香川高専

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
パーミュテー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	雛元 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 is useful in situations where multiple speech needs to be separated into the individual speech sources. It can also be used to separate a target speech signal and the background noise. One of the popular source separation methods is frequency-domain independent conponent 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 (DNNs) 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. 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: frequency-domain independent conponent 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 conponent 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: DNN）を用いたパーミュテーション問題の解決法が提案されたが，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，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

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，複数人が会話を行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単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 $\boldsymbol{A}$（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 $\boldsymbol{A}$ の逆系である分離系 $\boldsymbol{W}$ を推定し，観測信号 $\boldsymbol{X}$ 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，優決定条件下の 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

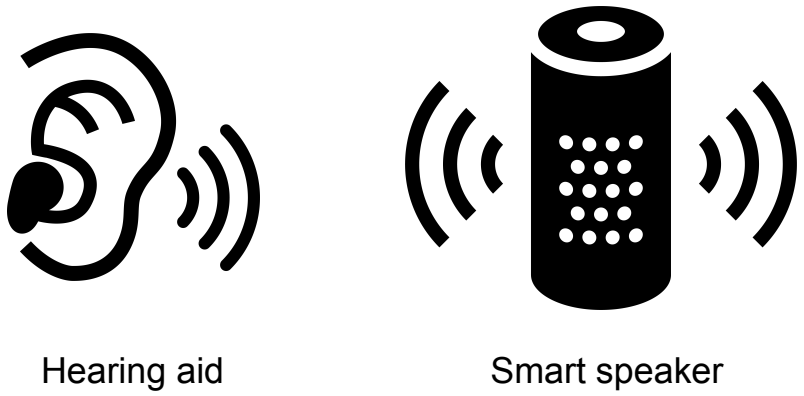

Fig. 1.1. Examples of application using speech source separation.

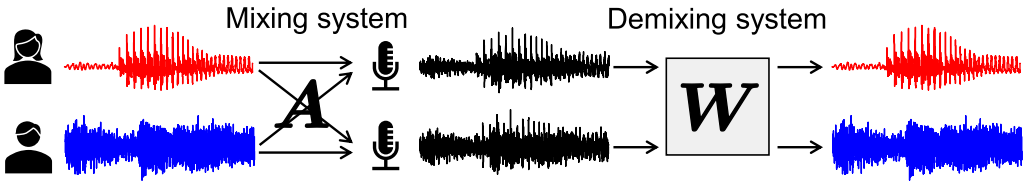

Fig. 1.2. Overview of BSS.

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]，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 [5]，及びその両者を組み合わせた手法 [6] 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）[7, 8] 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）[9] 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）[10, 11] 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．

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 [12] が提案されてきた．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~~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を Fig. 1.3 に示す．本論文では，Fig. 1.4 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．

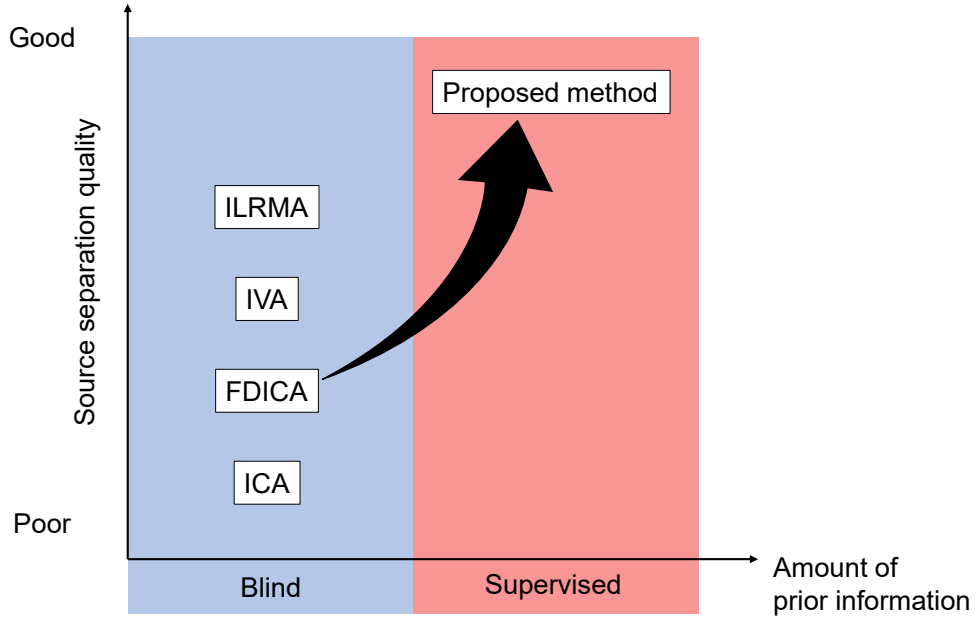

Fig. 1.3. Scope of this thesis.

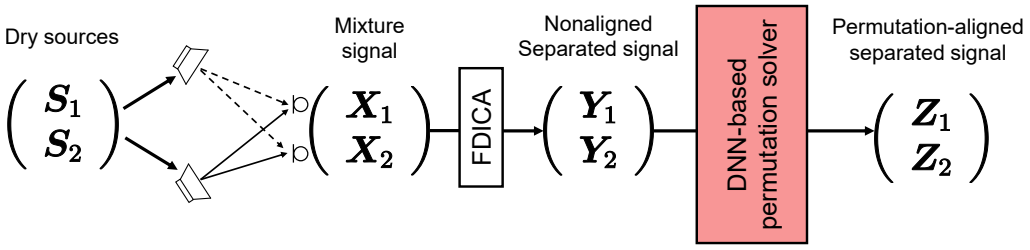

Fig. 1.4. Objective of this thesis.

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され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，$l = 1, 2, \cdots, L$ 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．

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

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，$n = 1, 2, \cdots, N$，及び $m = 1, 2, \cdots, M$ 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．

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{2.2}$$

この行列 $\boldsymbol{A}$ は混合行列と呼ばれる．観測信号ベクトル $\boldsymbol{x}(l) = (x_1(l), x_2(l))^T [x_1(l), x_2(l)]^{\mathrm{T}}$ 信号源ベクトル $\boldsymbol{s}(l) = (s_1(l), s_2(l))^T [s_1(l), s_2(l)]^{\mathrm{T}}$ 及び混合行列 $\boldsymbol{A}$ 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．

$$\boldsymbol{x}(l) = \boldsymbol{A}\boldsymbol{s}(l) \tag{2.3}$$

ここで，$\cdot^{\mathrm{T}}$ はベクトルや行列の転置を表す．分離信号を $\boldsymbol{y}(l) = (y_1(l), y_2(l))^T [y_1(l), y_2(l)]^{\mathrm{T}}$，分離行列を $\boldsymbol{W}$ 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．

$$\boldsymbol{y}(l) = \boldsymbol{W}\boldsymbol{x}(l) \tag{2.4}$$

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

$$\begin{aligned} \boldsymbol{y}(l) &= \boldsymbol{W}\boldsymbol{x}(l) && (2.5) \\ &= \boldsymbol{A}^{-1}\boldsymbol{x}(l) && (2.6) \\ &= \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{s}(l) && (2.7) \\ &= \boldsymbol{s}(l) && (2.8) \end{aligned}$$

このように，混合行列 $\boldsymbol{A}$ の逆行列である分離行列 $\boldsymbol{W}$ 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 $\boldsymbol{A}$ 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 $\boldsymbol{W}$ 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の重要な概念として，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．

$$p(s_1, s_2) = p(s_1)p(s_2) \tag{2.9}$$

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．

$$p(y_1, y_2) = p(y_1)p(y_2) \tag{2.10}$$

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 $\boldsymbol{W}$ 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．

$$\arg\min_{\boldsymbol{W}} \mathfrak{I}(\boldsymbol{W}) \tag{2.11}$$

$$\mathfrak{I}(\boldsymbol{W}) = \mathfrak{D}_{KL\mathrm{KL}}[p(y_1, y_2)||p(y_1)p(y_2)] \tag{2.12}$$

ここで，$\mathfrak{D}_{KL\mathrm{KL}}[p(s)||q(s)]$ はカルバックライブラ・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．

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

また，分離フィルタ行列 $\boldsymbol{W}$ で観測信号を線形変換する前（$\boldsymbol{x}$）と後（$\boldsymbol{y}$）の確率変数を考えたとき，それぞれの同時分布 $p(\boldsymbol{y}) = p(y1, y2y_1, y_2)$ と $p(\boldsymbol{x}) = p(x1, x2x_1, x_2)$ の間には，次式が成立する．

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

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 $\mathfrak{I}(\boldsymbol{W})$ は以下のように書ける．

$$\begin{aligned}\mathfrak{I}(\boldsymbol{W}) = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1, x_2)\log\, p(x_1, x_2)dx_1dx_2 - \log\,|\det\,\boldsymbol{W}| \\ &- \int_{-\infty}^{\infty} p(y_1)\log\, p(y_1)dy_1 - \int_{-\infty}^{\infty} p(y_2)\log\, p(y_2)dy_2\end{aligned} \tag{2.15}$$

ICA では, 式 (2.15) が最小化される分離行列 $\boldsymbol{W}$ を求めることでを $\boldsymbol{W}$ について最小化することで, 信号源を分離する．

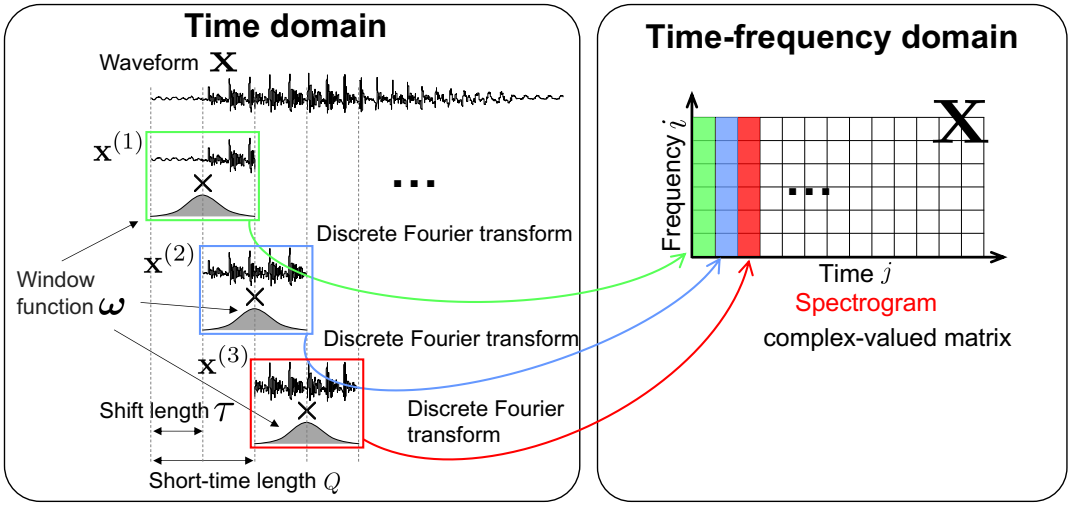

Fig. 2.1. Mechanism of STFT. Each of windowed short-time signals are transformed to frequency domain by discrete Fourier transform.

2.2.3　aaa

2.3　STFT

STFT は Fig. 2.1 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．

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

STFT の分析窓関数の長さ及びシフト長をそれぞれ Q 及び τ としたとき，時間領域の信号 ~~$z(l)$~~$\mathbf{x}$ 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．

$$\cancel{z}\mathbf{x}^{(j)} = [\cancel{z}\mathrm{x}((j-1)\tau+1), \cancel{z}\mathrm{x}((j-1)\tau+2), \cdots, \cancel{z}\mathrm{x}((j-1)\tau+Q)]^{\cancel{T}\mathrm{T}} \tag{2.17}$$

$$= \left[\cancel{z}\mathrm{x}^{(j)}(1), \cancel{z}\mathrm{x}^{(j)}(2), \cdots, \cancel{z}\mathrm{x}^{(j)}(q), \cdots, \cancel{z}\mathrm{x}^{(j)}(Q)\right]^{\cancel{T}\mathrm{T}} \in \mathbb{R}^{Q} \tag{2.18}$$

ここで，$j = 1, 2, \cdots, J$ 及び $q = 1, 2, \cdots, Q$ 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．

$$J = \frac{L}{\tau} \tag{2.19}$$

~~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．

$$\cancel{Z}\mathbf{X} = \mathrm{STFT}_{\boldsymbol{\omega}}(\cancel{z}\mathbf{x}) \in \mathbb{C}^{I \times J} \tag{2.20}$$

~~スペクトログラム $\boldsymbol{Z}$ の~~ここで，$\mathbf{X}$ は（複素）スペクトログラムと呼ばれ，Fig. 2.1 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．

$$~~\mathrm{x}_{ij}~~ = \sum_{q=1}^{Q} \omega(q)~~\mathrm{x}^{(j)}~~(q)\exp\left\{\frac{-\iota 2\pi(q-1)(i-1)}{F}\right\} \tag{2.21}$$

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~~，$i = 1, 2, \cdots, I$ は周波数ビンのインデクス~~を~~，ι は虚数単位~~を~~，~~$\boldsymbol{\omega}$~~ $= [\omega(1), \omega(2), \ldots, \omega(Q)]^{\mathrm{T}} \in \mathbb{R}^Q$ 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解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~~$\boldsymbol{Z}$ で表すことができる．~~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

~~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．

$$\boldsymbol{s}_{ij} = \left[s_{ij~~,1~~}, s_{ij~~,2~~}, \cdots, s_{ij~~,n~~}, \cdots, s_{ij~~,N~~}\right]^{\mathrm{T}} \in \mathbb{C}^N \tag{2.22}$$

$$\boldsymbol{x}_{ij} = \left[x_{ij~~,1~~}, x_{ij~~,2~~}, \cdots, x_{ij~~,m~~}, \cdots, x_{ij~~,M~~}\right]^{\mathrm{T}} \in \mathbb{C}^M \tag{2.23}$$

$$\boldsymbol{z}_{ij} = \left[z_{ij~~,1~~}, z_{ij~~,2~~}, \cdots, z_{ij~~,n~~}, \cdots, z_{ij~~,N~~}\right]^{\mathrm{T}} \in \mathbb{C}^N \tag{2.24}$$

~~と表す．~~

~~また，複素スペクトログラム行列 $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$, $\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$ $\boldsymbol{Z}_n \in \mathbb{C}^{I\times J}$ 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 $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$，及び $\boldsymbol{Z}_n \in \mathbb{C}^{I\times J}$ と定義する．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を推定することで BSS を実現する手

法である．しかし，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，線形時不変な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．

$$\boldsymbol{x}(l) = \sum_{n} \sum_{l'=0}^{R-1} \tilde{\boldsymbol{a}}_n(l') \boldsymbol{s}_n(l - l') \tag{2.25}$$

ここで，$\tilde{\boldsymbol{a}}_n(l)$ 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 までのインパルス応答をまとめたもの）である．これを分離するためには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，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) 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．

FDICA では，周波数毎の時不変な混合行列 $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1}\ \boldsymbol{a}_{i,2}\ \cdots\ \boldsymbol{a}_{i,n}\ \cdots, \boldsymbol{a}_{i,N}) \in \mathbb{C}^{M \times N}$ を定義し，混合信号が次式で表現できると仮定する．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{2.26}$$

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する．以後，決定的な系（$M = N$）を仮定すると，混合行列 $\boldsymbol{A}_i$ が正則であれば，分離行列 $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = (\boldsymbol{w}_{i,1}\ \boldsymbol{w}_{i,2}\ \ldots\ \boldsymbol{w}_{i,n}\ \ldots\ \boldsymbol{w}_{i,N})^{\mathrm{H}}$ を用いて，分離信号を次式で表せる．

$$\boldsymbol{z}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{2.27}$$

ここで，$.^{\mathrm{H}}$ 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 $\boldsymbol{w}_{i,n} \in \mathbb{C}^M$ は，周波数 i において，観測信号から n 番目のみの音源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 $\boldsymbol{x}_{ij}$ の各周波数ビンに対しそれぞれ独立に ICA を適用することで，周波数毎の分離行列 $\boldsymbol{W}_i$ を全周波数にわたって推定することで音源分離を行う．

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，分離音源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 $\hat{\boldsymbol{W}}_i$ とすると，次式のような不定性が残る．

$$\hat{\boldsymbol{W}}_i = \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{W}_i \tag{2.28}$$

ここで，$\boldsymbol{P}_i \in \{0, 1\}^{N \times N}$ は分離行列 $\boldsymbol{W}_i$ の行ベクトル $\boldsymbol{w}_{i,n}$ 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．$\boldsymbol{D}_i \in \mathbb{R}^{N \times N}$ は，$\boldsymbol{w}_{i,n}$ 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

$$\boldsymbol{y}_{ij} = \hat{\boldsymbol{W}}_i \boldsymbol{x}_{ij} \tag{2.29}$$

$$= [y_{ij,1}, y_{ij,2}, \cdots, y_{ij,n}, \cdots, y_{ij,N}]^{\mathrm{T}} \ \in \mathbb{C}^N \tag{2.30}$$

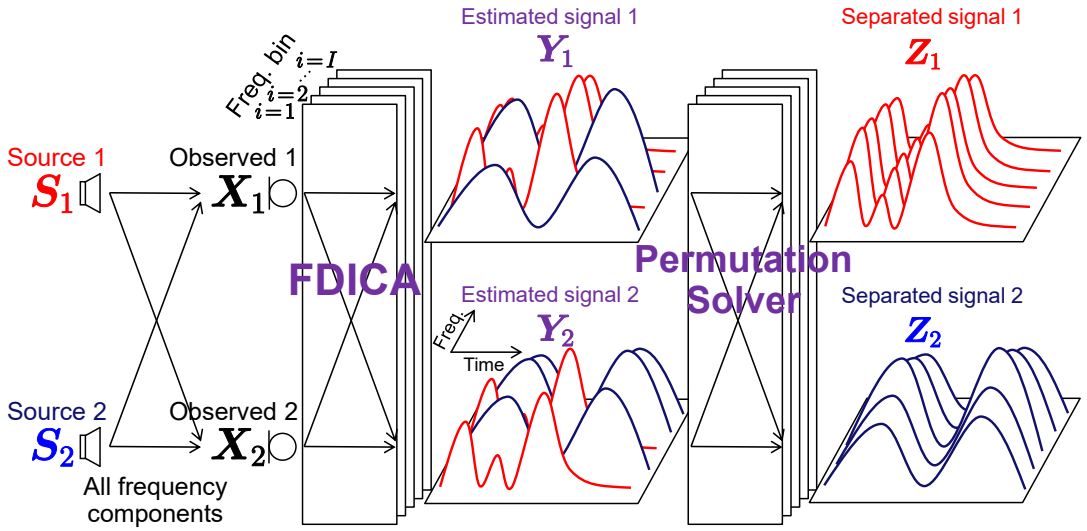

Fig. 2.2. Permutation problem in FDICA, where $N = M = 2$.

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，$\boldsymbol{D}_i$ 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 [14] で復元可能である．一方で，$\boldsymbol{P}_i$ によって生じる分離信号の順番の任意性（パーミュテーション）を純粋に復元することは，組み合わせ爆発が発生するため容易ではない．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. 2.2 に示す．ここで，FDICA で推定される分離信号 $\boldsymbol{y}_{ij}$ の音源毎の複素スペクトログラム行列を $\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$ で表している．FDICA 直後の $\boldsymbol{Y}_n$ 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，異なるグループの分離信号が 1 つの時間周波数構造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．

$$\boldsymbol{z}_{ij} = \boldsymbol{P}_i^{-1}\boldsymbol{D}_i^{-1}\boldsymbol{y}_{ij} \tag{2.31}$$

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されてきた．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] がある．これは，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，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 [5] 及び両者を組み合わせたパーミュテーション解決法も提案されている．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

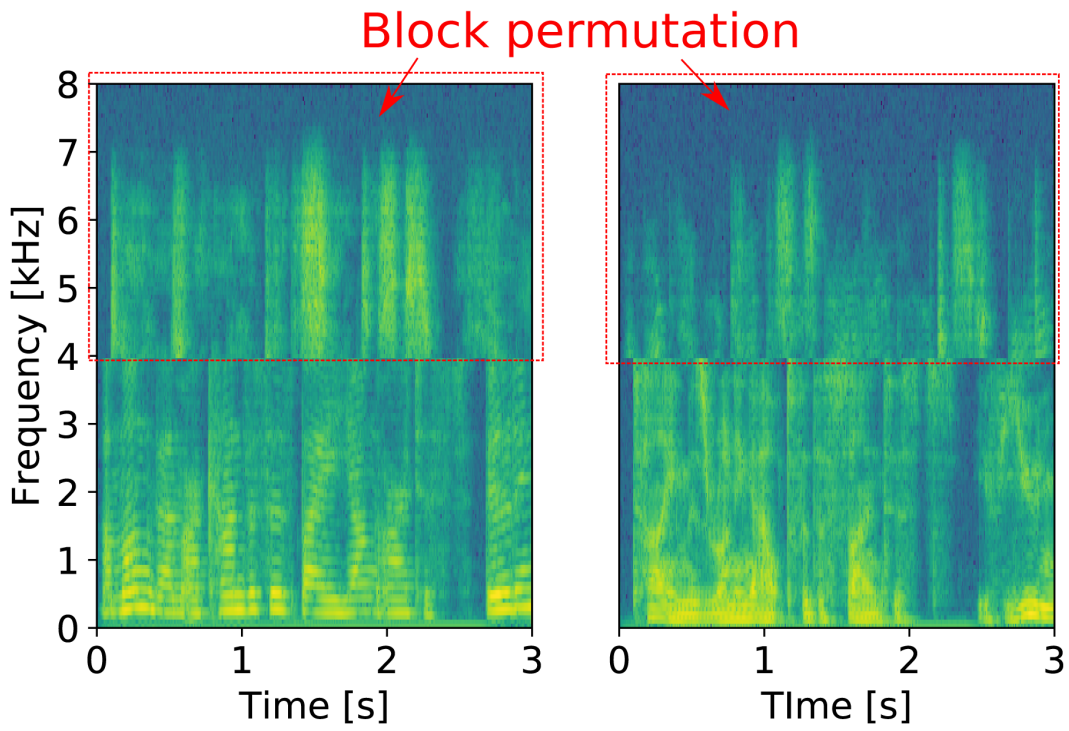

Fig. 2.3. Example of block permutation problem.

いても完璧にパーミュテーション問題を解くことは非常に難しく，とくに複数音声の混合信号における高精度なパーミュテーション問題の解決はいまだできていない．

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 [7, 8] 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．パーミュテーション問題を回避することが期待できる．また，NMF [9] と IVA を組み合わせた BSS である ILRMA [10, 11] 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．

しかし，音声と音声の混合信号の様な分離タスクの場合，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．また，IVA や ILRMA においても，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）[15] が報告されている．Fig. 2.3 にブ

ロックパーミュテーション問題の様子を示す．Fig. 2.3 では，4000hz 以上の周波数帯がまとめて反転していることが分かる．そのため，依然としてパーミュテーション問題の解決が不十分であることが分かる．

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 $\boldsymbol{X}_n$ 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 $\boldsymbol{Y}_n$ が得られる．これらのパワースペクトログラム $|\boldsymbol{Y}_n|^{\cdot 2}$ 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，各サブバンドベクトル内の 2 値は（「0」及び「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，参照周波数の変化を共に，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．

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．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，FDICA に伴い生じるパーミュテーション問題と従来の深層パーミュテーション解決法について説明した．本章では，組み合せ爆発を起こすことのない，DNN 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．3.2 節では，IVA や ILRMA のようなブラインド（教師無し）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．3.3 節及び 3.4 節で，提案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構造を説明する．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ラベルの取得方法を説明する．3.7 節で本章のまとめを述べる．

3.2 動機

文献 [13] では, BSS の STFT における最適な窓長を実験的に検討している．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．縦軸は信号対歪み比（source-to-distortion ratio: SDR）[16] の改善量であり，これは即ち分離性能を表している．この結果より，IVA 及び ILRMA では，残響状態 $T_{60} = 470$ ms の条件では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 $\boldsymbol{s}_{ij}$ 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であっても，$\hat{\boldsymbol{W}}_i$ は FDICA で正確に推定でき，$\boldsymbol{P}_i^{-1}$ の推定のみ失敗していることを示している．また，従来の深層パーミュテーション解決法では，全周波数帯域中の局所的な狭帯域のおけるパーミュテー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，ある参照周波数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 DNN に基づくデータ駆動型（教師あ

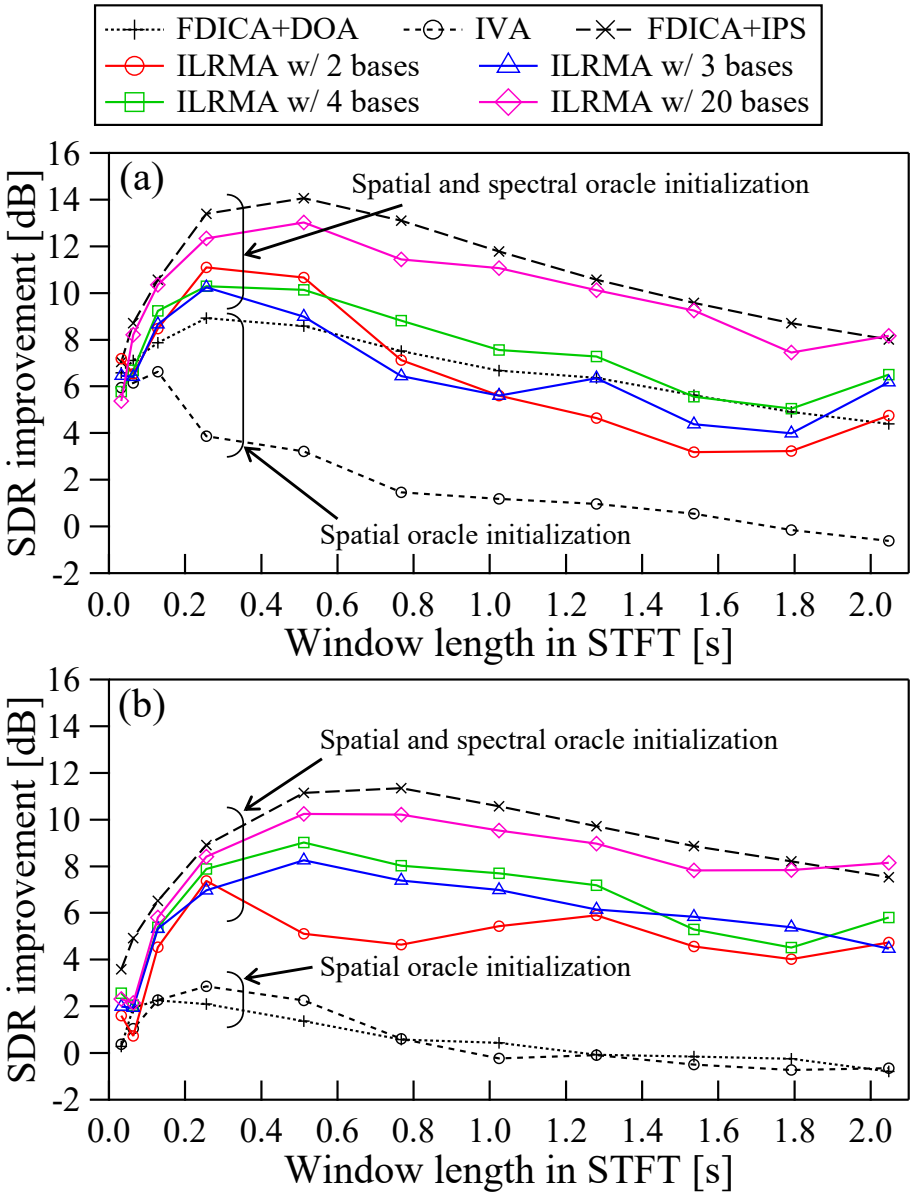

Fig. 3.1. Average source separation results for speech signals using random initialization: (a) E2A (T_{60} =300 ms) and (b) JR2 (T_{60}=470 ms) impulse responses [13].

り）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に，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を用いてパーミュテーション問題の解決を考える．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．提案するパーミュテーション解決法の概要は以下の通りである．

- 分離信号 $\boldsymbol{Y}_1$ 及び $\boldsymbol{Y}_2$ 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
- $\boldsymbol{Y}_1$ 及び $\boldsymbol{Y}_2$ 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）$\widehat{\boldsymbol{L}}$ は，予測値の時間方向への多数決結果から決定される

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，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 $\boldsymbol{Z}_n$ 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 $\boldsymbol{X}_n$ に FDICA を適用すると，パーミュテーション問題が生じた分離信号 $\boldsymbol{Y}_n$ 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．

$$\widehat{\boldsymbol{Y}}_1 = \frac{|\boldsymbol{Y}_1|^2}{|\boldsymbol{Y}_1|^2 + |\boldsymbol{Y}_2|^2} \tag{3.1}$$

$$\widehat{\boldsymbol{Y}}_2 = \frac{|\boldsymbol{Y}_2|^2}{|\boldsymbol{Y}_1|^2 + |\boldsymbol{Y}_2|^2} \tag{3.2}$$

ここで，DNN の入力に用いる値をそれぞれ $\widehat{\boldsymbol{Y}}_1 \in \mathbb{R}_{\geq 0}^{I \times J}$，$\widehat{\boldsymbol{Y}}_2 \in \mathbb{R}_{\geq 0}^{I \times J}$ とする．この時，$i = 1, \ldots, I$ 及び $j = \tau, \ldots, J - \tau$ 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 の範囲を限定的にしている．ここで，行列の $|\cdot|^{\cdot 2}$ 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．

$$\widehat{\boldsymbol{y}}_{1j} = [\widehat{y}_{11j}, y_{12j}, \cdots, \widehat{y}_{1Ij}]^{\mathrm{T}} \tag{3.3}$$

$$\widehat{\boldsymbol{y}}_{2j} = [\widehat{y}_{21j}, y_{22j}, \cdots, \widehat{y}_{2Ij}]^{\mathrm{T}} \tag{3.4}$$

ここで，$\widehat{y}_{1ij}$ は $\widehat{\boldsymbol{Y}}_1$ の ij 要素であり，$\widehat{y}_{2ij}$ は $\widehat{\boldsymbol{Y}}_2$ の ij 要素を表す．DNN の入力として与える情報は，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．これを $\boldsymbol{x}_j$ とおくと，次式のように構成される．また，j 近傍の時間フレームの幅は τ とする．

$$\boldsymbol{x}_j = [\widehat{\boldsymbol{y}}_{1(j-\tau)}^{\mathrm{T}}, \cdots, \widehat{\boldsymbol{y}}_{1(j+\tau)}^{\mathrm{T}}, \widehat{\boldsymbol{y}}_{2(j-\tau)}^{\mathrm{T}}, \cdots, \widehat{\boldsymbol{y}}_{2(j+\tau)}^{\mathrm{T}}]^{\mathrm{T}} \tag{3.5}$$

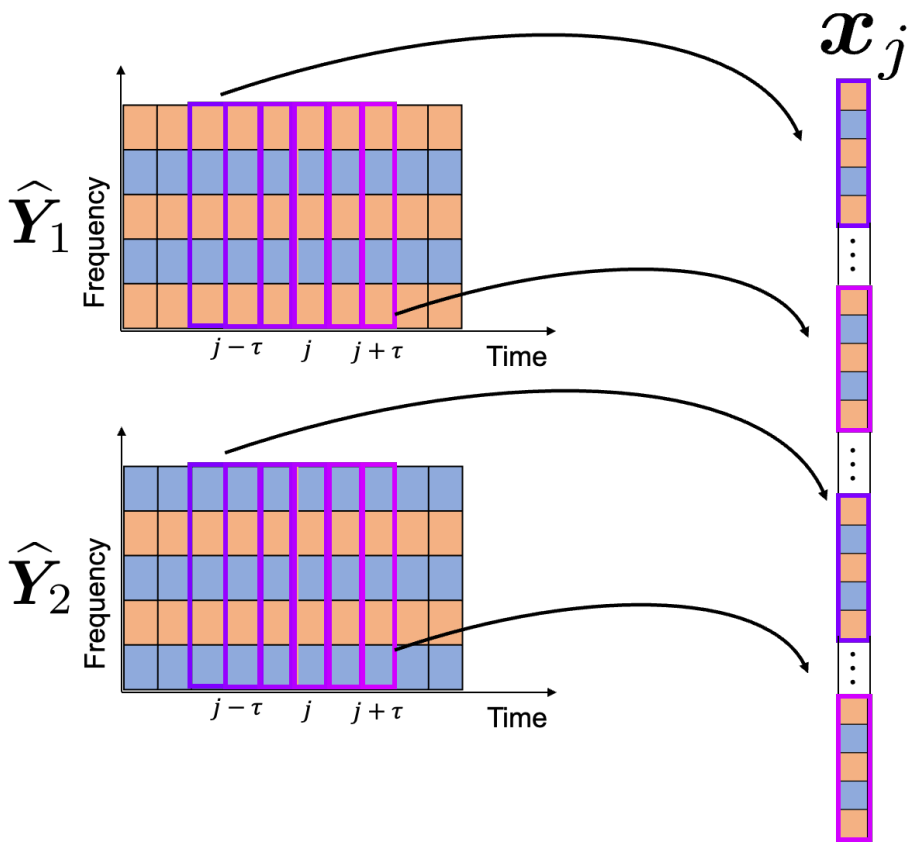

Fig. 3.2. Input vector of DNN.

Fig. 3.2 に示すように，上記の $\boldsymbol{x}_j$ が DNN の入力ベクトルとなる．

$$\boldsymbol{L} = \mathrm{DNN}(\boldsymbol{x}_j) \tag{3.6}$$

DNN が出力する予測は $\boldsymbol{L} \in \mathbb{R}_{[0,1]}^{2\times I}$ であり，確率値を示す．$\boldsymbol{L}$ 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．

$$\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2]^{\mathrm{T}} \tag{3.7}$$

$$\boldsymbol{l}_n = [\widehat{l}_{n1}, \cdots, \widehat{l}_{nI}]^{\mathrm{T}} \tag{3.8}$$

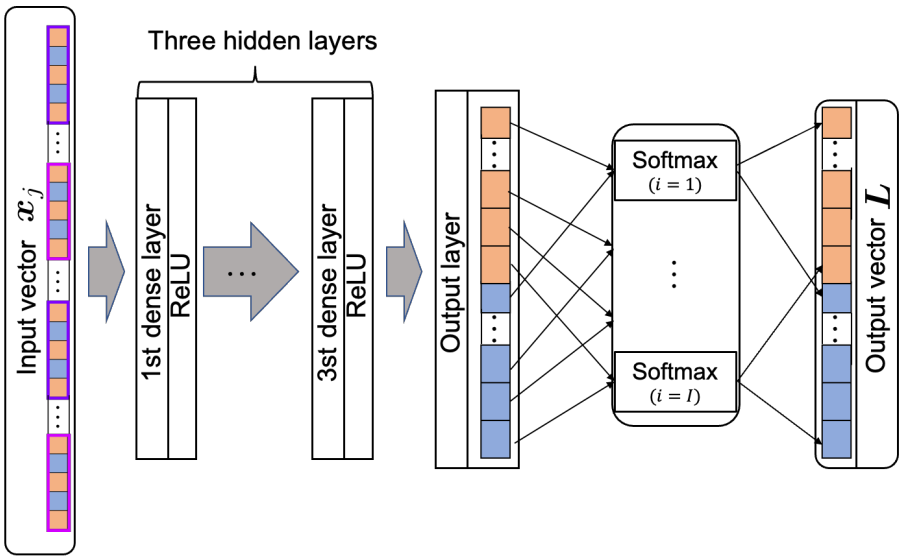

Fig. 3.3. DNN architecture.

3.4　DNN の構造

Fig. 3.3 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 層からなる全結合構成となっており，1〜5 番目の隠れ層には rectified linear unit (ReLU) [17] 関数，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　損失の取り方

誤差逆伝播を行う際の損失は，DNN が出力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，$\widehat{\boldsymbol{l}}_n$ 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 $\widetilde{\boldsymbol{Y}}_{1j} \in \mathbb{R}_{\geq 0}^{I\times(2\tau+1)}$ と $\widetilde{\boldsymbol{Y}}_{2j} \in \mathbb{R}_{\geq 0}^{I\times(2\tau+1)}$ の 2 種類を導く．

$$\widetilde{\boldsymbol{y}}_{1j} = [y_{11j}\widehat{l}_{11} + y_{21j}\widehat{l}_{21}, \cdots, y_{1Ij}\widehat{l}_{1I} + y_{2Ij}\widehat{l}_{2I}]^{\mathrm{T}} \tag{3.9}$$

$$\widetilde{\boldsymbol{y}}_{2j} = [y_{11j}\widehat{l}_{21} + y_{21j}\widehat{l}_{11}, \cdots, y_{1Ij}\widehat{l}_{2I} + y_{2Ij}\widehat{l}_{1I}]^{\mathrm{T}} \tag{3.10}$$

$$\widetilde{\boldsymbol{Y}}_{1j} = [\widetilde{\boldsymbol{y}}_{1(j-\tau)}, \cdots, \widetilde{\boldsymbol{y}}_{1(j+\tau)}] \tag{3.11}$$

$$\widetilde{\boldsymbol{Y}}_{2j} = [\widetilde{\boldsymbol{y}}_{2(j-\tau)}, \cdots, \widetilde{\boldsymbol{y}}_{2(j+\tau)}] \tag{3.12}$$

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

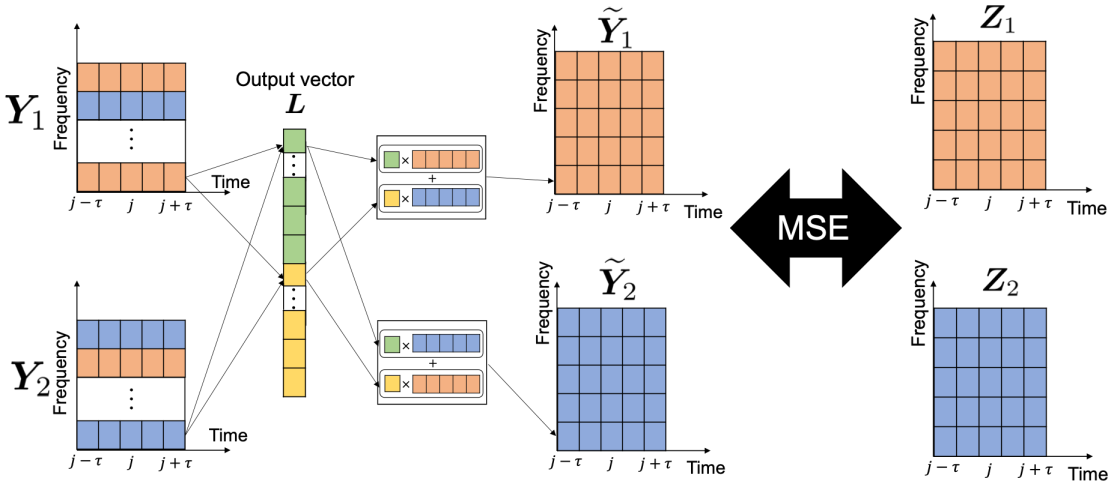

Fig. 3.4. The process of calculating losses.

$\boldsymbol{Z}_{2j} \in \mathbb{R}_{\geq 0}^{I\times(2\tau+1)}$ 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ため，次式で損失を計算する．

$$\text{Loss} = \text{MIN}\{\text{MSE}(\widetilde{\boldsymbol{Y}}_{1j}, \widetilde{\boldsymbol{Z}}_{1j}) + \text{MSE}(\widetilde{\boldsymbol{Y}}_{2j}, \widetilde{\boldsymbol{Z}}_{2j}) \ , \ \text{MSE}(\widetilde{\boldsymbol{Y}}_{1j}, \widetilde{\boldsymbol{Z}}_{2j}) + \text{MSE}(\widetilde{\boldsymbol{Y}}_{2j}, \widetilde{\boldsymbol{Z}}_{1j})\} \tag{3.13}$$

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 Fig. 3.4 に示す．最終的なラベル $\widetilde{\boldsymbol{L}}$ は，$\boldsymbol{L}$ の列成分を比較し，値が大きい方のインデクスを含んだベクトルとなる．

$$\widetilde{l}_i = \begin{cases} 1 \ (\widehat{l}_{1i} \leq \widehat{l}_{2i}) \\ 0 \ (\widehat{l}_{2i} \leq \widehat{l}_{1i}) \end{cases} \tag{3.14}$$

$$\widetilde{\boldsymbol{L}} = [\widetilde{l}_1, \cdots, \widetilde{l}_I]^{\mathrm{T}} \tag{3.15}$$

3.6 時間方向への多数決

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 $\widehat{\boldsymbol{Y}}_1$ や $\widehat{\boldsymbol{Y}}_2$ 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. 3.5 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．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 $\widehat{\boldsymbol{L}}$ を得る．この処理は，次のように示される．

$$\widehat{\boldsymbol{L}} = \text{round}\left(\frac{1}{J-13}\sum_{\text{J}} \widetilde{\boldsymbol{L}}_j \in \{0,1\}^I\right) \tag{3.16}$$

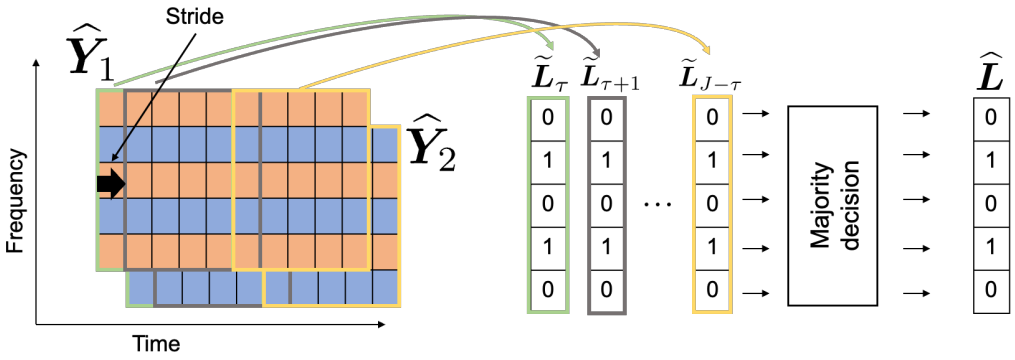

Fig. 3.5. DNN predictions for all short-time subbands and their majority decision.

$\widehat{L}$ の値に従って，パーミュテーション行列を並び替えることで，パーミュテーション問題の解決を行う．

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．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．実験データとして，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．行列のサイズは全て 100 行 100 列とした．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，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，95% の割合で 1 行ごとにシャッフルしそれ以外は 2 行ごとにシャッフルした場合と，99% の割合で 1 行ごとにシャッフルしそれ以外は 2 行ごとにシャッフルした場合の実験も行った．加えて，実際の音声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 [18] の英語の音声信号（男性 1 名及び女性 1 名）2 種類を使用した．音声信号に対する STFT は，fft サイズ 2048 ms，シフトサイズ 1024 ms に設定した．また，音声データに対しては 8 行ごとにシャッフルした場合と 16 行ごとにシャッフルした場合の 2 種類の実験を行った．学習データには，完全分離信号 $\boldsymbol{Z}_1$ と $\boldsymbol{Z}_2$ の各周波数成分をランダムでシャッフルしたものを用いた．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 $\boldsymbol{Z}_1$ と $\boldsymbol{Z}_2$ の各周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[19] を用い，ハイパーパラメータはそれぞれ

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

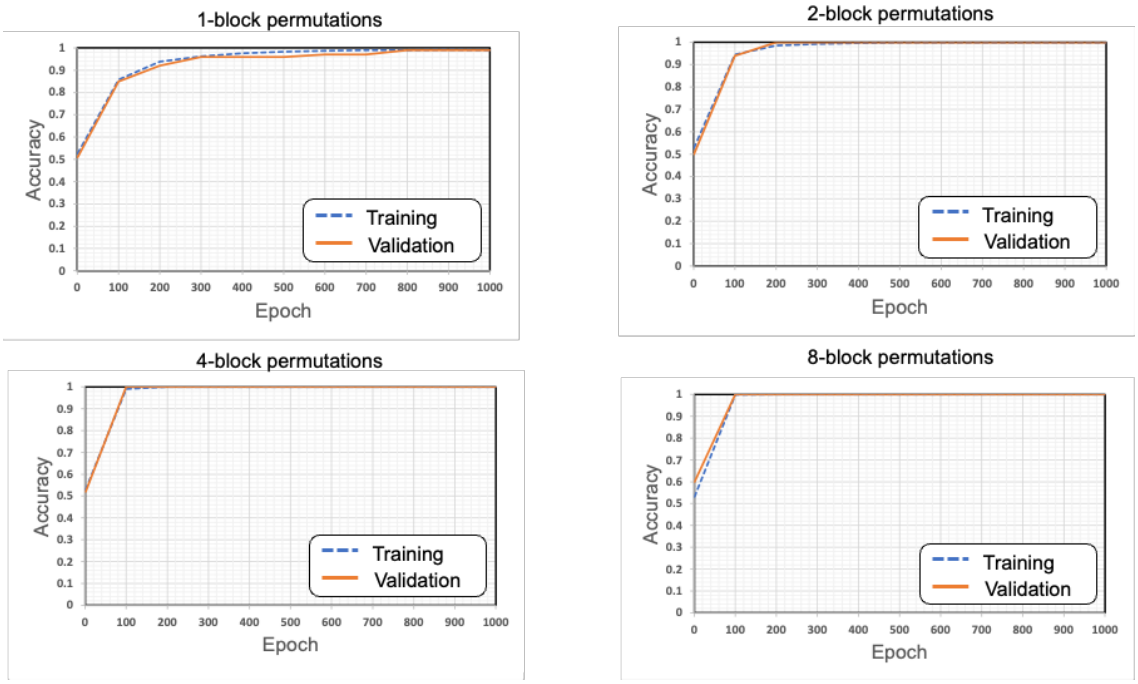

Fig. 4.1. Accuracy curves of DNN for training and validation with datasets of a matrix with only zeros and a matrix with only ones.

$\varepsilon = 1.0 \times 10^{-8}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ 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．

4.3　実験結果

Fig. 4.1〜Fig. 4.3 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それぞれ 1 行，2 行，4 行，8 行分をまとめてシャッフルを行った時の結果を示す．この結果から，各周波数成分の 8 行分をまとめてシャッフルした場合，実験を行った全てのパターンにおいてどれも正答率が 100% に近い値となっている．しかし，1 列毎に 0 と 1 の値が入れ替わる行列に対して各行ごとにシャッフルを行った場合は，正答率が 54% 程度となった．即ち，提案手法において各行ごと

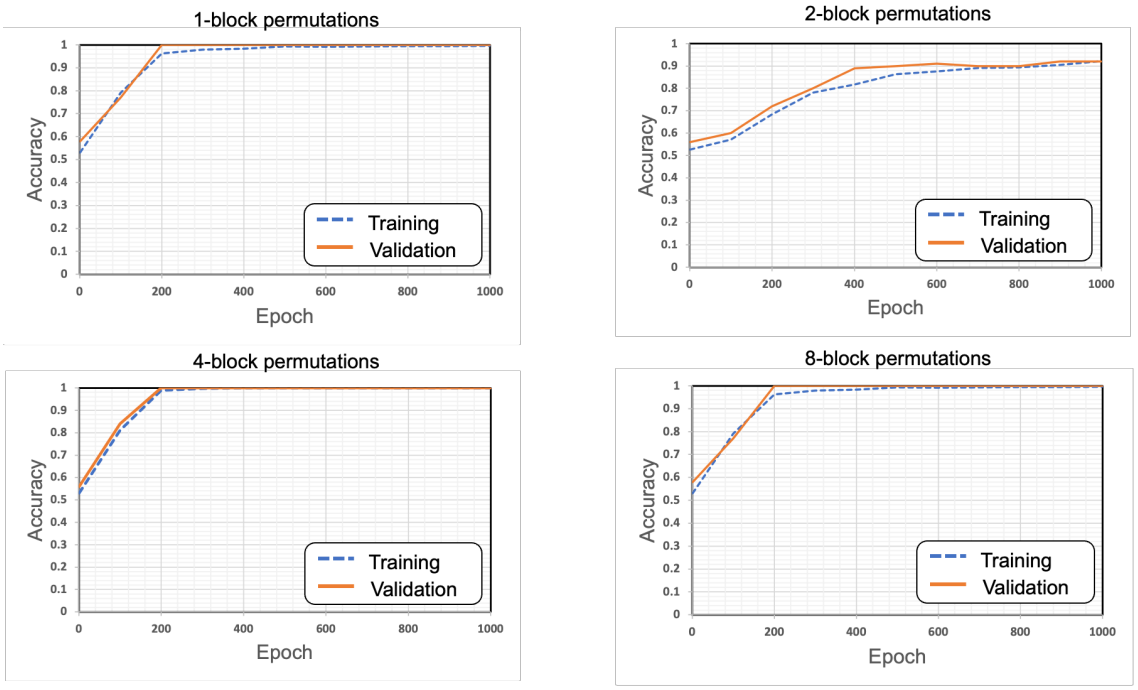

Fig. 4.2. Accuracy curves of DNN for training and validation with datasets of two matrices with 0 and 1 values swapped every 25 columns.

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. 4.4 と Fig. 4.5 より，1 列毎に 0 と 1 の値が入れ替わる行列に対して，95% の割合で 1 行ごとにシャッフルしそれ以外は 2 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，99% の割合で 1 行ごとシャッフルしそれ以外は 2 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．Fig. 4.6 と Fig. 4.7 より，音声データに対して 8 行ごとと 16 行ごとのブロック単位でのパーミュテーション問題として考え実験を行った結果，どちらのグラフでも正答率が 80% を超えた．

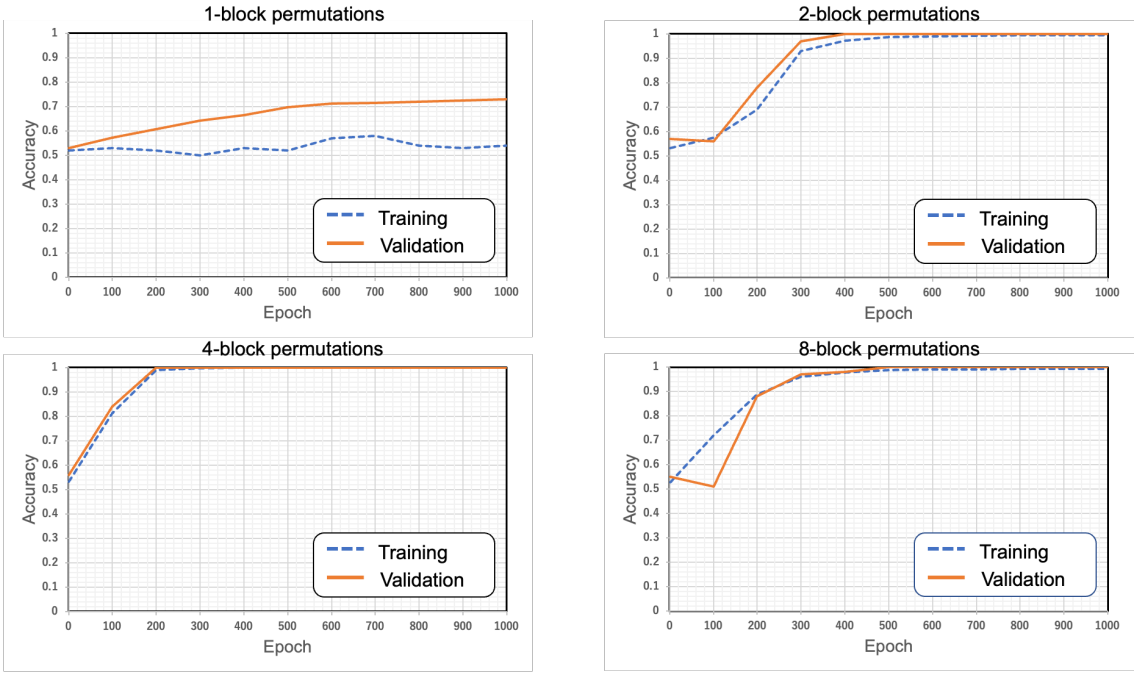

Fig. 4.3. Accuracy curves of DNN for training and validation with datasets of two matrices with 0 and 1 values swapped per column.

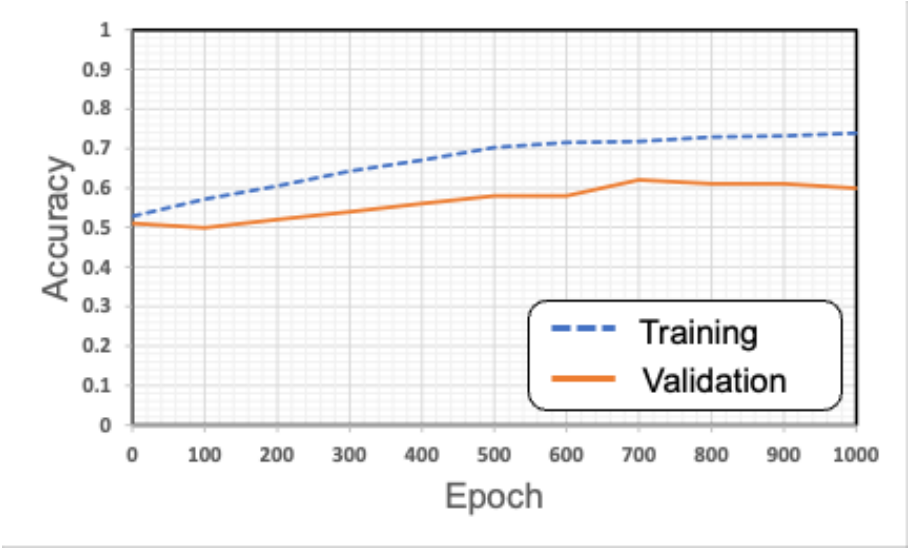

Fig. 4.4. Accuracy curves of DNN for training and validation with datasets of two matrices with 0 and 1 values swapped in each column with a 1% probability of shuffling two rows together and the rest shuffled row by row.

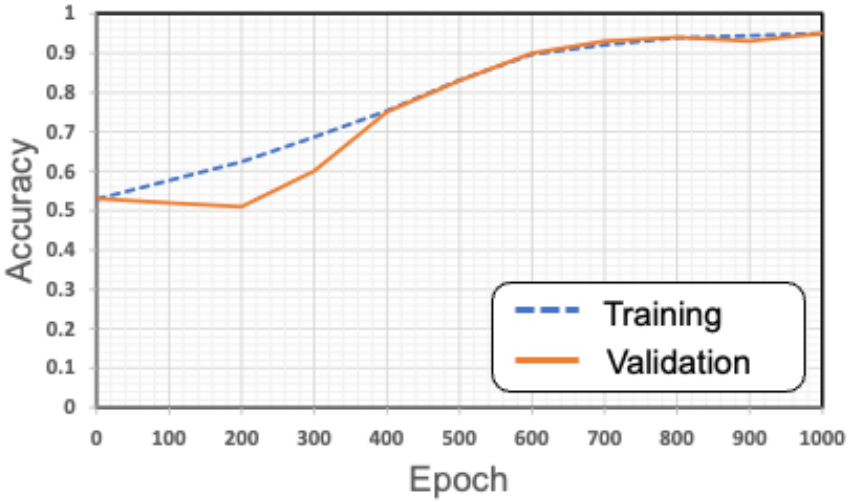

Fig. 4.5. Accuracy curves of DNN for training and validation with datasets of two matrices with 0 and 1 values swapped per column with a 5% probability of shuffling two rows together and the rest shuffled row by row.

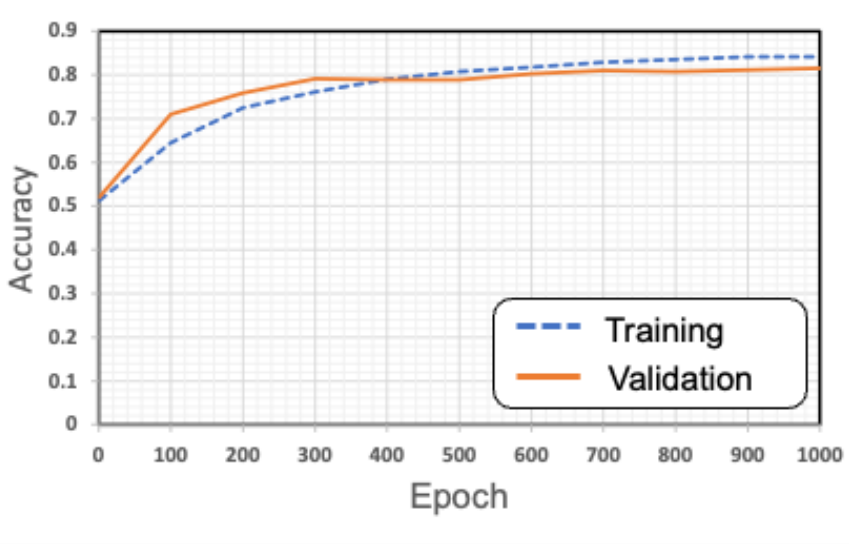

Fig. 4.6. Accuracy curves of DNN for training and validation with datasets of audio data shuffled into groups of 8 lines each.

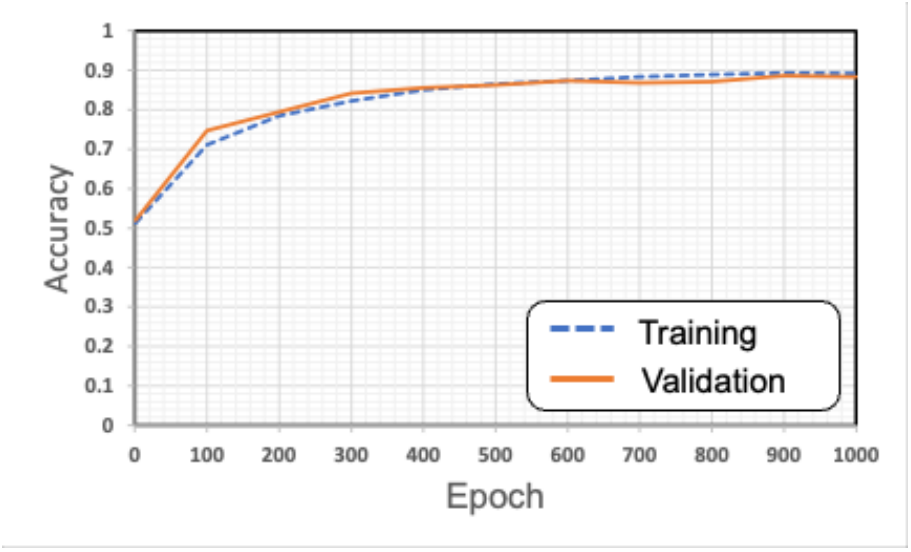

Fig. 4.7. Accuracy curves of DNN for training and validation with datasets of audio data shuffled into groups of 16 lines each.

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ことを示した．

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 音源以上で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．心よりありがたくお礼申し上げます．

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. 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.

[6] 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.

[7] 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.

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

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

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," IEEE/ACM Trans. ASLP, vol. 24, no. 9, pp. 1626–1641, 2016.

[11] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in Audio Source

Separation, S. Makino, Ed., pp. 125–155. Springer, Cham, 2018.

[12] 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.

[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] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," Proc. ICA, pp. 722–727, 2001.

[15] 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.

[16] 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.

[17] V. Nair, and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," Proc. ICML, 2010.

[18] 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.

[19] D. P. kingma, and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv, pp. 1412–6980, 2014.