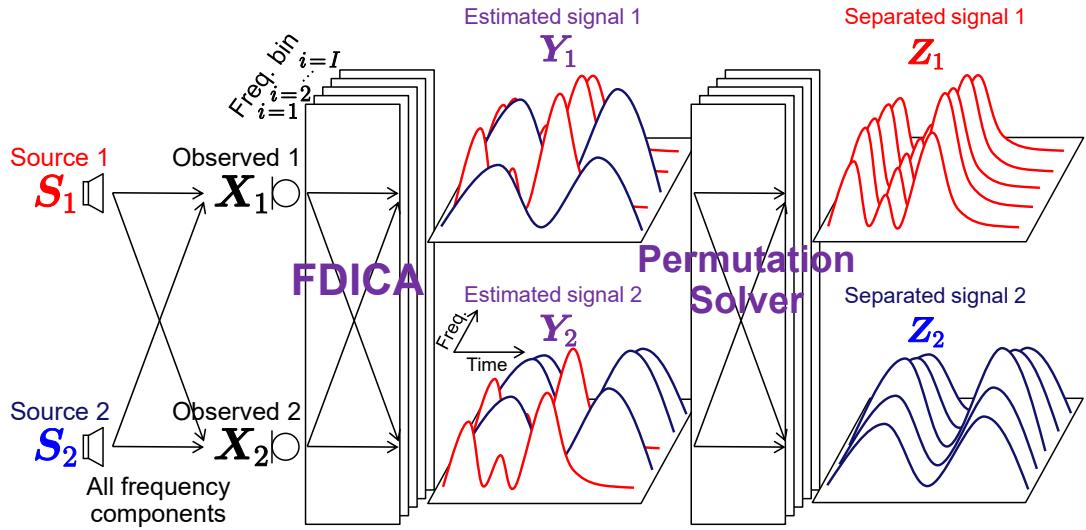
$$\mathbf{y}_{ij} = \hat{\mathbf{W}}_i \mathbf{x}_{ij} \quad (2.29)$$

$$= \left[ y_{ij/1}, y_{ij/2}, \dots, y_{ij/n}, \dots, y_{ij/N} \right]^T \in \mathbb{C}^N \quad (2.30)$$

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、[プロジェクションバック法](#)時間領域での ICA の場合と同様に[プロジェクションバック法](#) [15] で復元可能である。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[純粹に](#) $I$  個の全周波数ビンに関して復元することは、組み合わせ爆発が[発生するため](#)生じるため容易ではない。具体的には、 $I$  個の周波数ビンのそれだけで  $N$  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[推定される分離得られる推定信号](#)  $\mathbf{y}_{ij}$  の[音源毎の複素スペクトログラム行列を](#) $n$  番目のスペクトログラムを  $\mathbf{Y}_n \in \mathbb{C}^{I \times J}$  [で表していると定義](#)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[グループ](#)音源の[分離信号](#)分離成分が 1 つの[時間周波数構造](#)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ビンにわたって正しく並べ直す必要がある。[パーミュテーション問題を解決して得られる分離信号は次式となる。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。](#)

$$\mathbf{z}_{ij} = \mathbf{P}_i^{-1} \mathbf{D}_i^{-1} \mathbf{y}_{ij} \quad (2.31)$$

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。したがって、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  $\mathbf{P}^{-1}$  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] [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。[またこのとき](#)、離れた周波数においても、同じ音源のアクティ

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

ペ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、とくに複数音声の混合信号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、同じ音源の分離信号は全周波数でまとめて出力されるような分離モデルとなっており、実際に複数の周波数ビンで同時に共起する成分が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。また、NMF [10] と IVA を組み合わせた BSS である ILRMA [11, 12] は同一源の時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おり IVA と同様にパーミュテーション問題を音源モデルに基づいて可能な限り回避するようなモデルとなっている。さらに、IVA の音源信号の

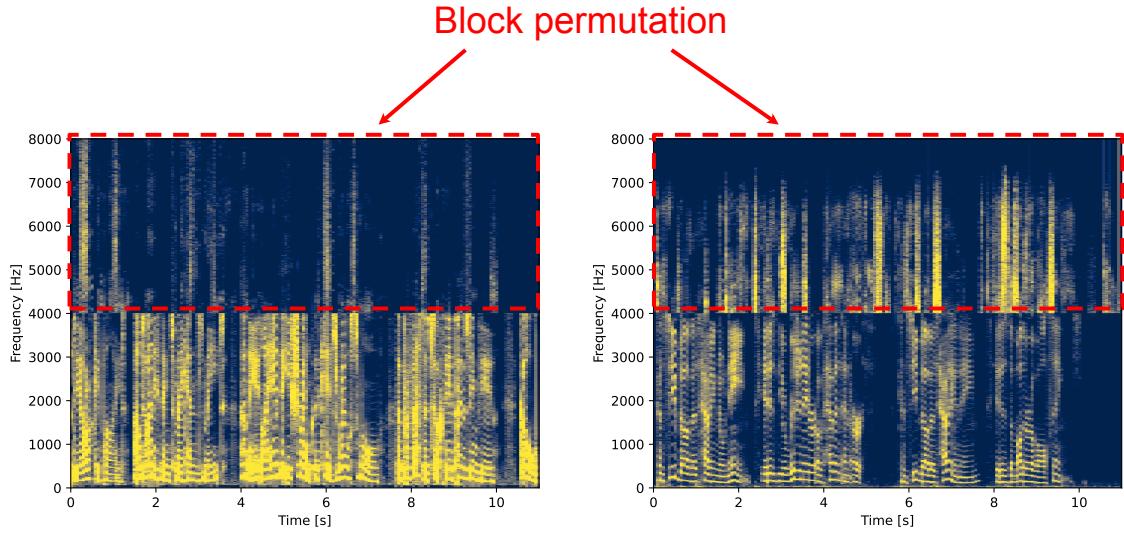


Fig. 2.4: Example of block permutation problem.

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、音声と音声の混合信号の様な分離タスクの場合、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、音声信号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、音声信号のパワースペクトログラムを低ランクで表現することが難しいことが原因と予想される。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。また例えば、IVAやILRMAにおいても、まとまった周波数帯域でパーティションが入れ替わる問題（ブロックパーティション問題）[16]が報告されている。Fig. 2.4にブロックパーティション問題の様子を示す。Fig. 2.4では、4000Hz以上4 kHz以上の周波数帯がまとめて反転していることが分かる。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。そのため、このような事実からも、依然としてパーティション問題の解決が不十分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ティションであれば、ユーザアノテーションにより修正するような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、多くの帯域にブロックパーティ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

近年では、DNNを用いたパーミュテーション問題解決法が登場している。観測された混合信号 $\mathbf{X}_n$ にFDICA適用すると、パーミュテーション問題が生じた分離信号 $\mathbf{Y}_n$ が得られる。これらのパワースペクトログラム $|\mathbf{Y}_n|^2$ から全周波数帯域中の局所的な狭帯域（サブバンド）を定義し、サブバンド毎にデータをDNNに入力し、パーミュテーション問題を解決する。サブバンド毎に参照周波数を定義し、その近傍周波数が参照周波数に対して同一音源か否かを判断し、同一音源である場合はDNNの出力として「0」を出力し、同一音源でない場合はDNNの出力として「1」を出力する。この結果を時間方向にずらして、全時間フレームに対するDNNの予測処理を走査する。そして、DNNの予測結果を時間方向に対して多数決処理を行うことで、より信頼性の高いサブバンドベクトルを取得する。サブバンドベクトルは、基準周波数 $i$ をシフトすることにより全周波数を推定する。ただし、各サブバンドベクトル内の2値は（「0」と「1」）は異なる意味を持つ可能性がある。これはサブバンド内の周波数成分が、参照周波数の成分と同一音源か否かを示しているに過ぎず、参照周波数の変化と共に、対応音源が変化する。2音源の場合を考えるとサブバンドベクトル内の値が「1」、つまり同一音源ではない時、必然的にもう一方の音源を指すこととなる。但し、3音源以上になるとサブバンドベクトル内の値が「1」の時、残りのどの音源のことを指すのかが判断できない。3音源以上になると組み合せ爆発を起こしてしまい、計算量の観点から3音源以上の音源分離は難しい。

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、合致しなければブ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 値分類 DNN を学習しており、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、DNNに基づくパーミュテーション問題の解決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、深層パ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 に伴い生じるパーミュテーション問題について説明した。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ような手法である、IVA と ILRMA について説明した。次章以降では、より簡潔に精度の良い BSS を達成するために 2.5節で導入した 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を新たに提案する。

## 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に伴い生じるのパーミュテーション問題と従来の深層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、従来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、組み合せ爆発を起こすことのない、DN $N$ を用いたデータ駆動型パーミュテーション解決法を新たに提案する。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。3.2節では、IVAやILRMAのようなブラインド（教師なし）なパーミュテーション解決法における課題と従来の深層パーミュテーション解決法における課題を述べ、データ駆動型の教師ありパーミュテーション解決法を新たに提案する動機について明らかにする。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、提案本論文で提案する深層パーミュテーション解決法における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の並び替えに用いる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ラベルの取得方法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における最適な窓長を短時間区間長（窓長） $Q$ について実験的に検討調査している。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、これは即ち分離性能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、残響状態時間が $T_{60}=470\text{ ms}$ の条

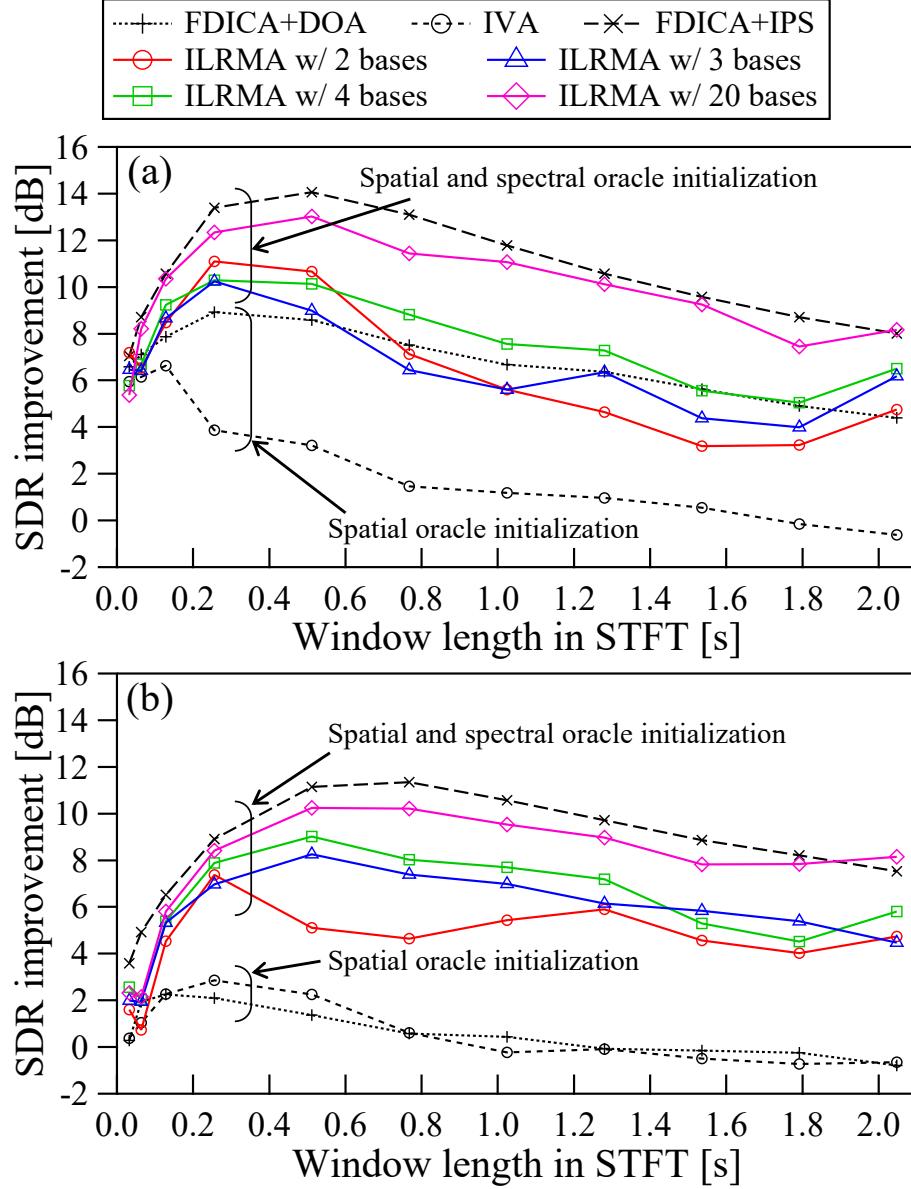


Fig. 3.1: Average source separation results for speech signals using random initialization: (a) E2A ( $T_{60} = 300$  ms) and (b) JR2 ( $T_{60} = 470$  ms) impulse responses. For details of this figure, see [13].

件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、高残響下での音声混合信号音声信号の混合という難しい観測条件であっても、 $\hat{W}_i$  は FDICA で正確に推定でき、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ることを示して

いる。すなわち、残る課題は推定信号  $y_{ij}$  を正しい順番に並び変えるパーティション問題の解決 ( $P_i^{-1}$  の推定のみ失敗していることを示している。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、従来の深層パーティション解決法では、全周波数帯域中の局所的な狭帯域におけるサブバンド内のパーティション問題の解決を全時間方向と全周波数方向に行う際に、~~を~~を解決する際に、ある参照周波数ビンに対して同一か否かで音源を判断しているため、3音源以上の分離等の拡張性に欠ける。その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 ( $NC_2$ ) 分適用せねばならず、さらに後段のサブバンド間のパーティション問題の解決（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ティション問題を正確に解くことに焦点を当て、新しいDNNに基づくデータ駆動型（教師あり）パーティション解決法深層パーティション解決法を提案する。以後、本論文では、提案するパーティ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に、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に模倣した人工データと実際の音声データ音響信号を用いてパーティション問題の解決を考える解決性能を実験的に調査する。この際、音源数  $N = 2$  及びチャネル数  $M = 2$  と仮定し、実験を行う。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。

~~提案する本論文で提案する深層パーティション解決法の概要を適用する処理の概要は以下の通りである。~~

- 分離信号  $\mathbf{Y}_1$  及び  $\mathbf{Y}_2$  から全周波数成分の値を保持するミニ振幅スペクトログラムを取り出し、それぞれに対して全ての音源のパワースペクトログラムの値を基準にして正規化を行った値をDNNに入力する
- DNNは入力された2つのミニ振幅スペクトログラムの値がどの音源の値かを予測し、0~1の間の確率値として出力する
- DNNから出力された確率値に従ってシャッフルされたスペクトログラムを並び替えた行列と、完全に分離されたスペクトログラムとの間で損失を取得する
- $\mathbf{Y}_1$  及び  $\mathbf{Y}_2$  の全時間方向に対してDNNが適用される
- 最終的な推定値（ラベル） $\hat{\mathbf{L}}$ は、予測値の時間方向への多数決結果から決定される

- (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 (2.31) 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

~~提案するパーミュテーション解決法では、全周波数成分を持ったミニ振幅スペクトログラムに対して、どの音源の成分が入っているかをDNNで予測し、その予測結果に基づいてパーミュテーション解決を行う。また、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。DNNには大量の学習用データが必要であるが、IPSで理想的にパーミュテーション解決された分離信号  $\mathbf{Z}_n$  を周波数毎にランダムにシャッフルすることで、容易かつ大量に生成することができる。~~

### 3.3 DNN の入出力

~~観測された混合信号  $\mathbf{X}_n$  に FDICA を適用すると、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生じた分離信号  $\mathbf{Y}_n$  が得られる。が発生した状態の推定信号  $(\mathbf{Y}_1, \mathbf{Y}_2)$  が得られる。DNNへの入力は、各分離信号のパワースペクトログラム成分を全ての分離信号のパワースペクトログラム成分で割った値を用いる。即ち、2音源の場合 DNN の入力に用いる信号成分は次のようになる。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  $(\mathbf{Y}_1, \mathbf{Y}_2)$  のパワースペクトラム ( $|\mathbf{Y}_1|^2, |\mathbf{Y}_2|^2$ ) の比率に変換する正規化 [5] を施す。この処理~~

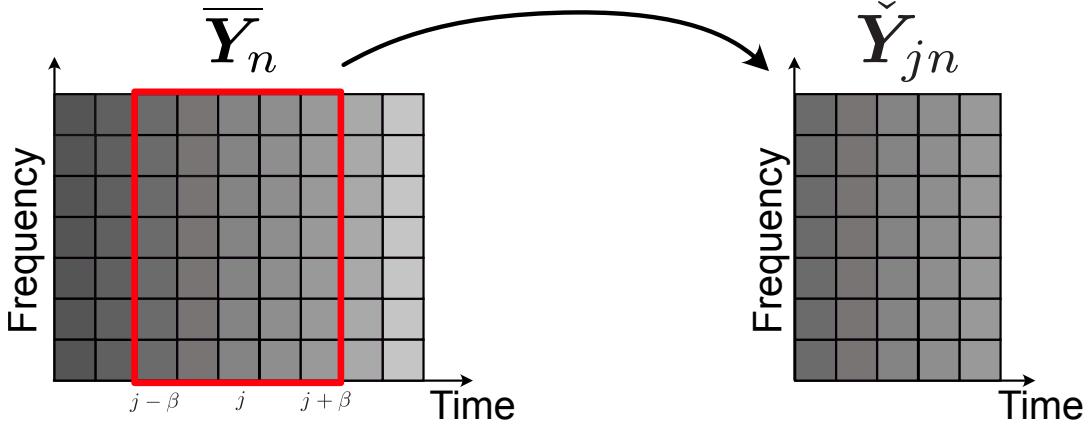


Fig. 3.2: Create a local time-amplitude spectrogram.

は次式で表される。

$$\hat{\mathbf{Y}}_1 = \frac{|\mathbf{Y}_1|^2}{|\mathbf{Y}_1|^2 + |\mathbf{Y}_2|^2} \quad (3.1)$$

$$\hat{\mathbf{Y}}_2 = \frac{|\mathbf{Y}_2|^2}{|\mathbf{Y}_1|^2 + |\mathbf{Y}_2|^2} \quad (3.2)$$

$$\bar{\mathbf{Y}}_1 = \frac{|\mathbf{Y}_1|^2}{|\mathbf{Y}_1|^2 + |\mathbf{Y}_2|^2} \in [0, 1]^{I \times J} \quad (3.3)$$

$$\bar{\mathbf{Y}}_2 = \frac{|\mathbf{Y}_2|^2}{|\mathbf{Y}_1|^2 + |\mathbf{Y}_2|^2} \in [0, 1]^{I \times J} \quad (3.4)$$

ここで、DNNの入力に用いる値をそれぞれ  $\hat{\mathbf{Y}}_1 \in \mathbb{R}_{\geq 0}^{I \times J}$ ,  $\hat{\mathbf{Y}}_2 \in \mathbb{R}_{\geq 0}^{I \times J}$  とする。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数乗は要素毎の指数乗、及び行列間のベクトルは要素毎の商を示している。この時、 $i = 1, \dots, I$  及び  $j = \tau, \dots, J - \tau$  はそれぞれ全周波数帯域の周波数ビン及び時間フレームのインデックスである。DNNの入力にはミニ振幅スペクトログラムを用いるので、元のスペクトログラムの値からはみ出ることがないように  $j$  の範囲を限定的に入れている。ここで、行列の  $|\cdot|^2$  は、要素ごとの絶対値の二乗を示す。時間フレーム  $j$  における分離信号を次式で表す。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  $(\bar{\mathbf{Y}}_1, \bar{\mathbf{Y}}_2)$  から、Fig. 3.2 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。

$$\hat{\mathbf{y}}_{1j} = [\hat{y}_{11j}, \hat{y}_{12j}, \dots, \hat{y}_{1Ij}]^T \quad (3.5)$$

$$\hat{\mathbf{y}}_{2j} = [\hat{y}_{21j}, \hat{y}_{22j}, \dots, \hat{y}_{2Ij}]^T \quad (3.6)$$

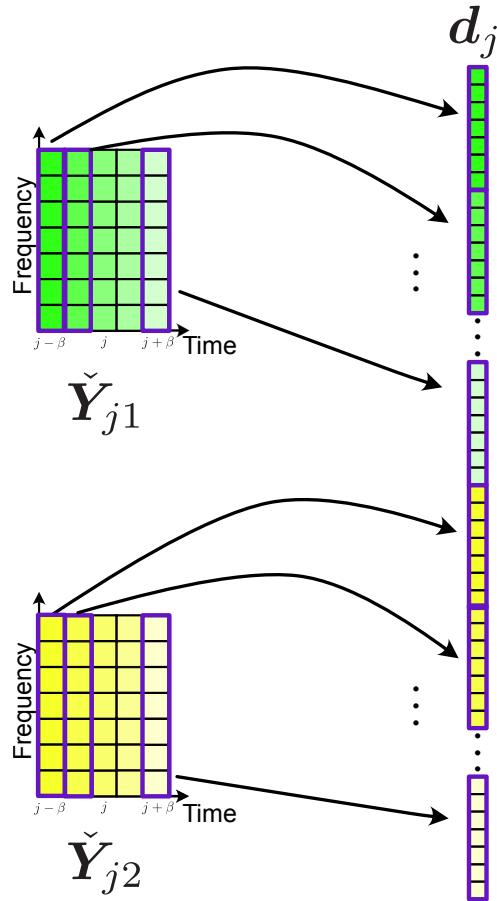


Fig. 3.3: Input vector of DNN.

$$\check{Y}_{j1} = [\bar{y}_{(j-\beta)1}, \bar{y}_{(j-\beta+1)1}, \dots, \bar{y}_{(j-1)1}, \bar{y}_{j1}, \bar{y}_{(j+1)1}, \dots, \bar{y}_{(j+\beta)1}] \in [0, 1]^{I \times (2\beta+1)} \quad (3.7)$$

$$\check{Y}_{j2} = [\bar{y}_{(j-\beta)2}, \bar{y}_{(j-\beta+1)2}, \dots, \bar{y}_{(j-1)2}, \bar{y}_{j2}, \bar{y}_{(j+1)2}, \dots, \bar{y}_{(j+\beta)2}] \in [0, 1]^{I \times (2\beta+1)} \quad (3.8)$$

ここで、 $\hat{y}_{1ij}$  は  $\check{Y}_1$  の  $ij$  要素であり、 $\hat{y}_{2ij}$  は  $\check{Y}_2$  の  $ij$  要素を表す。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ル）を表す。また、 $\beta$  ( $0$  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。DNN の入力として与える情報提案手法では、DNN の入力ベクトルは、 $j$  近傍の時間フレームの列ベクトルを結合したベクトルとする。式 (3.7) 及び (3.8) 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. 3.3 のように一次元に整形（ベクトル化）したベクトルである。これを  $x_j$  とおくと、次式のように構成される。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。また、 $j$ 近傍の時間フレームの幅は $\tau$ とする。

$$\mathbf{x}_j = [\widehat{\mathbf{y}}_{1(j-\tau)}^T, \dots, \widehat{\mathbf{y}}_{1(j+\tau)}^T, \widehat{\mathbf{y}}_{2(j-\tau)}^T, \dots, \widehat{\mathbf{y}}_{2(j+\tau)}^T]^T \quad (3.9)$$

$$\mathbf{d}_j = \text{vec}([\check{\mathbf{Y}}_{j1} \check{\mathbf{Y}}_{j2}]) \in [0, 1]^{2I(2\beta+1)} \quad (3.10)$$

Fig. 3.3に示すように、上記の $\mathbf{x}_j$ がDNNの入力ベクトルとなる。ここで、 $[\check{\mathbf{Y}}_{j1} \check{\mathbf{Y}}_{j2}] \in [0, 1]^{I \times (4\beta+2)}$ は2つの信号の正規化局所時間振幅スペクトログラムを列方向に結合した行列を表す。

DNNによる予測は次式で表される。

$$\underline{\mathbf{L}} = \text{DNN}(\mathbf{x}_j) \quad (3.11)$$

$$\hat{\mathbf{l}}_j = \text{DNN}(\mathbf{d}_j) \in [0, 1]^{2I} \quad (3.12)$$

DNNが出力する予測は $\mathbf{L} \in \mathbb{R}_{[0,1]}^{2 \times I}$ であり、確率値を示す。 $\mathbf{L}$ の1行目には、各周波数成分における音源1である確率値、2行目には、各周波数成分における音源2である確率値が代入される。ここで、 $\hat{\mathbf{l}}_j = [\hat{l}_{11j}, \hat{l}_{21j}, \dots, \hat{l}_{11j}, \hat{l}_{12j}, \hat{l}_{22j}, \dots, \hat{l}_{12j}]$ は出力である予測ベクトルを表す。Fig. 3.4のように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。

$$\hat{\mathbf{L}}_j = \text{mat}(\hat{\mathbf{l}}_j) \in [0, 1]^{I \times 2} \quad (3.13)$$

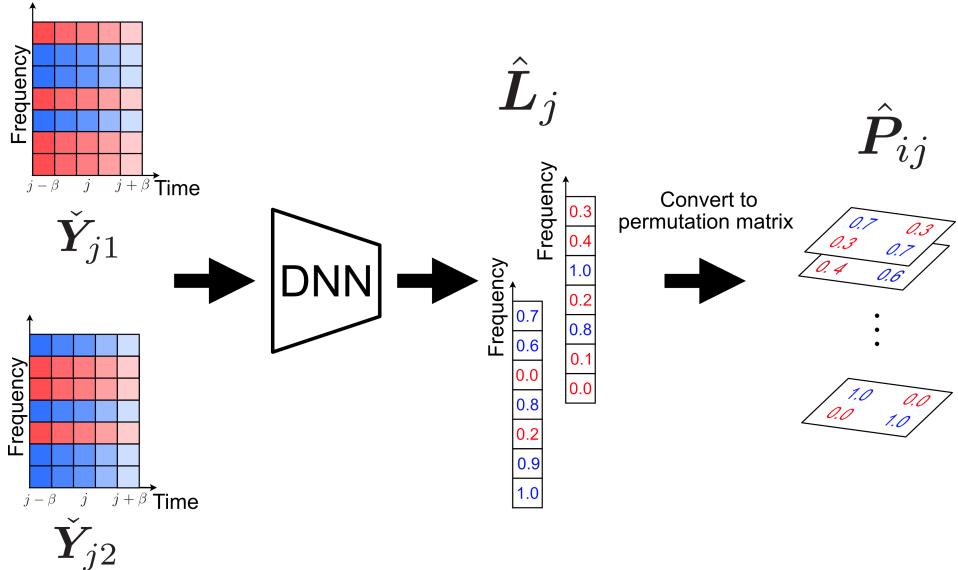
再成型された行列 $\hat{\mathbf{L}}_j$ はFig. 3.4に示すように、2つのパーミュテーション問題が生じている入力信号 $(\check{\mathbf{Y}}_{j1}, \check{\mathbf{Y}}_{j2})$ の各周波数成分のそれぞれが「1番目の音源の成分である確率 $l_{i1}$ 」と「2番目の音源の成分である確率 $l_{i2}$ 」を $\mathbf{d}_j$ から予測したものと定義し、提案手法ではこの定義に基づいて正確な予測ができるDNNを学習する。ここで、 $(l_{i1}, l_{i2})$ は離散確率値であるため $l_{i1} + l_{i2} = 1$ 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。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 $\mathbf{P}_i$ 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 $\mathbf{d}_j$ の予測結果 $\hat{\mathbf{L}}_j$ 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 $\hat{\mathbf{L}}$ （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。

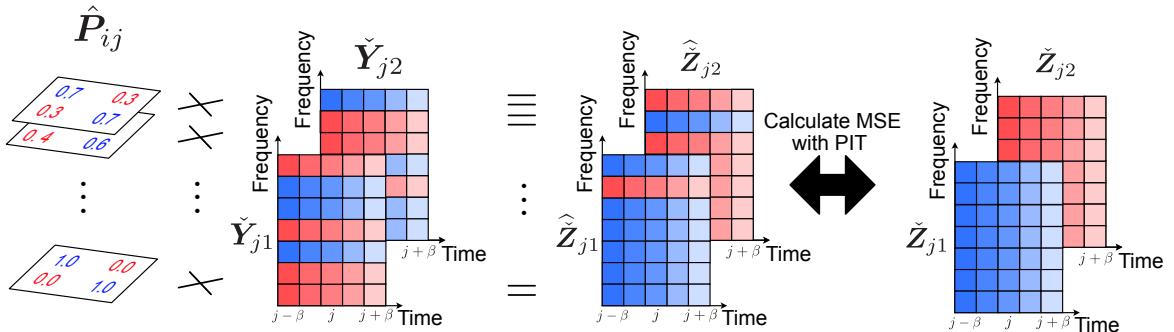
$$\hat{\mathbf{P}}_i = \begin{bmatrix} \hat{l}_{i1} & \hat{l}_{i2} \\ \hat{l}_{i2} & \hat{l}_{i1} \end{bmatrix} \quad (3.14)$$

ここで、 $\hat{l}_{i1}$ 及び $\hat{l}_{i2}$ は $\hat{\mathbf{L}}$ 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、単位行列及び1行列をそれぞれ $\mathbf{I}$ 及び $\mathbf{1}$ と表記すると、 $N = 2$ の場合は

$$\mathbf{P}_i = \mathbf{I} \quad \text{or} \quad \mathbf{1} - \mathbf{I} \quad (3.15)$$



(a) Calculation of predicted permutation matrix.



(b) Calculation of MSE with PIT

Fig. 3.4: Process fo calculating loss function value.

のいずれかとなる。推定パーミュテーション行列  $\hat{P}_i$  は式 (3.14) であるため、予測が不完全であれば  $I$  又は  $1 - I$  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して、どのパーミュテーション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ができる。

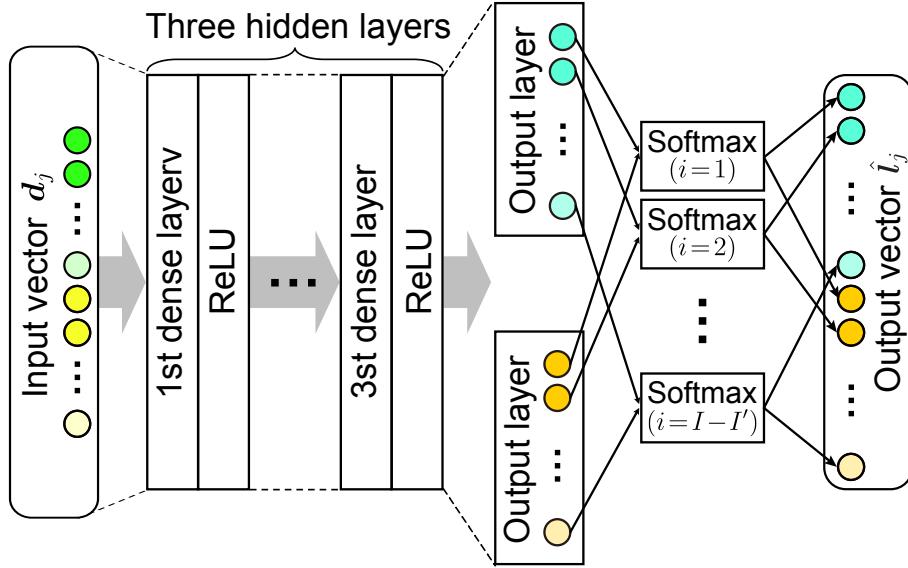


Fig. 3.5: DNN architecture.

## 3.4 DNN の構造

Fig. 3.5 に提案手法の DNN の構造を示す。提案する 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 (dense layer) からなる全結合構成 MLP となっており、1~5番目の隠れ層には隠れ層の 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<sup>\*1</sup>。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。最終隠れ層には softmax 関数を適用している。各隠れ層の次元数は全て、4096 である。

## 3.5 DNN 学習時の損失関数

誤差逆伝播を行う際の損失は、DNN が output した確率値に従ってパーミュテーション行列を並び替えた行列と完全に分離された行列との間で平均二乗誤差 (mean squared error: MSE) を使用して得る。3.3 節より、 $\hat{l}_n$  は、全周波数帯域における音源  $n$  の成分が含まれる割合で構成されている。また、次式でパーミュテーション行列の並び替えを行い、予測行列

<sup>\*1</sup> 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。

$\tilde{\mathbf{Y}}_{1j} \in \mathbb{R}_{\geq 0}^{I \times (2\tau+1)}$  と  $\tilde{\mathbf{Y}}_{2j} \in \mathbb{R}_{\geq 0}^{I \times (2\tau+1)}$  の2種類を導く。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<sup>\*2</sup>の負対数尤度関数であ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 $(\mathbf{Y}_1, \mathbf{Y}_2)$ 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、実際はパーミュテーション問題を解決しなくと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 $(\mathbf{Y}_1, \mathbf{Y}_2)$ 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）を用いる。

$$\text{MSE}(\hat{\mathbf{V}}, \mathbf{V}) = \frac{1}{RC} \|\hat{\mathbf{V}} - \mathbf{V}\|_{\text{Fr}}^2 \quad (3.16)$$

$$= \frac{1}{RC} \sum_{r,c} (\hat{v}_{rc} - v_{rc})^2 \quad (3.17)$$

ここで、 $\hat{v}_{rc}$  及び  $v_{rc}$  はそれぞれ行列  $\hat{\mathbf{V}}$  及び  $\mathbf{V}$  の要素、 $r = 1, 2, \dots, R$  及び  $c = 1, 2, \dots, C$  はそれぞれ行列  $\hat{\mathbf{V}}$  及び  $\mathbf{V}$  の行と列のインデクス、 $\|\cdot\|_{\text{Fr}}$  は Frobenius ノルムである。次に、DNNの入力である正規化局所時間振幅スペクトログラム $(\check{\mathbf{Y}}_{j1}, \check{\mathbf{Y}}_{j2})$ に対する予測結果  $\mathbf{L}_j$  と式(3.14)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  $\hat{P}_{ij}$  を構成し、式(2.28)と同様に音源パーミュテーションを並び替えた予測分離信号を $(\hat{Z}_{j1}, \hat{Z}_{j2})$ と定義する。さらに、この予測分離信号に対する正解ラベル（Fig. 3.2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。これらの定義と式(3.17)を用いて、前述の誤差関数  $\mathcal{L}$  は、 $(\hat{Z}_{j1}, \hat{Z}_{j2})$  及び  $(\check{Z}_{j1}, \check{Z}_{j2})$  間の MSE として次式で表せる。

$$\mathcal{L} = \text{MSE}(\hat{Z}_{j1}, \check{Z}_{j1}) + \text{MSE}(\hat{Z}_{j2}, \check{Z}_{j2}) \quad (3.18)$$

ここで、 $y_{1ij}$  は  $\mathbf{Y}_1$  の  $ij$  要素であり、 $y_{2ij}$  は  $\mathbf{Y}_2$  の  $ij$  要素を表す。 $\tilde{\mathbf{Y}}_{1j} \in \mathbb{R}_{\geq 0}^{I \times (2\tau+1)}$ 、 $\tilde{\mathbf{Y}}_{2j} \in \mathbb{R}_{\geq 0}^{I \times (2\tau+1)}$  と完全に分離された信号のミニスペクトログラム成分  $Z_{1j} \in \mathbb{R}_{\geq 0}^{I \times (2\tau+1)}$ 、 $Z_{2j} \in \mathbb{R}_{\geq 0}^{I \times (2\tau+1)}$

<sup>\*2</sup> 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。

でMSEを使用し、損失を得る。また、本論文ではパーミュテーション問題を解くことを考えており、分離信号の順番には触れないと、式で損失を計算する。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(3.18)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、損失関数を次式で定義する。

$$\mathcal{L} = \min \left( \text{MSE}(\hat{\mathbf{Z}}_{j1}, \check{\mathbf{Z}}_{j1}) + \text{MSE}(\hat{\mathbf{Z}}_{j2}, \check{\mathbf{Z}}_{j2}), \text{MSE}(\hat{\mathbf{Z}}_{j1}, \check{\mathbf{Z}}_{j2}) + \text{MSE}(\hat{\mathbf{Z}}_{j2}, \check{\mathbf{Z}}_{j1}) \right) \quad (3.19)$$

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、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の出力の順序には依存しないような学習が可能となる。

$$\text{Loss} = \text{MIN}\{\text{MSE}(\tilde{\mathbf{Y}}_{1j}, \tilde{\mathbf{Z}}_{1j}) + \text{MSE}(\tilde{\mathbf{Y}}_{2j}, \tilde{\mathbf{Z}}_{2j}), \text{MSE}(\tilde{\mathbf{Y}}_{1j}, \tilde{\mathbf{Z}}_{2j}) + \text{MSE}(\tilde{\mathbf{Y}}_{2j}, \tilde{\mathbf{Z}}_{1j})\} \quad (3.20)$$

DNNは上記のLossを最小化するように学習を行う。これらの処理をこの損失関数の計算の処理を Fig. 3.4 に示す。最終的なラベル  $\tilde{\mathbf{L}}$  は、 $\mathbf{L}$  の列成分を比較し、値が大きい方のインデックスを含んだベクトルとなる。

### 3.6 学習済の DNN のテストデータへの適用

音声信号は本来、無音区間が多く存在することから、一定区間の長さの成分を持つ  $\hat{\mathbf{Y}}_1$  や  $\hat{\mathbf{Y}}_2$  はほぼ零ベクトルになる可能性があり、その場合DNNの予測は不安定になる。この問題に対処するために、Fig. 3.6に示すように、長さ  $2\tau + 1$  の入力ベクトルをストライド幅1でシフトさせて、全時間フレームに対してDNNの予測処理を走査する。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  $(\mathbf{Y}_1, \mathbf{Y}_2)$  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  $\mathbf{P}_i$  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  $\hat{\mathbf{P}}_{ij}$  もバイナリ行列に変換する
- (b) FDICA等の時不变な分離行列  $\mathbf{W}_i$  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 ( $\mathbf{P}_i$  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  $\hat{\mathbf{P}}_{ij}$  を時間方向に多数決処理し、時不变な行列  $\hat{\mathbf{P}}_i$  に変換する

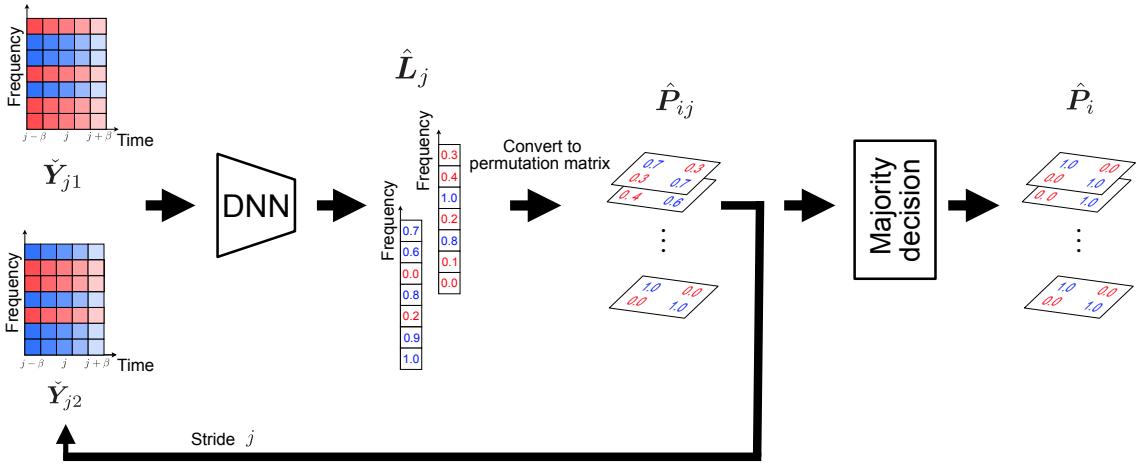


Fig. 3.6: DNN predictions for all short-time ~~subbands~~ spectrograms and their majority decision.

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。

$$\hat{P}_{ij} \leftarrow \text{round}(\hat{P}_{ij}) \in \{0, 1\}^{N \times N} \quad (3.21)$$

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 (3.21) 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 (3.21) 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 (3.21) の処理を適用している。そして、DNN の予測結果を時間軸に関して多数決することで、より信頼性の高いラベル  $\hat{L}$  を得る。この処理は、次のように示される。

一方、上記 (b) については、純粹にパーティ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  $(\check{Y}_{1j}, \check{Y}_{2j})$  は推定信号  $(Y_1, Y_2)$  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  $(\check{Y}_{1j}, \check{Y}_{2j})$  の抽出範囲をストライドさせ、その全てである  $((\check{Y}_{1j}, \check{Y}_{2j}))_{j=1}^J$  を個々に DNN に入力し、全ての予測結果  $((\hat{Z}_{1j}, \hat{Z}_{2j}))_{j=1}^J$  を得ることができる。これらの予測結果を推定パーティション行列  $(\hat{P}_{ij})_{j=1}^J$  に変換し、次式の多数決処理を適用する。

$$\hat{P}_i = \text{round} \left( \frac{1}{J} \sum_{j=1}^J \hat{P}_{ij} \right) \quad (3.22)$$

$\hat{L}$  の値に従って、パーティション行列を並び替えることで、パーティション問題の解決を行う。なお、式 (3.21) 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 (3.22) を計算することで時間方向の平均化ができるため、式 (3.22) は上記 (a) の適用の有無にかかわらず計算する。

### 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。提案手法は、各音源のパワースペクトログラムに対して全ての分離信号のパワースペクトログラムで割ったものをDNNの入力として用いる。また、DNNの出力である確率値を用いてパーミュテーション行列を並び替え、後に完全に分離されたスペクトログラムとの間でMSEを行うことと、時間方向への多数決処理を用いることで、より精度の高い予測ができる。

## 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, FDICA で分離した信号を模倣した行列と実際の音声ファイルを用意し, 提案パーミュテーション解決法を適用する. 後に, 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 DNN に基づく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並び替えができるかを実験的に確認した. 実験には, 人工データと実際の音響信号の 2 つを用いた.

#### 4.2.1 人工データを用いた実験の条件

実験データとして, Fig. 4.1–Fig. 4.3 に示すように, 全ての成分が 0 と 1 の行列, 25 列毎に 0 と 1 の値が入れ替わる行列, 1 列毎に 0 と 1 の値が入れ替わる行列の 3 パターンを使用した. 用意した 3 パターンの行列は, 全て分離信号 ( $\mathbf{Y}_1, \mathbf{Y}_2$ ) とみなして実験を行う. この時の行列のサイズは  $I = J = 100$  とした. また, ブロックパーミュテーションと呼ばれる, ブロック単位で音源分離に失敗することをふまえ, Fig. 4.4 のように 2 行, 4 行, 8 行ごとに各周波数成分をシャッフルした実験も同時に実施した. 1 列毎に 0 と 1 の値が入れ替わる行列に対しては, 4.5 のように 5% の割合で 2 行ごとにシャッフルしそれ以外は 1 行ごとにシャッフルした場合と, 1% の割合で 2 行ごとにシャッフルしそれ以外は 1 行ごとにシャッフルした場合の実験も行った.

学習データには, 完全分離信号  $\check{\mathbf{Z}}_1$  と  $\check{\mathbf{Z}}_2$  の各周波数成分をランダムでシャッフルしたもの用いた. 検証データには学習データにはないシャッフルパターンを用いて  $\check{\mathbf{Z}}_1$  と  $\check{\mathbf{Z}}_2$  の各周

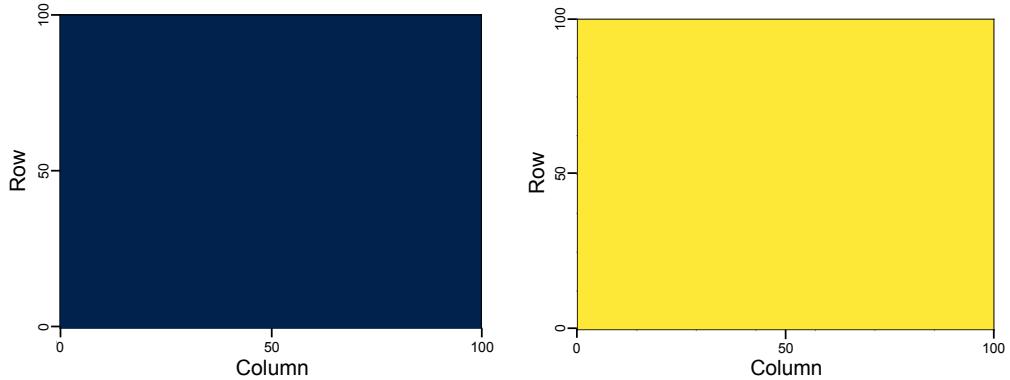


Fig. 4.1: 0 and 1 matrices.

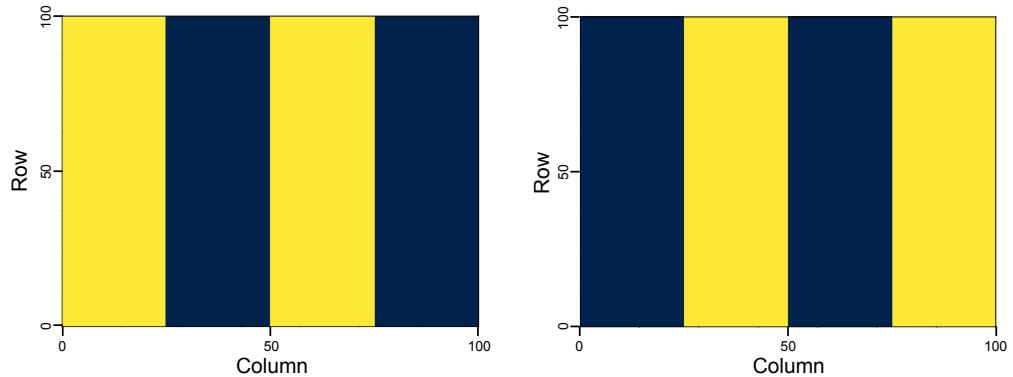


Fig. 4.2: Two matrices with 0 and 1 values changing every 25 columns.

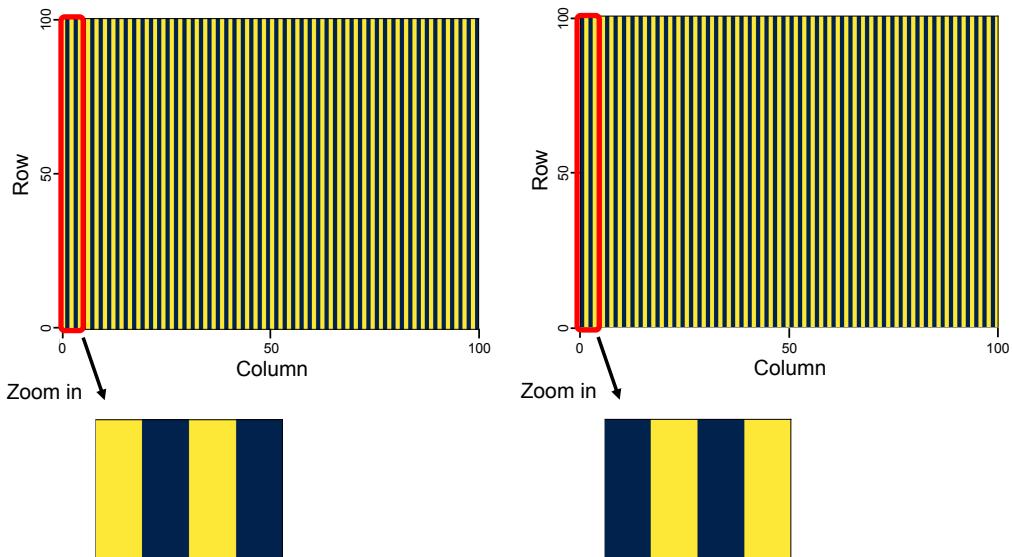


Fig. 4.3: Two matrices with 0 and 1 values changing per column

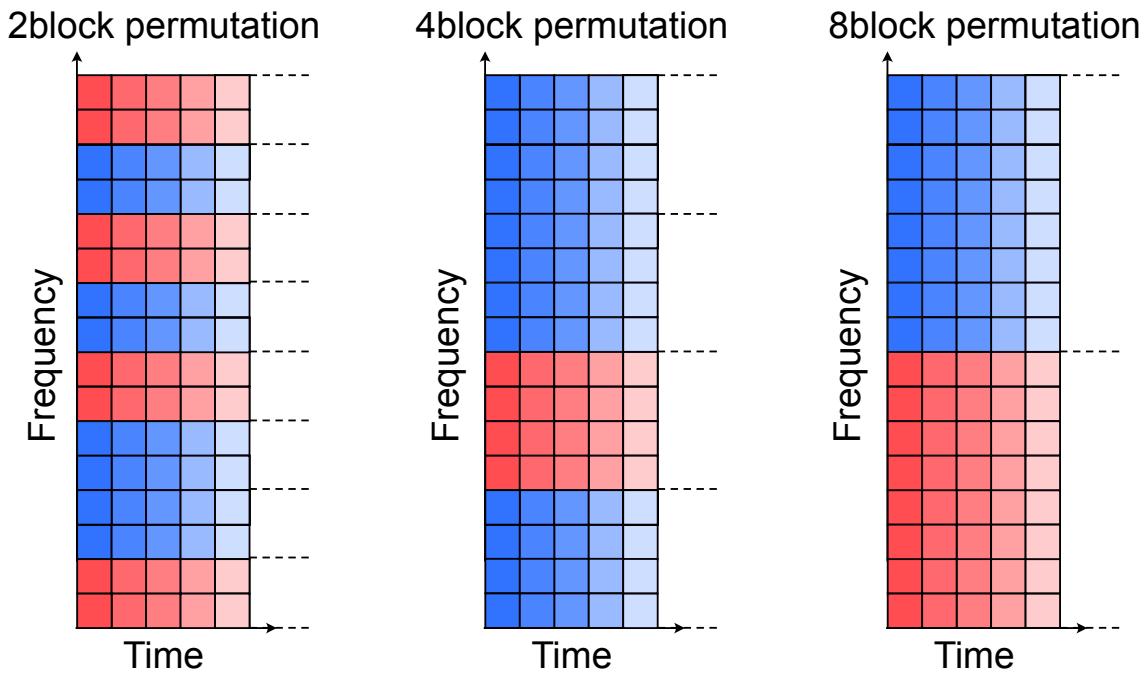


Fig. 4.4: Example of block permutations used in the experiment.

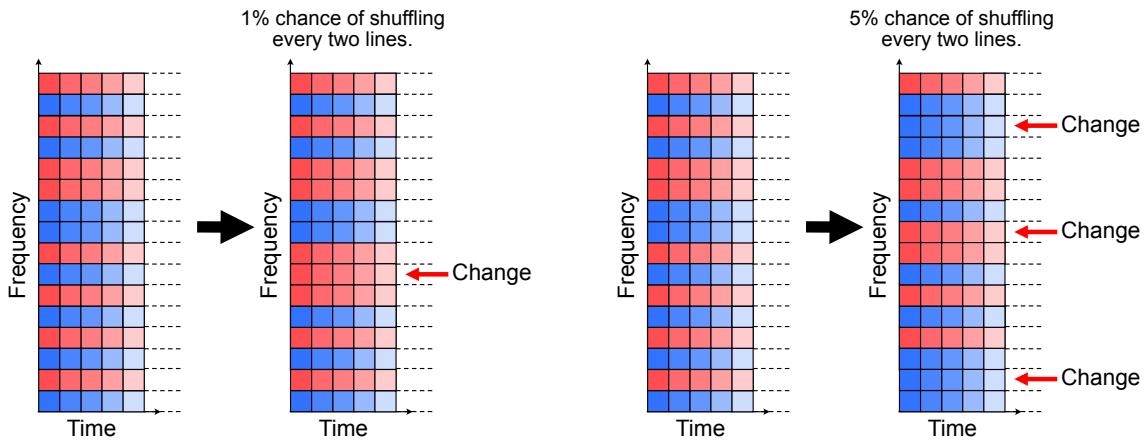


Fig. 4.5: Example of permuting in two blocks according to proportions.

波数成分をシャッフルさせることで作成した。DNN の最適化法には Adam[23] を用い、ハイパーパラメータはそれぞれ  $\varepsilon = 1.0 \times 10^{-8}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  及び学習率  $\eta = 0.001$  とした。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, 学習に用いるシャッフルパターンを 300 として誤差逆伝搬学習を行った。主観評価として、各周波数成分において正しく並び替えを行う割合、即ち検証データに対する正答率を用いる。

Table 4.1: Speech sources obtained from SiSEC2011

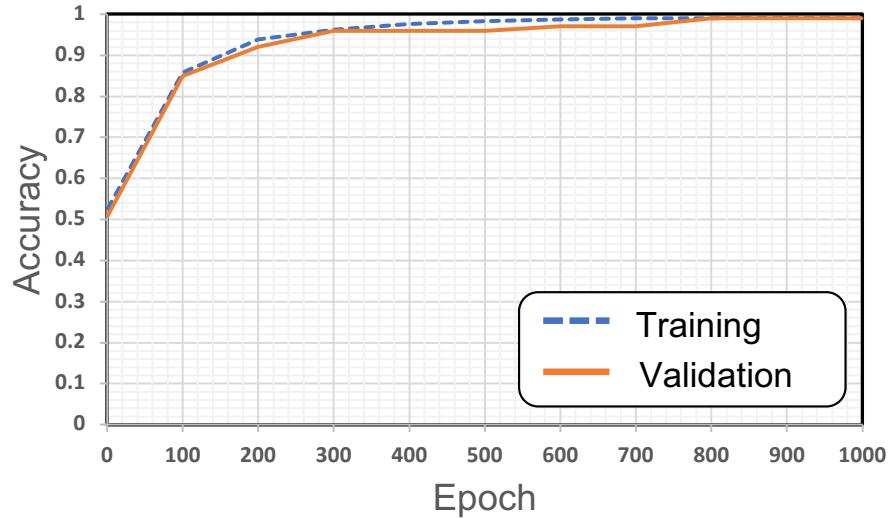
Signal	Language	Data name	Length [s]
Speech	English	dev3_female4_src_2	10.0
Speech	English	dev2_male4_src_2	10.0
Piano		dev2_nodrums_liverec_250ms_src_3	10.0
Drum		dev2_wdrums_liverec_250ms_src_3	10.0

#### 4.2.2 実際の音響信号を用いた実験の条件

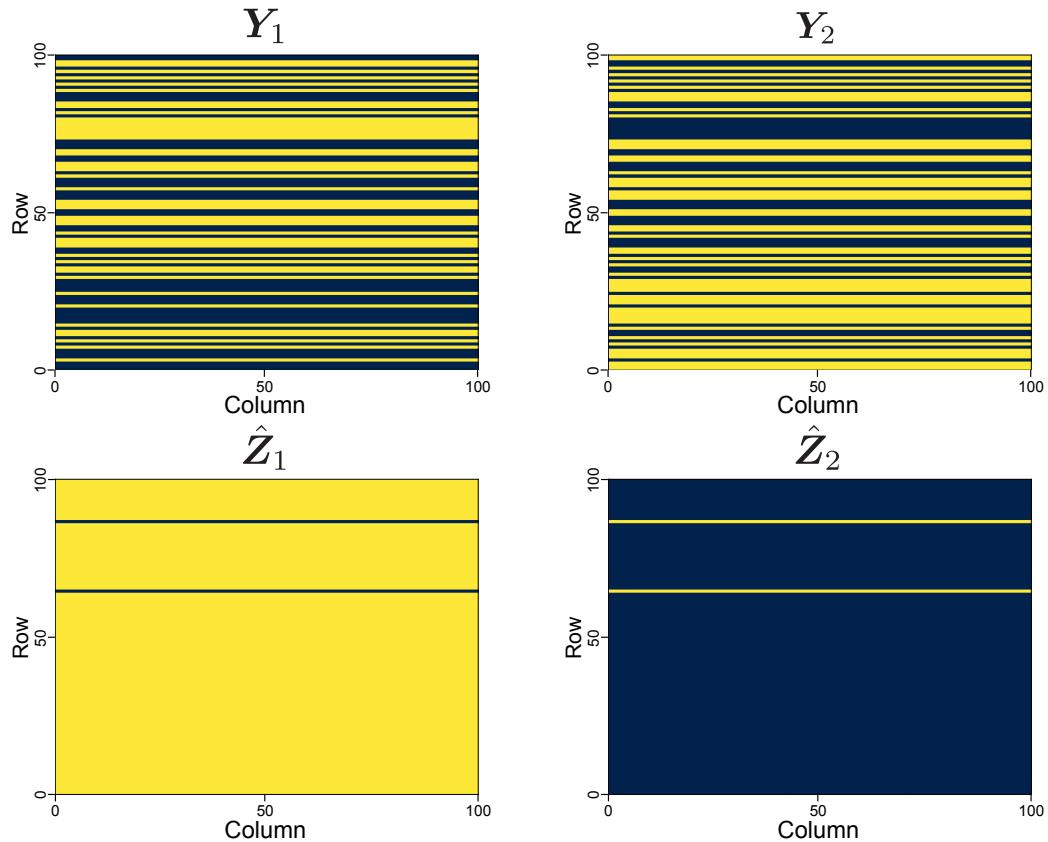
実際の音響信号に対して提案手法がどの程度適用できるかを調べるために、Table 4.1 に示すように SiSEC2011 [22] の英語の音声信号（男性 1 名及び女性 1 名）2 種類と楽器音（ピアノとドラム）2 種類を使用した。音響信号に対する STFT は、fft サイズ 2048 ms, シフトサイズ 1024 ms に設定した。音声信号と楽器音の分離に対しては、ブロックパーミュテーションを解くことを想定し、16 行ごとに周波数成分をシャッフルした場合の実験を行った。最適化法やハイパーパラメータについては、4.2.1 項の条件と同じである。

### 4.3 人工データに対する実験結果

Fig. 4.6–Fig. 4.8 には、全ての成分が 0 と 1 の行列、25 列毎に 0 と 1 の値が入れ替わる行列、1 列毎に 0 と 1 の値が入れ替わる行列の周波数成分に対して各周波数毎にシャッフルを行った時の結果を示す。この結果から、各周波数成分毎にシャッフルした場合、Fig. 4.1 や Fig. 4.2 に示すような、比較的簡易的な行列に対しては、それぞれ正答率が 100% に近い値となっていることがわかる。Fig. 4.1 と Fig. 4.2 の推定分離信号  $\hat{Z}_1$ ,  $\hat{Z}_2$  は少しの間違いは含んでいるものの、おおよそ並び替えができることがわかる。しかし、Fig. 4.3 に示すような、1 列毎に 0 と 1 の値が入れ替わる行列に対して各周波数成分毎にシャッフルを行った場合は、Fig. 4.8 に示すように正答率が 54% 程度となった。 $\hat{Z}_1$ ,  $\hat{Z}_2$  を見ても並び替えができていないことがわかる。即ち、提案手法において各行ごとにシャッフルを行なった行列に対してはパーミュテーション問題を解決することが難しいが、ブロック単位でのパーミュテーション問題は容易に解けると言える。Fig. 4.9a と Fig. 4.10a は、Fig. 4.5 のように 5% の割合で 2 行ごとにシャッフルしそれ以外は 1 行ごとにシャッフルした行列と、1% の割合で 2 行ごとシャッフルしそれ以外は 1 行ごとにシャッフルした行列を用いて実験を行った結果を示す。Fig. 4.9 と Fig. 4.10 より、5% の割合で 2 行ごとにシャッフルしそれ以外は 1 行ごとにシャッフルした場合の正答率は 93% 程度となったが、1% の割合で 2 行ごとシャッフルしそれ以外は 1 行ごとにシャッフルした場合の正答率は 60% 程度となった。このことより、DNN は少しでもブロック単位でシャッフルが行われていると学習が容易となることがわかる。

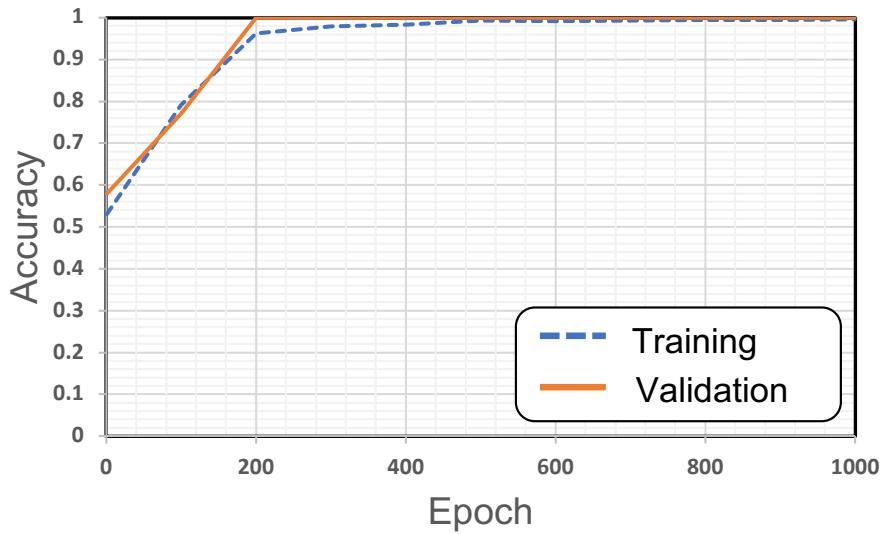


(a) Percentage of correct answers for training and validation data.

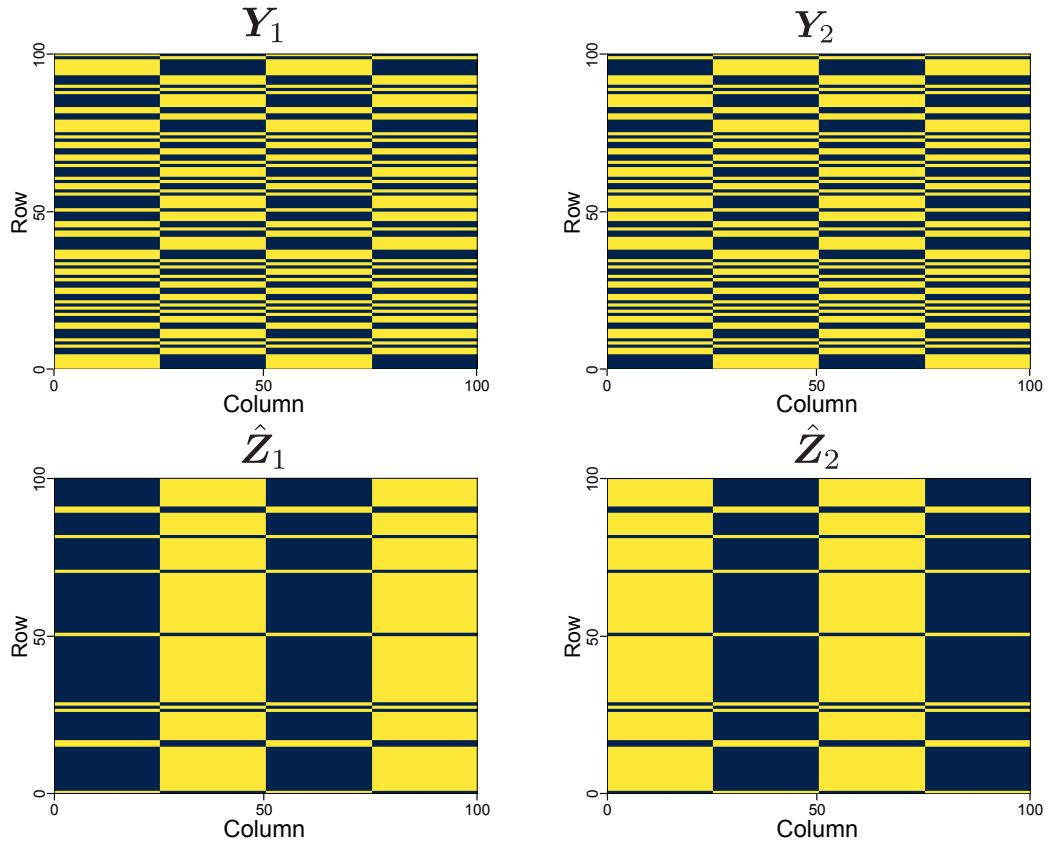


(b) Spectrogram of estimated signal and predictive separation signal

Fig. 4.6: Experimental results using matrix in Fig 4.1 (random shuffle for each frequency).

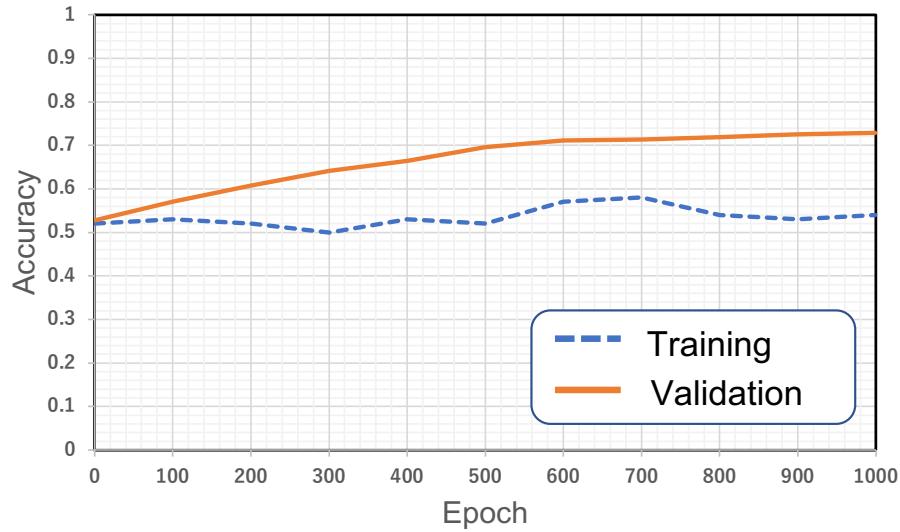


(a) Percentage of correct answers for training and validation data.

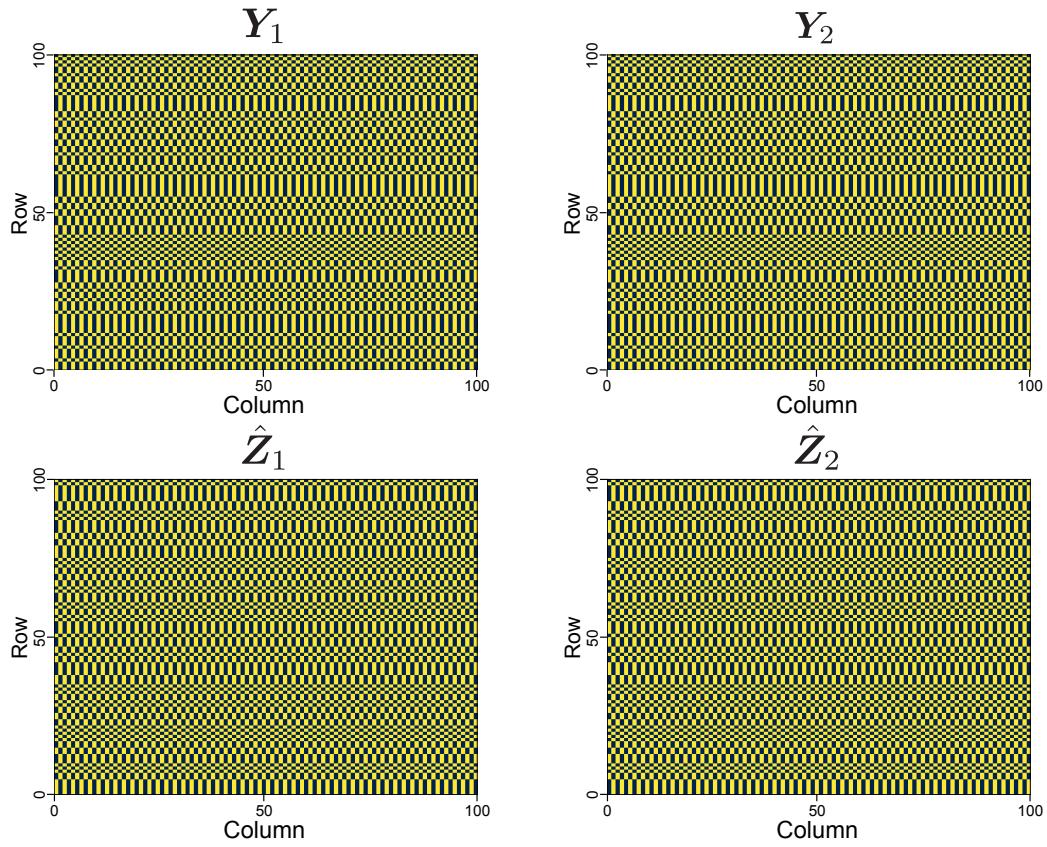


(b) Spectrogram of estimated signal and predictive separation signal

Fig. 4.7: Experimental results using the matrix in Fig 4.2 (random shuffle for each frequency).

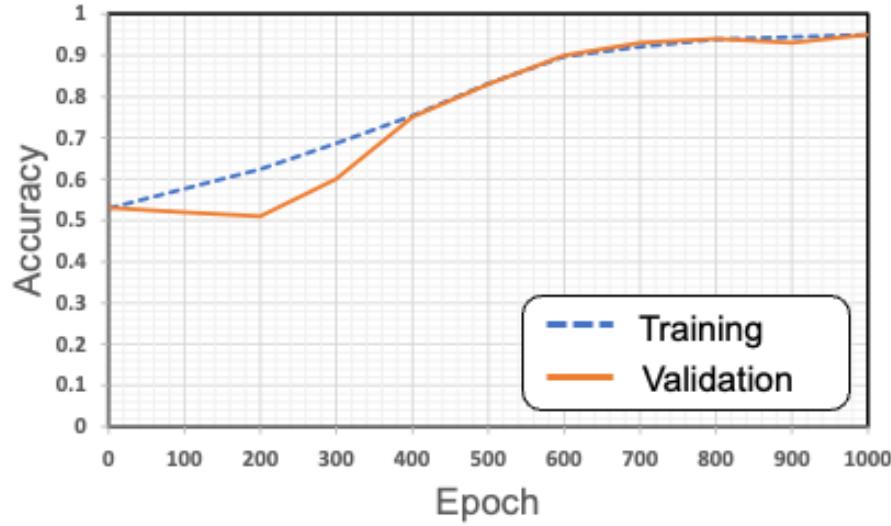


(a) Percentage of correct answers for training and validation data.

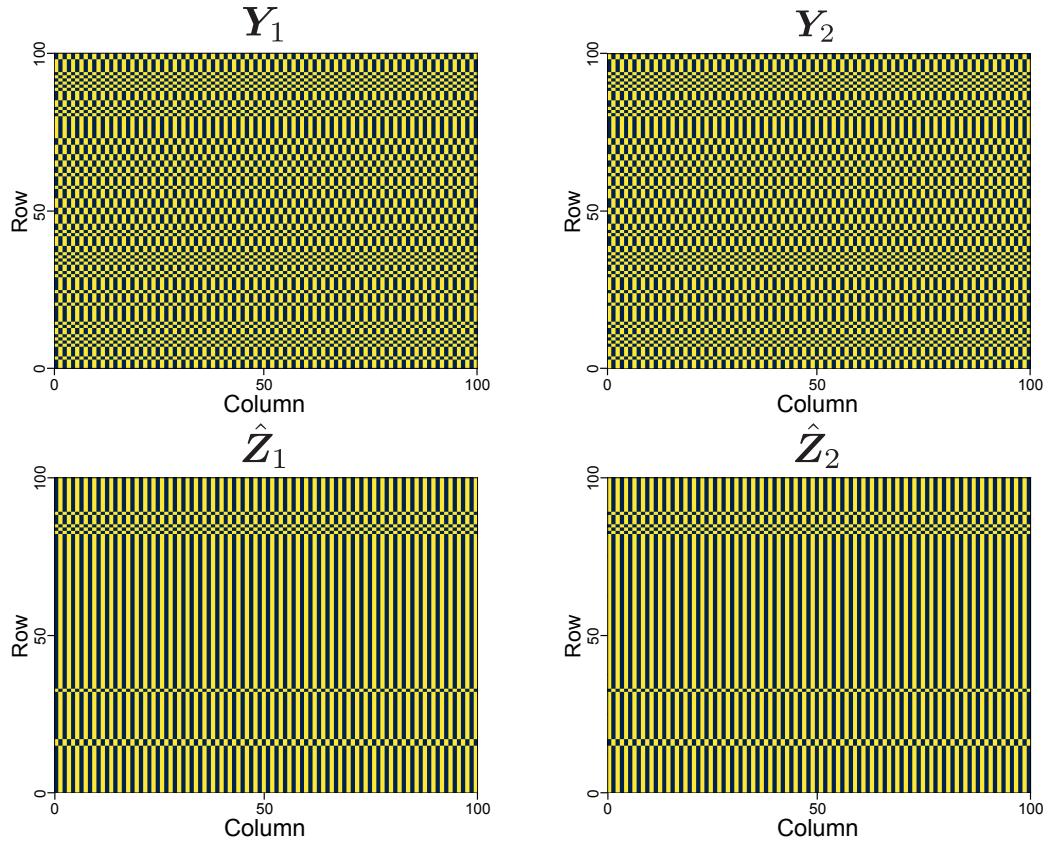


(b) Spectrogram of estimated signal and predictive separation signal

Fig. 4.8: Experimental results using the matrix in Fig 4.3 (random shuffle for each frequency).

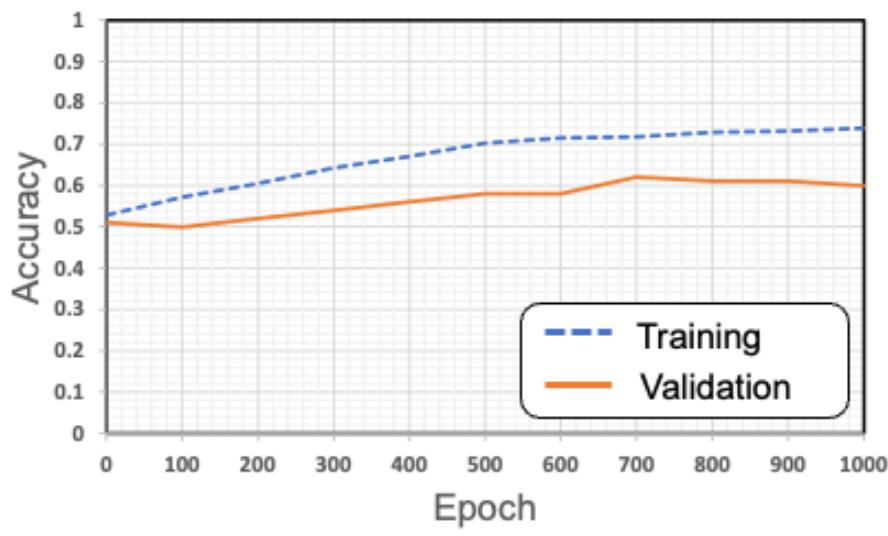


(a) Percentage of correct answers for training and validation data.

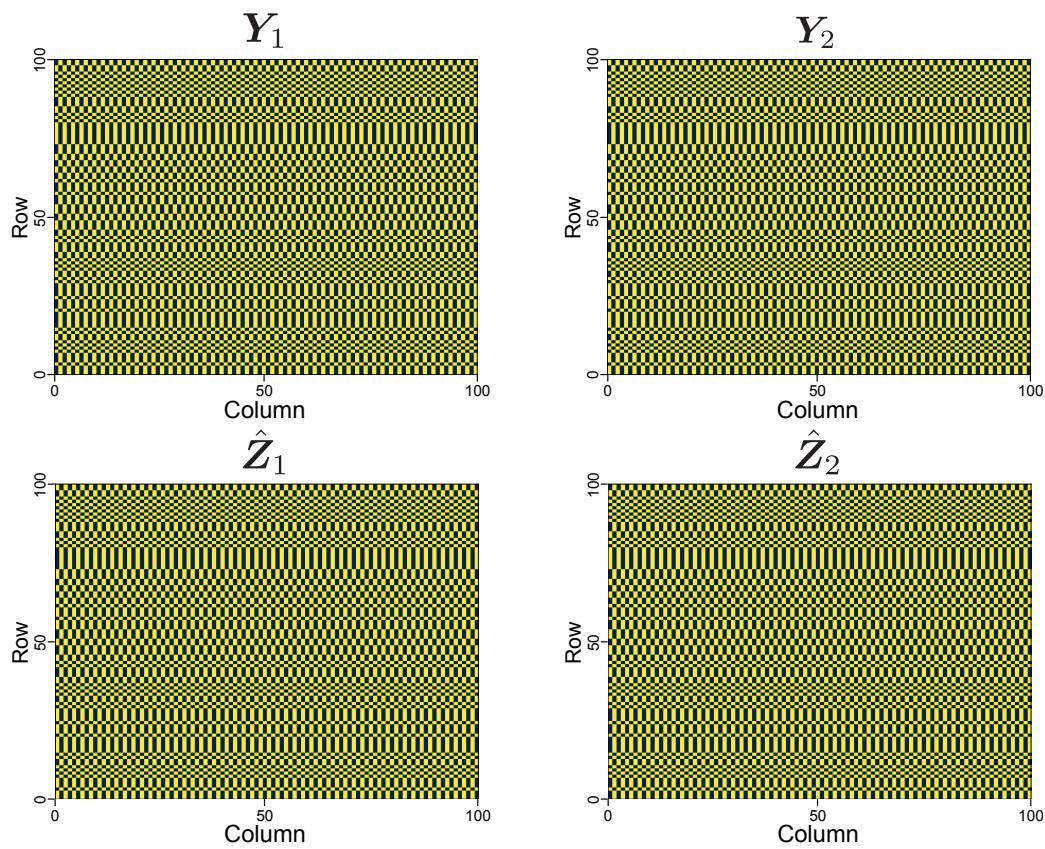


(b) Spectrogram of estimated signal and predictive separation signal

Fig. 4.9: Experimental results using the matrix in Fig 4.3 (random shuffle for each frequency).



(a) Percentage of correct answers for training and validation data.



(b) Spectrogram of estimated signal and predictive separation signal

Fig. 4.10: Experimental results using the matrix in Fig 4.3 (random shuffle for each frequency).

## 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、FDICA を適用した後のパーミュテーション行列を模倣した人工データと実際の音声データを用いて、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。実際の音声データに対しても、ブロック単位でシャッフルが行われていると 80% を超える正答率になることを示した。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、DNN を用いた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、ミニ振幅スペクトログラム成分を用いた。テストデータに対しては DNN の入力となるミニ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。また、誤差逆伝播の際に、スペクトログラム同士で平均二乗誤差を行い DNN のモデルを最適化した。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事を示した。

最後に今後の展望を述べる。本論文では、DNN を用いた 3 音源以上にも対応できる新しい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。ただ、実際に 3 音源以上の実験は行っていないことに加えて、実行時間や計算量等はあまり考慮されていない。今回行った 2 音源での実験の拡張版として、今後は 3 音源以上に対する実験も行っていきたい。また、リアルタイムでの音源分離に適用する場合は、DNN モデル及び損失取得時の計算アルゴリズムを改良する必要がある。

## 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ー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した。

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、研究の進め方につい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ー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 Transactions on Signal and Information Processing*, 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. 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.
- [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. ASLP*, 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. SAP*, 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. ASLP*, 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. WASPAA*, 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. ASLP*, 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. EUSIPCO*, 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*, pp. 781–787, 2020.
- [15] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” *Proc. ICA*, 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. ICASSP*, pp. 241-245, 2017.
- [20] E. Vincent, R. Gribonval, and C. F, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [21] V. Nair, and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” *Proc. ICML*, 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. LVA/ICA*, pp. 414–422, 2012.
- [23] D. P. kingma, and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv*, pp. 1412–6980, 2014.