特別研究論文

（査読済み）

研　究　題　目

局所時間周波数構造に基づく深層パーミュテーション解決法の実験的評価

提出年月日	2021 年　1 月　25 日	
氏　　名		山地　修平
主　　査		北村　大地
副　　査		村上　幸一
副　　査		雛元　洋一

香川高等専門学校
専攻科
創造工学専攻

Experimental evaluation of deep permutation solver based on local time-frequency structure

Shuhei Yamaji

Advanced Course in Industrial and Systems 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, in FDICA, the order of the estimated signal in each frequency is not aligned among all frequencies, resulting in the so-called permutation problem. Thus, FDICA requires a permutation solver as a post-process. In recent years, independent vector analysis and independent low-rank matrix analysis (ILRMA) have been proposed. These methods introduce a source model to FDICA to avoid encountering the permutation problem. Although these methods can avoid the permutation problem to some extent, for the mixture with strong reverberation, they often fail to separate the sources. On the other hand, it has been confirmed in the previous study that FDICA can separate very high quality for each frequency. A remaining issue is only the permutation problem.

In this thesis, we propose a new method for solving the permutation problem using deep neural networks (DNNs).The DNN learns the features of the time-frequency structures of audio sources and predicts whether the permutation error occur. Then, the permutation alignment is performed based on the predicted results of the DNN. To evaluate the performance of the proposed permutation solver, source separation experiment using two-speech mixtures with strong reverberation have been conducted. The experimental results show that FDICA with the proposed DNN-based permutation solver can achieve about 4 dB improvement from the state-of-the-art algorithm, ILRMA, in terms of a sources-to-distortion ratio. I also show that the separation accuracy of the proposed DNN-based permutation solver does not change even when the spatial arrangement of the sources (mixture condition) in the test dataset is different from that in the training dataset. Therefore, the proposed method can be used as a general permutation solver that does not depend on the locations of each source. In this thesis, I only focus on a speech source separation problem in two-source mixture case. The extension of the proposed method to three or more sources is a future work.

Key Words: Independent component analysis, Deep neural network, Permutation problem

（和訳）

本論文は，音源分離技術について取り扱う．音源分離とは，様々な音源が混ざった観測信号から，混ざる前の個々の音源信号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場合などで役立つ．代表的な音源分離手法の 1 つとして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，このパーミュテーション問題を可能な限り回避するような音源分離手法が提案されており，特に独立ベクトル分析（independent vector analysis: IVA）や独立低ランク行列分析が有名である．これらの手法では，パーミュテーション問題をある程度避けながら音源分離が可能であるが，より残響の強い観測信号に対しては，しばしば分離に失敗することが確認されている．一方で，FDICA では，周波数毎の音源分離は非常に高精度で実現できることが先行研究から確認でき，パーミュテーション問題だけが残された課題であることが分かっている．

本論文では，深層ニューラルネットワーク（deep neural networks: DNNs）を用いたパーミュテーション問題の解決法を提案する．提案手法の DNN は，音源信号の複雑な時間周波数構造の特徴を事前に学習し, 複数の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. この DNN の予測結果を用いて，全周波数のパーミュテーション問題を解決する新しいアルゴリズムを提案する．提案手法のパーミュテーション問題の解決性能を評価するために，高残響下における 2 話者音声の混合信号の音源分離実験を実施した．実験結果より，提案する DNN パーミュテーション解決法と FDICA を組み合わせた音源分離手法は，既存の最先端手法である ILRMA から，信号対歪み比において 4~dB もの改善があることが明らかになった．また，提案手法中のパーミュテーション解決法は，テスト時の観測信号中の各音源の音源到来方向が学習用データのそれと異なるという条件であっても，推定精度が変わらないことを実験的に示した．そのため，各音源の位置関係に依存することのない，一般的なパーミュテーション解決法として提案手法を扱うことができる．本論文は，2 音源の音源分離を対象としているが，3 音源以上への拡張は今後の課題である．

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，音源分離技術は近年ニーズが高まっており，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およ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，混合系 $\boldsymbol{A}$ に適用することで元の音源を推定する．

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．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

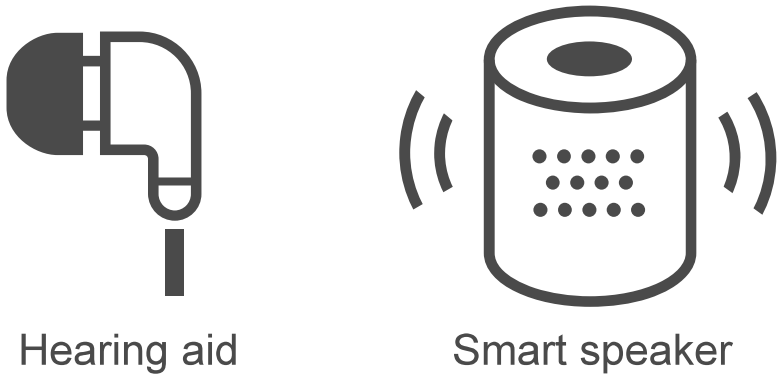


Fig. 1.1. Examples of application using speech source separation.

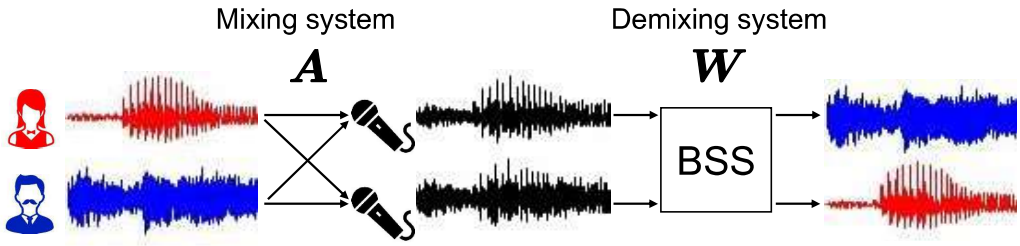


Fig. 1.2. Overview of BSS.

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．さらに，深層ニューラルネットワーク（deep neural networks: DNNs）を用いて音源の時間周波数構造の仮定を音源データから事前に学習し，FDICA に適用する独立深層学習行列分析 [12] も提案されている．

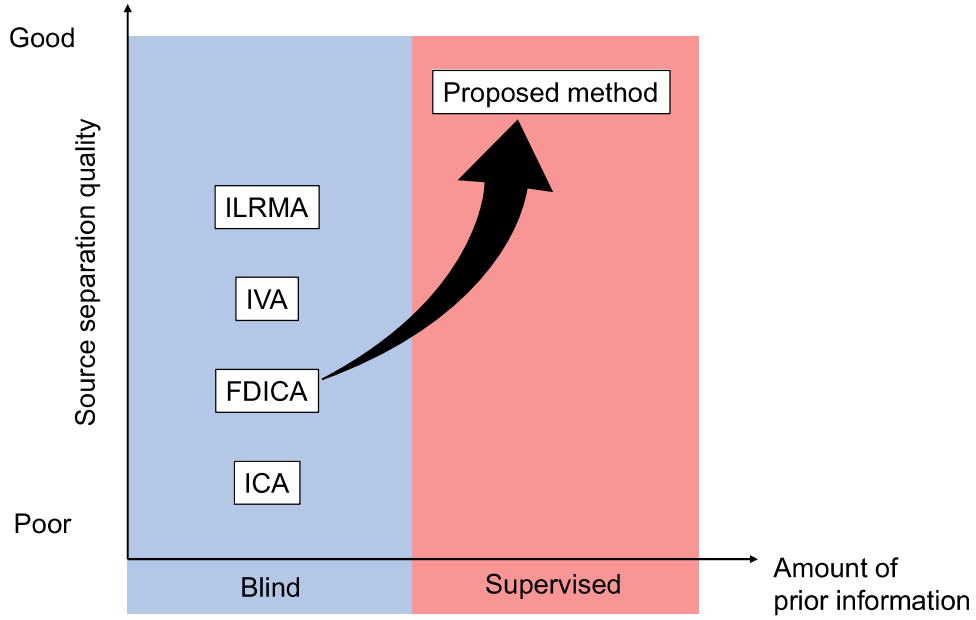


Fig. 1.3. Scope of this thesis.

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．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．この結果は，FDICA はパーミュテーション問題のみが課題であり，周波数毎の分離は非常に高い精度で達成されていることを示している．

そこで，本論文では，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を新たに提案する．この提案手法の既存手法に対する立ち位置を 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 を用いてのパーミュテーション解決を目的とする．ここでは，優決定条件下での複数音声の混合を対象とし，FDICA で周波数毎の分離信号を求めた後に DNN に基づくパーミュテーション解決法を適用することを考える．

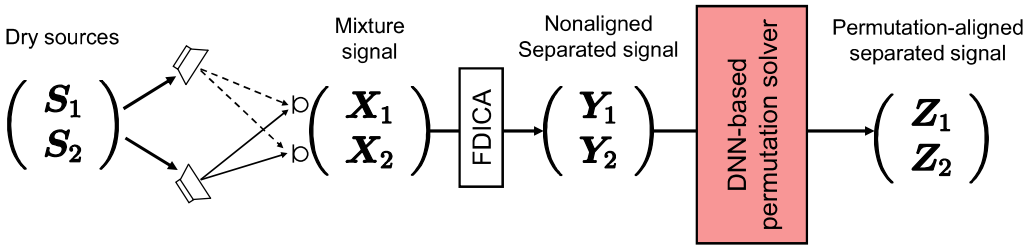


Fig. 1.4. Objective of this thesis.

1.3　本論文の構成

まず，2 章では，音源分離手法の 1 つである FDICA と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について詳しく説明する．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アルゴリズムの詳細について述べる．4 章では，音声の混合信号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及び他音源分離手法との比較検討を行う．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

従来手法

2.1 まえがき

まず，2.2 節では，提案手法において必要な基礎理論を説明するため，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 の定式化を導入する．2.5 節では，音源分離手法の 1 つである FDICA について説明する．2.6 節では，パーミュテーション問題と呼ばれる FDICA に伴う課題の説明と，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．

2.2 ICA の基本原理

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 t 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．

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

この行列 $\boldsymbol{A}$ は混合行列と呼ばれる．観測信号ベクトル $\boldsymbol{x}(l) = (x_1(l), x_2(l))^T$ 信号源ベクトル $\boldsymbol{s}(l) = (s_1(l), s_2(l))^T$ 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 $y(l) = (y_1(l), y_2(l))^T$，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 $\boldsymbol{W}$ を選択することで，信号源 $s(l)$ 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 $\boldsymbol{W}$ 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，$s_1(l)$ から $s_2(l)$ を推定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) 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．

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

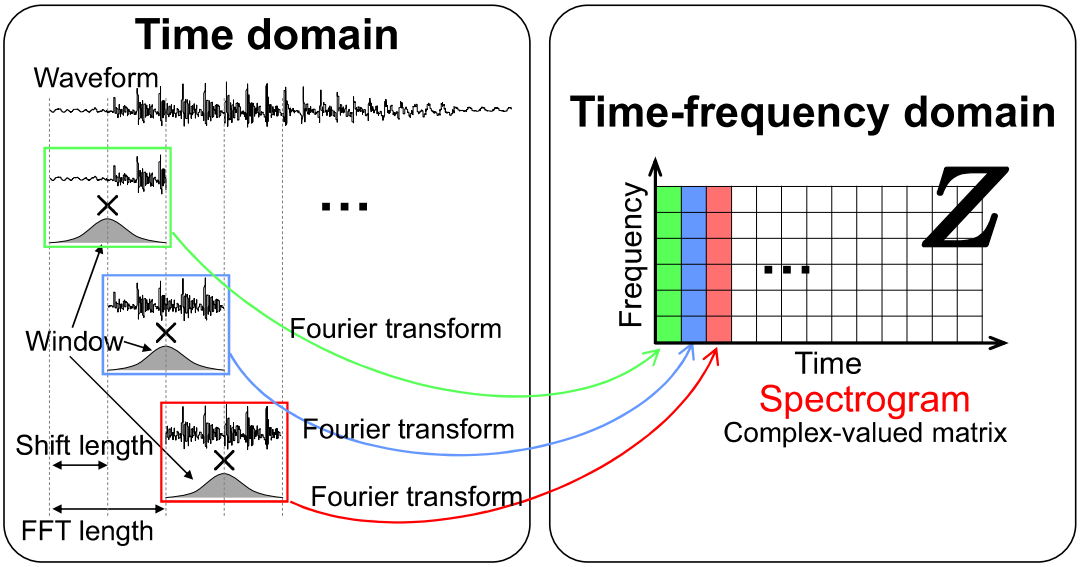


Fig. 2.1. Mechanism of STFT.

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

ここで，$\mathfrak{D}_{KL}[p(s)||q(s)]$ 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．

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

また，分離フィルタ $\boldsymbol{W}$ で線形変換する前（$\boldsymbol{x}$）と後（$\boldsymbol{y}$）の確率変数を考えたとき，それぞれの同時分布 $p(\boldsymbol{y}) = p(y1, y2)$ と $p(\boldsymbol{x}) = p(x1, x2)$ 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) を $\boldsymbol{W}$ について最小化することで，信号源を分離する．

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．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

$z(l)$ 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．

$$\boldsymbol{z}^{(j)} = \left(z\left((j-1)\tau+1\right), z\left((j-1)\tau+2\right), \cdots, z\left((j-1)\tau+Q\right)\right)^T \tag{2.16}$$

$$= \left(z^{(j)}(1), z^{(j)}(2), \cdots, z^{(j)}(q), \cdots, z^{(j)}(Q)\right)^T \in \mathbb{R}^Q \tag{2.17}$$

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．

$$J = \frac{L}{\tau} \tag{2.18}$$

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．

$$\boldsymbol{Z} = \mathrm{STFT}_{\boldsymbol{\omega}}(\boldsymbol{z}) \in \mathbb{C}^{I\times J} \tag{2.19}$$

また，スペクトログラム $\boldsymbol{Z}$ の (i, j) 番目の要素は次式で表される．

$$z_{ij} = \sum_{q=1}^{Q} \omega(q) z^{(j)}(q) \exp\left\{\frac{-\iota 2\pi(q-1)(i-1)}{F}\right\} \tag{2.20}$$

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，$\boldsymbol{\omega}$ は分析窓関数を示している．このように，時間領域の信号は一定幅の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，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 $\boldsymbol{Z}$ で表すことができる．

2.4　周波数領域における BSS の定式化

今一度，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

$$\boldsymbol{s}_{ij} = \left(s_{ij,1}, s_{ij,2}, \cdots, s_{ij,n}, \cdots, s_{ij,N}\right)^{\mathrm{T}} \in \mathbb{C}^N \tag{2.21}$$

$$\boldsymbol{x}_{ij} = \left(x_{ij,1}, x_{ij,2}, \cdots, x_{ij,m}, \cdots, x_{ij,M}\right)^{\mathrm{T}} \in \mathbb{C}^M \tag{2.22}$$

$$\boldsymbol{z}_{ij} = \left(z_{ij,1}, z_{ij,2}, \cdots, z_{ij,n}, \cdots, z_{ij,N}\right)^{\mathrm{T}} \in \mathbb{C}^N \tag{2.23}$$

と表す．ここで，$i = 1, 2, \cdots, I$，$j = 1, 2, \cdots, J$，$n = 1, 2, \cdots, N$，及び $m = 1, 2, \cdots, M$ はそれぞれ周波数，時間，音源，チャネルのインデクスを示す．また，複素スペクトログラム行列 $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$, $\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$ 及び $\boldsymbol{Z}_n \in \mathbb{C}^{I\times J}$ 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．

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．

$$\boldsymbol{x}(l) = \sum_{n} \sum_{l'=0}^{R-1} \tilde{\boldsymbol{a}}_n(l') \boldsymbol{s}_n(l - l') \tag{2.24}$$

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 (2.24) 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．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{2.25}$$

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 $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = (\boldsymbol{w}_{i,1}\ \boldsymbol{w}_{i,2}\ ...\ \boldsymbol{w}_{i,n}\ ...\ \boldsymbol{w}_{i,N})^{\mathrm{H}}$ を用いて，分離信号を次式で表せる．

$$\boldsymbol{z}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{2.26}$$

ここで，$\cdot^{\mathrm{H}}$ 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．

$$\hat{\boldsymbol{W}}_i = \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{W}_i \tag{2.27}$$

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．すなわち，FDICA で推定される分離信号

$$\boldsymbol{y}_{ij} = \hat{\boldsymbol{W}}_i \boldsymbol{x}_{ij} \tag{2.28}$$

$$= (y_{ij,1}, y_{ij,2}, \cdots, y_{ij,n}, \cdots, y_{ij,N})^{\mathrm{T}} \quad \in \mathbb{C}^N \tag{2.29}$$

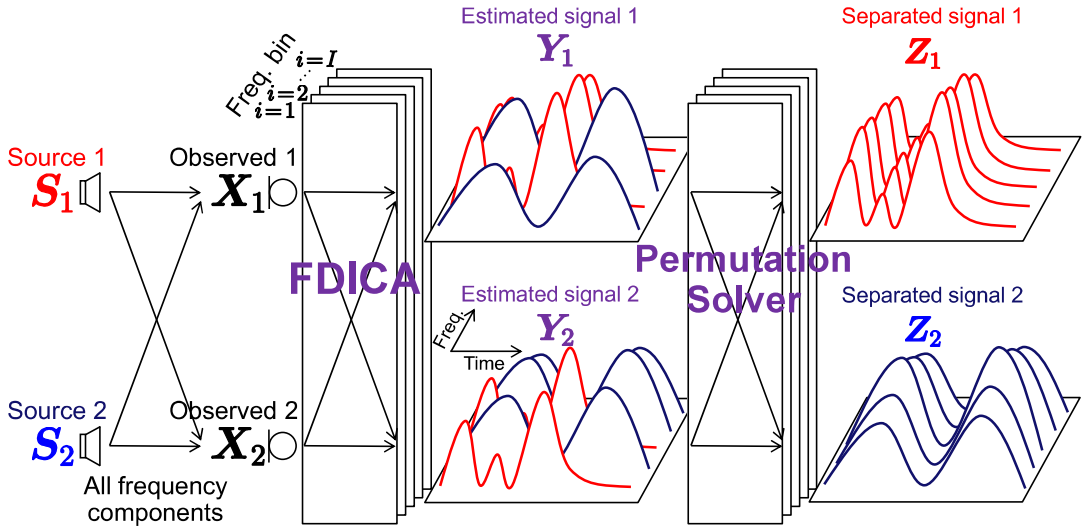


Fig. 2.2. Permutation problem in FDICA, where $N = M = 2$.

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．

$$\boldsymbol{z}_{ij} = \boldsymbol{P}_i^{-1}\boldsymbol{D}_i^{-1}\boldsymbol{y}_{ij} \tag{2.30}$$

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] およ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

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 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および各種従来手法について説明した．次章以降では，より高度な音声信号の 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．

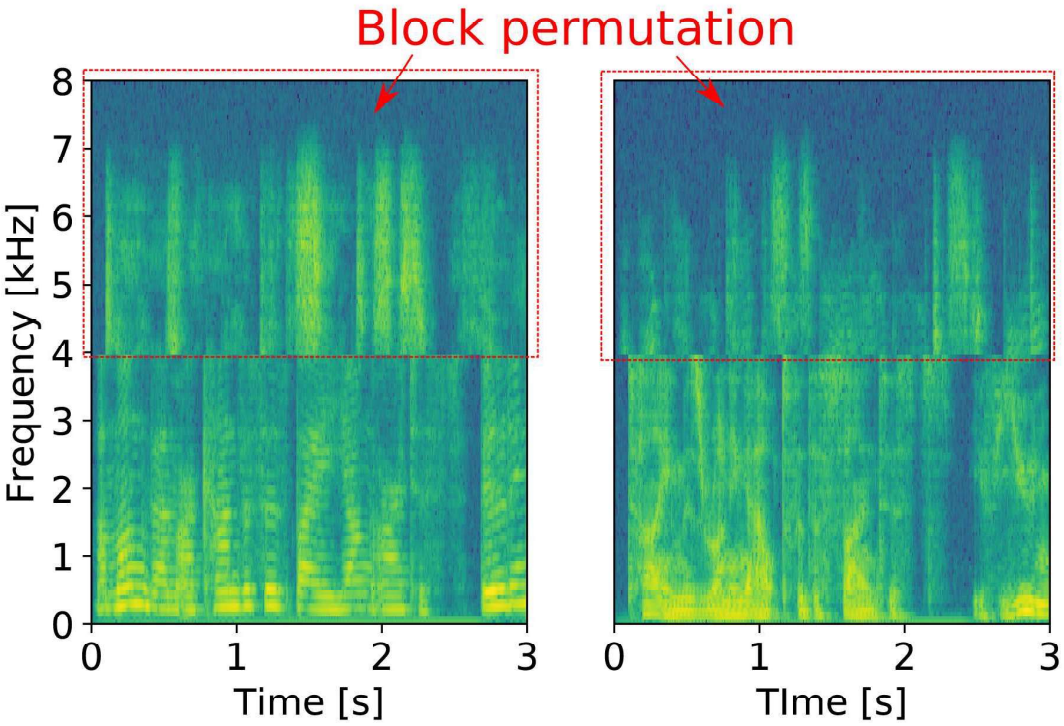


Fig. 2.3. Example of block permutation problem.

第 3 章

提案手法

3.1 まえがき

前章では，BSS 手法の 1 つである FDICA とそれによって生じるパーミュテーション問題について説明した．本章では，FDICA のポスト処理として，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，3.6 節及び 3.7 節では，DNN の推定結果を用いて推定信号 Y_n のパーミュテーションを正しく並べ替えるアルゴリズムの詳細を示す．3.8 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，これはすなわ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，IVA や ILRMA も周波数毎の分離は成功している一方で，ブロックパーミュテーション問題）[15] の様にパーミュテーション解決にのみ失敗している可能性も考えられる．これは，IVA や ILRMA で仮定されている音源モデルが音声音源に適していないためと考えられる．実際に，IVA の音源モデル，すなわち同一音源のすべての周波数成分が共起するといった仮定は，音声音源に対してはやや単純化されすぎており，パーミュテーション問題を完全に回避することは難しい．また，ILRMA の音源モデル，すなわち同一音源の低ランク時間周波数

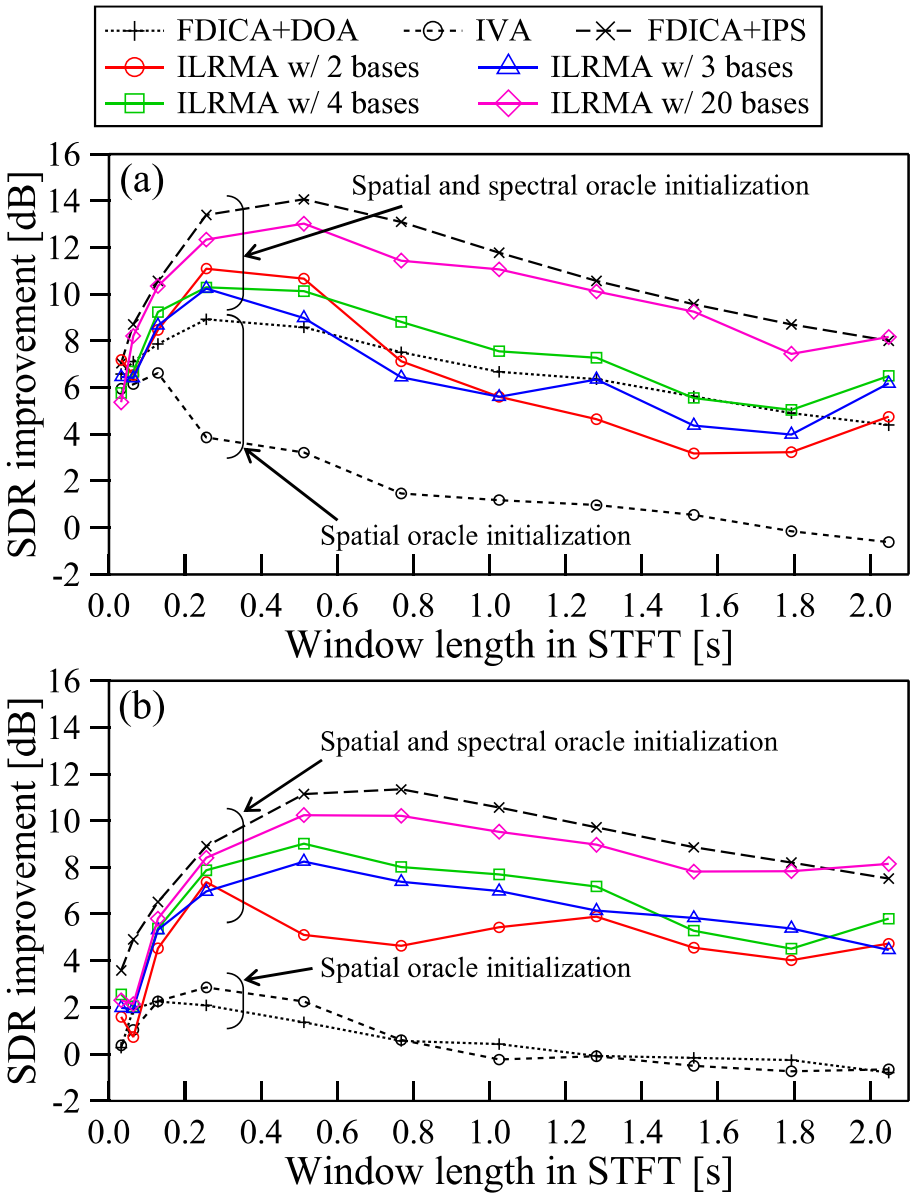


Fig. 3.1. Average source separation results for speech signals using random initialization: (a) E2A (T_{60} =300 ms) and (b) JR2 (T_{60}=470 ms) impulse responses [13].

構造の仮定は，動的かつ連続的にスペクトルを変化させる音声信号のパワースペクトログラムには適合しないことも考えられる．そこで，本論文では，パーミュテーション問題を正確に解くことにのみ焦点を当て，新しい DNN に基づくデータ駆動型（教師あり）パーミュテーション解決法を提案する．以後，本論文では，音源数 $N = 2$ 及びチャネル数 $M = 2$ を仮定したうえで，パーミュテーション問題の解決を考える．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 $\boldsymbol{Y}_2$ から共通する 2 つの周波数の時系列パワーをそれぞれ抽出し DNN に入力する

- DNN は入力された 2 つの時系列パワーが同一音源か否かを予測し 0 または 1 として出力する
- $\boldsymbol{Y}_1$ 及び $\boldsymbol{Y}_2$ の全時間及び全周波数に対して DNN が適用される
- 最終的な推定値 $\boldsymbol{P}_i^{-1}$ は，予測値の時間方向及び周波数方向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，近傍にある異なる 2 つの周波数アクティベーションのパーミュテーションが正しいか否かを 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．そのため，提案手法は，隣接した周波数アクティベーションの相関に基づいてパーミュテーション解決を行う従来手法 [4] の教師ありへの拡張として解釈することができる．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

提案する DNN モデルの入力ベクトルを Fig. 3.2 に示す．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．これらのパワースペクトログラム $|\boldsymbol{Y}_n|^{\cdot 2}$ から，2 つの周波数（i,$i+\omega$）の短時間時系列パワー（長さ τ）を以下のように集める．

$$\boldsymbol{d}_{i,\omega,\gamma} = (\tilde{\boldsymbol{r}}_{i,\gamma}^{\mathrm{T}}, \tilde{\boldsymbol{g}}_{i,\omega,\gamma}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{4\tau\times 1} \tag{3.1}$$

$$\tilde{\boldsymbol{r}}_{i,\gamma} = (\boldsymbol{r}_{i,\gamma,1}{}^{\mathrm{T}}, \boldsymbol{r}_{i,\gamma,2}{}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{2\tau\times 1} \tag{3.2}$$

$$\boldsymbol{r}_{i,\gamma,n} = (|y_{i,(\gamma-1)\eta+1,n}|^2, |y_{i,(\gamma-1)\eta+2,n}|^2, \cdots, |y_{i,(\gamma-1)\eta+\tau,n}|^2)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{\tau\times 1} \tag{3.3}$$

$$\tilde{\boldsymbol{g}}_{i,\omega,\gamma} = (\boldsymbol{g}_{i,\omega,\gamma,1}{}^{\mathrm{T}}, \boldsymbol{g}_{i,\omega,\gamma,2}{}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{2\tau\times 1} \tag{3.4}$$

$$\boldsymbol{g}_{i,\omega,\gamma,n} = (|y_{i+\omega,(\gamma-1)\eta+1,n}|^2, |y_{i+\omega,(\gamma-1)\eta+2,n}|^2, \cdots, |y_{i+\omega,(\gamma-1)\eta+\tau,n}|^2)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{\tau\times 1} \tag{3.5}$$

ここで，行列の $|\cdot|^{\cdot 2}$ は，要素ごとの絶対値の二乗を返す．また，$\omega = -\Omega, -\Omega+1, \cdots, -1, 0, 1, \cdots, \Omega$ は，$\boldsymbol{r}_{i,\gamma,n}$ と $\boldsymbol{g}_{i,\omega,\gamma,n}$ の周波数の差であり，η は，短時間セグメントの時間軸に沿ったストライド幅，$\gamma = 1, 2, \cdots, \Gamma$ は，短時間セグメントのインデクスである．なお，Γ は，短時間のアクティベーションの長さ τ とストライド幅 η によって決まる．ベクトル $\boldsymbol{r}_{i,\gamma,n}$ は，参照周波数 i の短時間時系列パワーに対応し，ベクトル $\boldsymbol{g}_{i,\omega,\gamma,n}$ は，Fig. 3.2 に示すように，隣接又は局所周波数 $i+\omega$ の短時間時系列パワーに対応する．DNN の入力ベクトルは，(3.1) を正規化したものとして次のようにして表す．

$$\tilde{\boldsymbol{d}}_{i,\omega,\gamma} = \frac{\boldsymbol{d}_{i,\omega,\gamma}}{\|\boldsymbol{d}_{i,\omega,\gamma}\|_2} \in \mathbb{R}_{\geq 0}^{4\tau\times 1} \tag{3.6}$$

提案する DNN モデルは，0 または 1 を出力する 2 値分類器である．推定結果が「0」の場合は，$\boldsymbol{r}_{i,\gamma,1}$ と $\boldsymbol{g}_{i,\omega,\gamma,1}$ が同一音源であることを意味し，同様に $\boldsymbol{r}_{i,\gamma,2}$ と $\boldsymbol{g}_{i,\omega,\gamma,2}$ も同一音源である．一方，推定結果が「1」の場合は $\boldsymbol{r}_{i,\gamma,1}$ と $\boldsymbol{g}_{i,\omega,\gamma,1}$（同様に $\boldsymbol{r}_{i,\gamma,2}$ と $\boldsymbol{g}_{i,\omega,\gamma,2}$）が異なる音

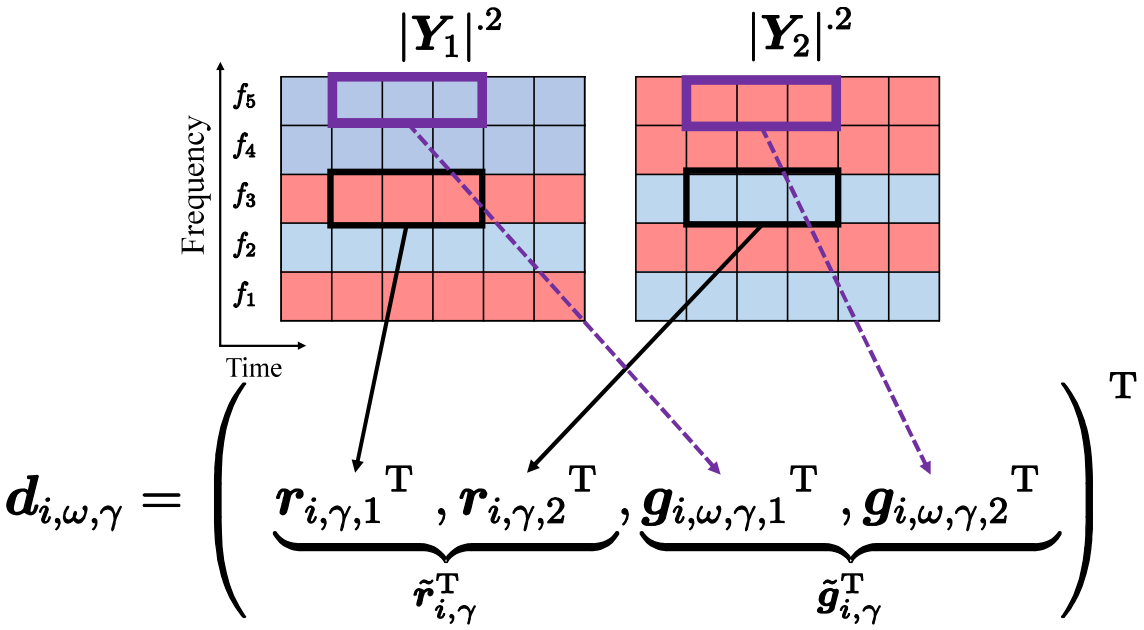


Fig. 3.2. Input vector of DNN. Matrices $|\boldsymbol{Y}_1|^{\cdot 2}$ and $|\boldsymbol{Y}_2|^{\cdot 2}$ are separated power spectrograms with permutation problem, and red and blue binwise activations (rows of $|\boldsymbol{Y}_1|^{\cdot 2}$ and $|\boldsymbol{Y}_2|^{\cdot 2}$) depict sourcewise components, e.g., red and blue slots respectively correspond to first and second source components.

源成分であることを意味している．これらの推定処理を Fig. 3.3 に示す．実際には，DNN の予測結果は 2 値ではなく次式のような値である．

$$q_{i,\omega,\gamma} = \mathrm{DNN}\left(\tilde{\boldsymbol{d}}_{i,\omega,\gamma}\right) \in [0,1] \tag{3.7}$$

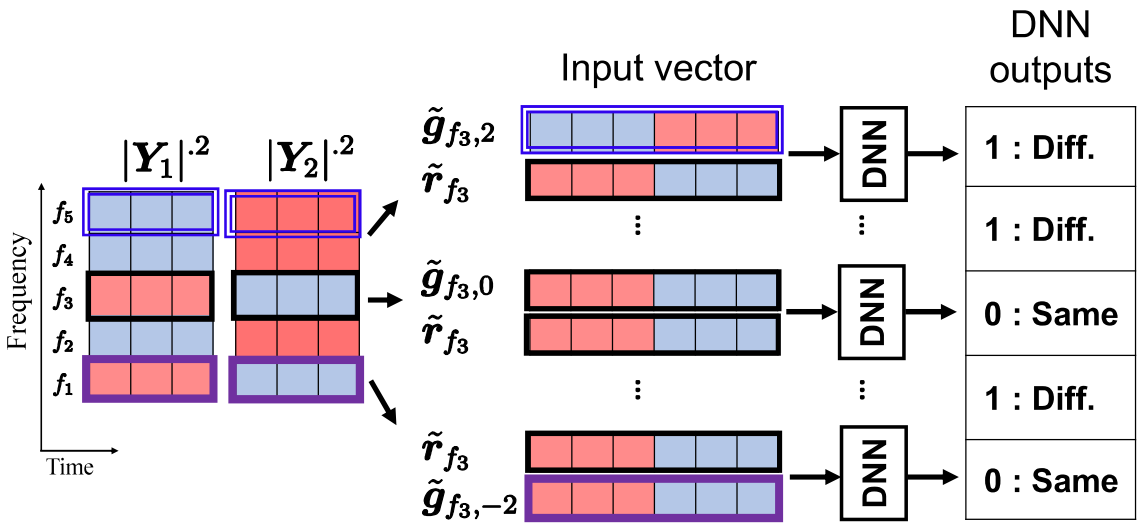


Fig. 3.3. DNN predictions in subband frequency bins, where $f_1, f_2, \cdots, f_5$ are frequency bins in subband frequency, and index of short-time activations, γ, is omitted for simplicity. Reference frequency bin is $i = f_3$, and adjacent or local frequency bins are $i + \omega = f_1, f_2, \cdots, f_5$, namely, $\Omega = 2$. When source permutation of $\tilde{\boldsymbol{r}}_i$ and $\tilde{\boldsymbol{g}}_{i,\omega}$ is correct, DNN ideally outputs zero as "same." In contrast, when source permutation of $\tilde{\boldsymbol{r}}_i$ and $\tilde{\boldsymbol{g}}_{i,\omega}$ is incorrect, DNN ideally outputs one as "different."

3.4　DNN の構造

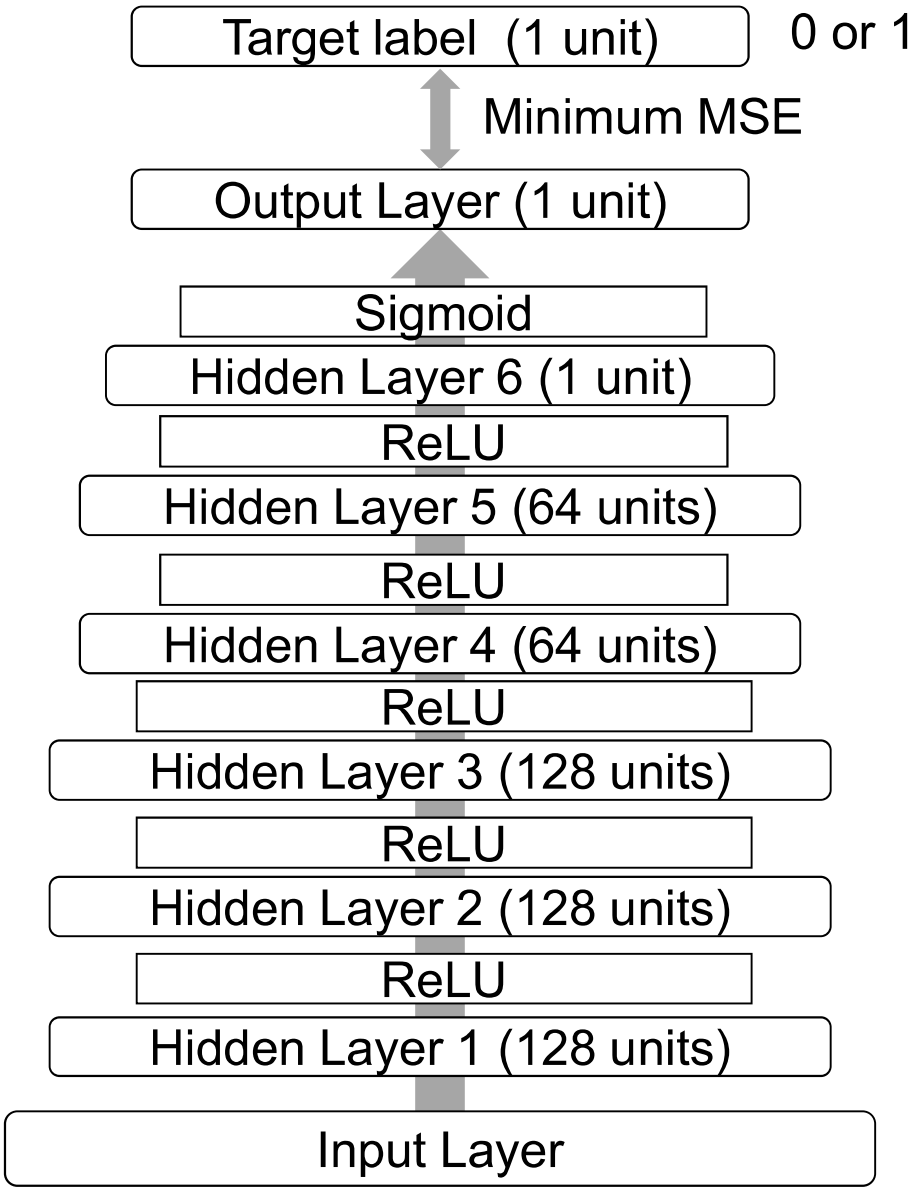


Fig. 3.4. DNN architecture.

Fig. 3.4 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 6 層，及び出力層の計 8 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 sigmoid 関数を適用している．各隠れ層の次元数は入力から順番に 128, 128, 128, 64, 64 である．予測結果 $q_{i,\omega,\gamma}$ と正解ラベルとの誤差関数には，平均二乗誤差（mean squared error: MSE）を使用している．

3.5　サブバンド領域での DNN 推定

DNN の予測例を Fig. 3.3 に示す．まず，提案手法では全周波数帯域中の局所的な狭帯域（サブバンド）におけるパーミュテーション問題の解決を考える．ここで，$f_1, f_2, \cdots, f_5$ はサブバンド内の周波数であり，簡単のために γ は省略している．Fig. 3.3 では，参照周波数を $i = f_3$ とし，その近傍周波数を $i + \omega = f_1, f_2, \cdots, f_5$ 及び $\Omega = 2$ と定義している．

$\boldsymbol{Y}_1$ に着目すると，参照周波数 f_3 及び近傍周波数 f_1 の成分は赤色の音源成分であり，$\boldsymbol{Y}_1$ の f_2，f_4，及び f_5 は青色の音源成分である．この内，2 本の短時間時系列パワー $(\tilde{\boldsymbol{r}}_i, \tilde{\boldsymbol{g}}_{i,\omega})$ の全組み合わせが DNN に入力される．入力された短時間時系列パワー（i と $i + \omega$）のパーミュテーションが正しい場合，DNN は理想的には「0」を出力する．逆に，パーミュテーションが正しくない場合は，DNN は理想的には「1」を出力する．その結果 Fig. 3.3 の右側に示すように，参照周波数 i に基づくパーミュテーション問題の発生個所の推定が可能となる．例としてこの図では，参照周波数 f_3 と同じ音源成分になっているのは f_1 及び f_3 であるため，DNN の出力が正しければ f_1 及び f_3 のみが「0」となっている．一方で，f_2，f_4 及び f_5 はパーミュテーションが正しくないので「1」が出力される．以後このベクトルは，**サブバンドベクトル**と呼ぶ．実際には，DNN の出力は $[0, 1]$ の範囲内の値となるので，サブバンドベクトルの生成時は以下の閾値処理を行う．

$$\tilde{q}_{i,\omega,\gamma} = \mathrm{round}(q_{i,\omega,\gamma}) \in \{0, 1\} \tag{3.8}$$

$$\tilde{\boldsymbol{q}}_{i,\gamma} = (\tilde{q}_{i,-\Omega,\gamma}, \tilde{q}_{i,-\Omega+1,\gamma}, \cdots, \tilde{q}_{i,-1,\gamma}, \tilde{q}_{i,0,\gamma}, \tilde{q}_{i,1,\gamma}, \cdots, \tilde{q}_{i,\Omega,\gamma})^{\mathrm{T}} \in \{0, 1\}^{2\Omega+1} \tag{3.9}$$

ここで，round$(\cdot)$ は，丸め演算子である．

このサブバンドベクトルに従って周波数成分を入れ替えることで，サブバンド内においてパーミュテーション解決が可能となる．しかし，これらの処理はサブバンド内の参照周波数に基づいて並び替えているに過ぎない．そのため，参照周波数が変わる度，つまりサブバンド領域が異なる場合に，並び替えた後の音源の順番が反転する可能性があることに注意しなければならない．このようなサブバンド間でのパーミュテーション解決については 3.7 節で説明する．

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，長さ τ の短時間時系列パワー $\tilde{\boldsymbol{r}}_{i,\gamma}$ や $\tilde{\boldsymbol{g}}_{i,\omega,\gamma}$ 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，提案手法では，Fig. 3.5 に示すように，長さ τ の入力ベクトルをストライド幅 η 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，より信頼性の高いサブバンドベクトル

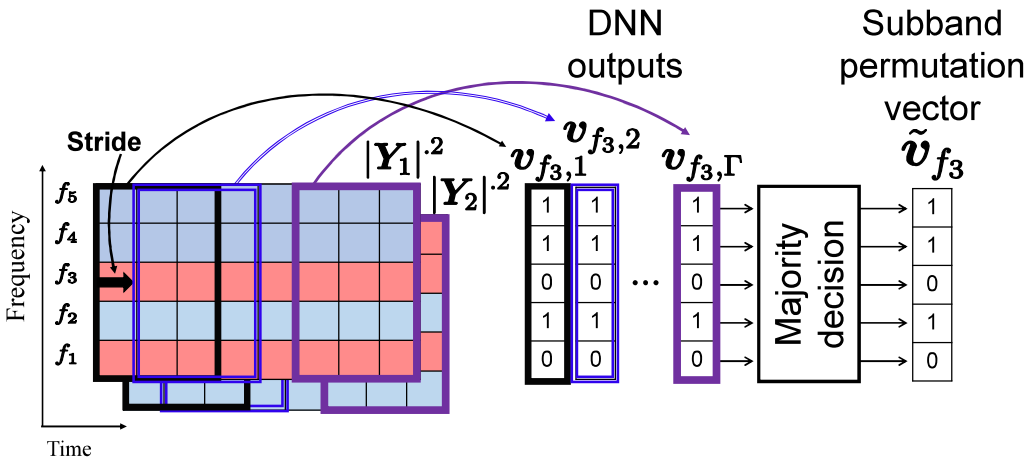


Fig. 3.5. DNN predictions for all short-time subbands and their majority decision.

$\tilde{\boldsymbol{v}}_i$ を得る．この処理は，次のように示される．

$$\boldsymbol{v}_i = \frac{1}{\Gamma}\sum_{\gamma}\tilde{\boldsymbol{q}}_{i,\gamma} \in \{0,1\}^{2\Omega+1} \tag{3.10}$$

$$\tilde{\boldsymbol{v}}_i = \mathrm{round}(\boldsymbol{v}_i) \in \{0,1\}^{2\Omega+1} \tag{3.11}$$

実際に，$\boldsymbol{Y}_1$ 及び $\boldsymbol{Y}_2$ 中の各周波数の音源の順番は時間に依存しない（同一周波数であればどの時刻も同一の音源順となっている）ため，多数決によって予測誤差の悪影響を大幅に軽減できる．

3.7　全周波数でのパーミュテーション解決

提案する DNN パーミュテーション解決法は（a）サブバンドのストライドによる全周波数のサブバンドベクトルの推定（3.7.1 項及び Fig. 3.6）及び（b）類似度比較と多数決に基づくフルバンドベクトルの構築（3.7.2 項及び Fig. 3.7）で構成される．

3.7.1　全周波数におけるサブバンドベクトルの推定

サブバンドベクトル $\tilde{\boldsymbol{v}}_i$ は，基準周波数 i をシフトすることにより全周波数で推定する．この処理を Fig. 3.6 に示す．全部で約 I 個のサブバンドベクトル $\tilde{\boldsymbol{v}}_1, \tilde{\boldsymbol{v}}_2, \cdots, \tilde{\boldsymbol{v}}_I$ を推定している．ここで，各サブバンドベクトル $\tilde{\boldsymbol{v}}_1, \tilde{\boldsymbol{v}}_2, \cdots, \tilde{\boldsymbol{v}}_I$ 内の 2 値（「0」及び「1」）は，異なる意味を持つ可能性があることに注意する．これはサブバンド内の周波数成分が，参照周波数 i 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 i の変化（サブバンドのシフト）とともに対応音源が変化するためである．

例えば，Fig. 3.6 の $\tilde{\boldsymbol{v}}_{f_3}$ の 0 と 1 は，それぞれ赤色と青色の音源成分を示している．一方で，$\tilde{\boldsymbol{v}}_{f_4}$ の 0 と 1 はそれぞれ青色と赤色の音源成分を示している．これらのサブバンドベクトルの整列は次項で処理される．

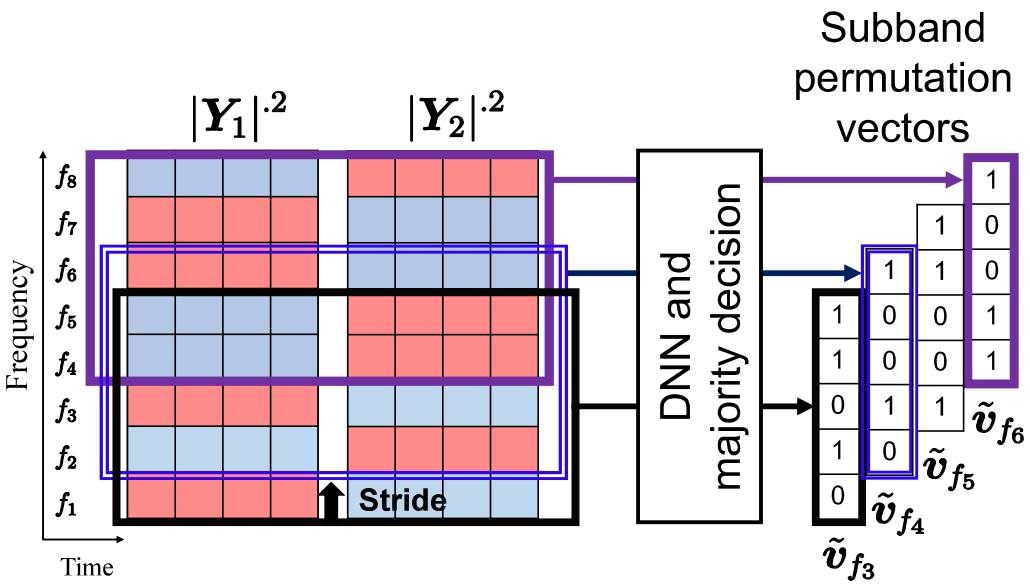


Fig. 3.6. Estimation of subband permutation vectors in all frequency bins.

3.7.2　フルバンドベクトルの作成

推定されたサブバンドベクトル $\tilde{\boldsymbol{v}}_1, \tilde{\boldsymbol{v}}_2, \cdots, \tilde{\boldsymbol{v}}_I$ から，次式で定義されたフルバンドベクトル $\boldsymbol{u}$ を構成する．

$$\boldsymbol{u} = (u_1, u_2, \cdots, u_I)^{\mathrm{T}} \in \{0,1\}^I \tag{3.12}$$

$\boldsymbol{u}$ の構成処理を Fig. 3.7 に示す．前項で述べた通り，サブバンドベクトル $\tilde{\boldsymbol{v}}_1, \tilde{\boldsymbol{v}}_2, \cdots, \tilde{\boldsymbol{v}}_I$ の 2 値は同じ意味を持たない．そのため，「0」と「1」の値がそれぞれ赤色の音源と青色の音源を示すように，全てのサブバンドベクトルを統一する必要がある．

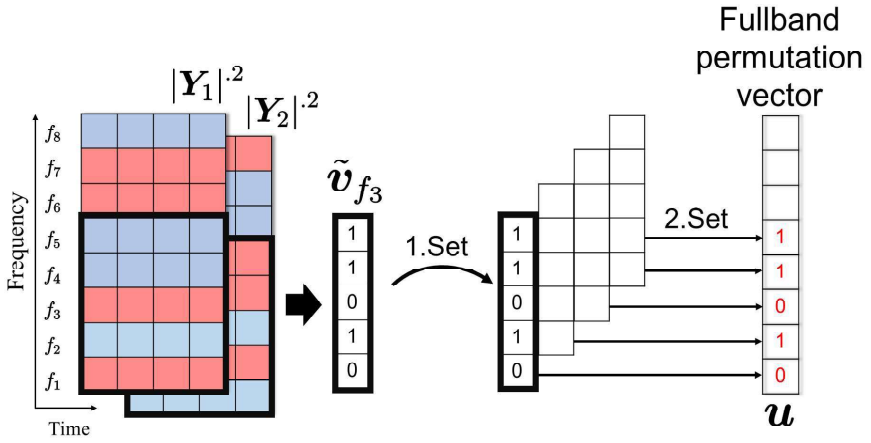
Fig. 3.7 (a) は，フルバンドベクトル $\boldsymbol{u}$ の構成における最初のステップを示している．図に示すように，最も低い周波数のサブバンドベクトル $\tilde{\boldsymbol{v}}_{i_\mathrm{s}}$ が，フルバンドベクトル $\boldsymbol{u}$ の対応する周波数に挿入される．ここで，i_s は，最も低い参照周波数のインデクスを表す．Fig. 3.7 (a) では，$i_\mathrm{s} = f_3$ 及び $\Omega = 2$ であり，$u_1, u_2, \cdots, u_5$ は $\tilde{\boldsymbol{v}}_{f_3}$ により決定される．

Fig. 3.7 (b) は Fig. 3.7 (a) の次のステップを示している．このステップでは，最も低い周波数のサブバンドに隣接するサブバンドベクトルを算出する．推定されたサブバンドベクトル $\tilde{\boldsymbol{v}}_{i_\mathrm{s}+1}$ 及びその論理反転ベクトル $\overline{\tilde{\boldsymbol{v}}_{i_\mathrm{s}+1}}$（Fig. 3.7 (b) 中の $\tilde{\boldsymbol{v}}_{f_4}$ と $\overline{\tilde{\boldsymbol{v}}_{f_4}}$）が用意される．

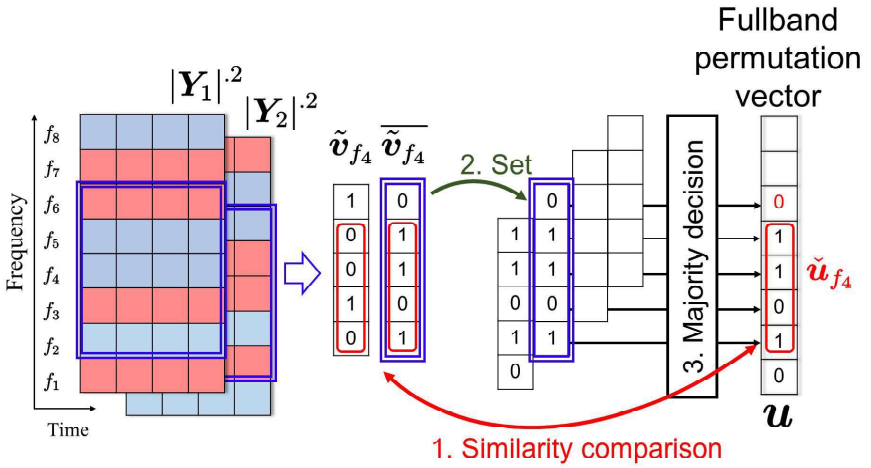
次に，$\boldsymbol{u}$ の一部を

$$\check{\boldsymbol{u}}_i = (u_{i-\Omega}, u_{i-\Omega+1}, \cdots, u_{i+\Omega-1})^{\mathrm{T}} \in \{0,1\}^{2\Omega} \tag{3.13}$$

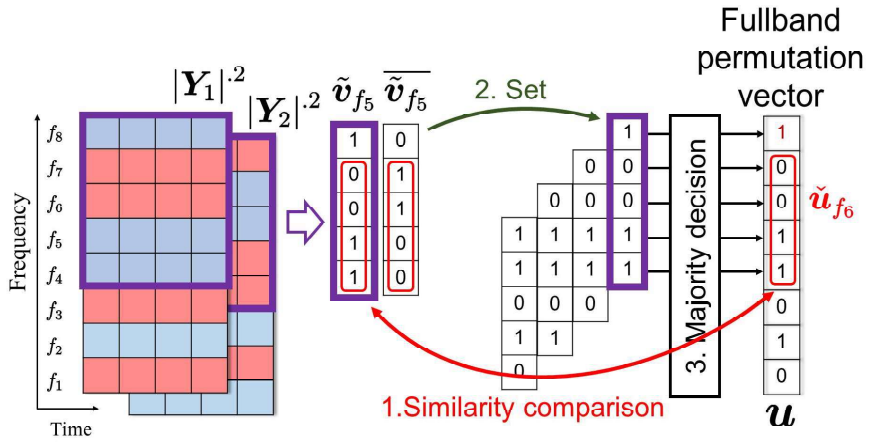
の様に定義し，$\mathrm{MSE}(\tilde{\boldsymbol{v}}_{i_\mathrm{s}+1}, \check{\boldsymbol{u}}_{i_\mathrm{s}+1})$ 及び $\mathrm{MSE}(\overline{\tilde{\boldsymbol{v}}_{i_\mathrm{s}+1}}, \check{\boldsymbol{u}}_{i_\mathrm{s}+1})$ を比較する．ここで $\mathrm{MSE}(\cdot,\cdot)$ は，2 つの入力ベクトル間の MSE 値である．その結果より，MSE が小さくなるベクトルを $\tilde{\boldsymbol{v}}_{i_\mathrm{s}+1}$ 又は $\overline{\tilde{\boldsymbol{v}}_{i_\mathrm{s}+1}}$ から選択してメモリに格納する．フルバンドベクトル $\boldsymbol{u}$ は，Fig. 3.7 (b) に示すように，メモリに格納されたベクトルを基に周波数毎に多数決処理を行って更新される．



(a)



(b)



(c)

Fig. 3.7. Reconstruction of fullband permutation label: (a) initialization, (b) second, and (c) last steps.

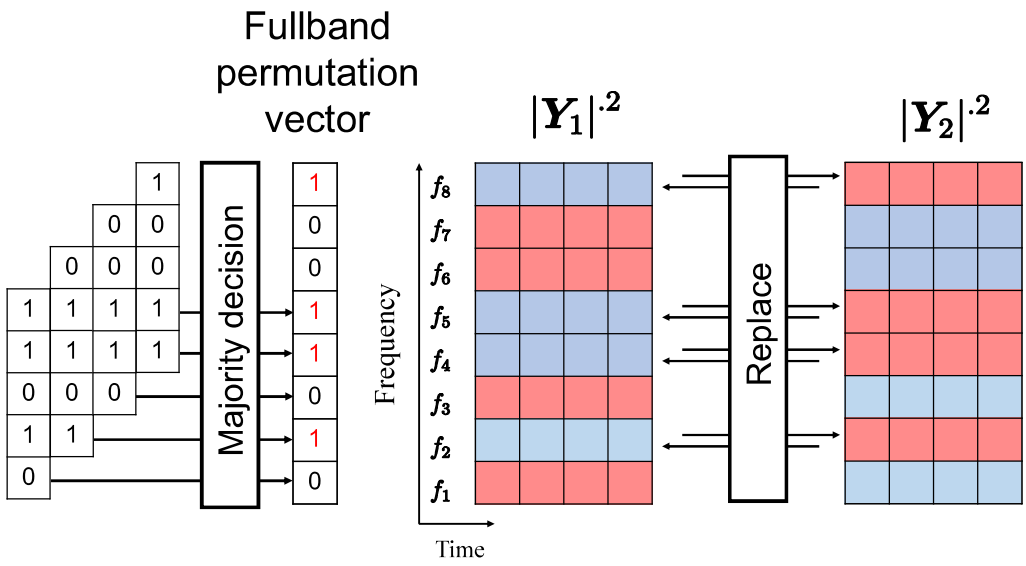


Fig. 3.8. Solving permutation problem.

以降，前述のステップを繰り返し，完全なフルバンドベクトル $\boldsymbol{u}$ が，Fig. 3.7 (c) のように構成される．提案手法では，$\boldsymbol{u}$ の構成過程の反復的多数決により，DNN 推定誤差の悪影響を軽減している．

前述のステップで構成されたフルバンドベクトル $\boldsymbol{u}$ は，$\boldsymbol{P}_i^{-1}$ の推定値そのものである．従って，Fig. 3.8 の様に，$\boldsymbol{u}$ に基づいて周波数毎の分離信号成分を入れ替えることで，全周波数におけるパーミュテーション解決が達成できる．

3.8 本章のまとめ

本章では，FDICA のポスト処理として DNN ベースのパーミュテーション解決法を新たに提案した．提案手法は（a）DNN を用いたサブバンドベクトルの推定及び（b）サブバンドベクトルを基にしたフルバンドベクトルの作成の 2 ステップで構成されている．最後に作成されたフルバンドベクトルに基づいてパーミュテーション解決を行う．

第 4 章

実験

4.1　まえがき

前章で提案した DNN に基づくデータ駆動型パーミュテーション解決法の有効性を確認するために，高残響下の音声の混合信号を FDICA で分離した後に，提案パーミュテーション解決法を適用し，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と他手法との比較結果を示している．4.4 節で本章のまとめを述べる．

4.2　実験条件

Table 4.1 に実験条件を示す．実験では IPS を用いた FDICA（FDICA + IPS）[13], ILRMA [11]，及び提案手法を用いた FDICA の 3 手法を比較した．IPS は完全に分離された音源信号 $\boldsymbol{s}_{ij}$ を利用しているため，FDICA に基づく BSS の上限性能を参考値として示している．本実験では JVS コーパス [18] の日本語の音声信号（nonpara30）を使用し，これらの音声信号に RWCP データベース [19] の JR2 インパルス応答を畳み込んで，1 ファイル当たり 10 s，残響長 470 ms の音声信号を男性 46 名 95 ファイル及び女性 48 名 95 ファイル分作成した．畳み込みに使用したインパルス応答は，Fig. 4.1 に示すように，文献 [13] に記載のマイク間隔 5.66 cm 及び音源方位 $(\theta_1, \theta_2) = (60°, 120°)$ のものを使用した．STFT は，窓長 512 ms 及びシフト長 128 ms に設定した．

Fig. 4.2 に DNN に用いた学習データの作成方法を示す．まず，観測された信号に窓長 512 ms 及びシフト幅 128 ms の STFT を適用し，$\boldsymbol{X}_1$ 及び $\boldsymbol{X}_2$ を得る．これらの信号に FDICA を適用することでパーミュテーション問題が未解決の（周波数毎の音源の順番がばらばらの）分離信号 $\boldsymbol{Y}_1$ 及び $\boldsymbol{Y}_2$ が推定される．推定信号 $\boldsymbol{Y}_n$ に IPS を適用し，パーミュテーション問題が理想的に解決された $\boldsymbol{Z}_1$ 及び $\boldsymbol{Z}_2$ を得る．この $\boldsymbol{Z}_n$ に対して，周波数毎の時系列成分を $\boldsymbol{Z}_1$ と $\boldsymbol{Z}_2$ 間でランダムにシャッフルすることで，パーミュテーション問題が模擬された信号 $\boldsymbol{Y}_1'$ 及び $\boldsymbol{Y}_2'$ とそれらの正解ラベル（各周波数の正しい音源順）を作成した．上記の処理で計 40 万ファイルの学習データを作成し，学習用及び検証用に 2 等分した．

Table 4.1. Experimental conditions

Window function in STFT	Hamming window
Window length in STFT	512 ms
Shift length in STFT	128 ms
Paramaters in Adam optimizer	Learning rate = 0.001 $\beta = 0.9$
Reverberation time	$T_{60} = 470$ ms
Source direction of training data	$(\theta_1, \theta_2) = (60°, 120°)$
Source direction of test data	$(\theta_1, \theta_2) = (60°, 120°)$ $(\theta_1, \theta_2) = (60°, 100°)$ $(\theta_1, \theta_2) = (70°, 110°)$

Table 4.2. Speech sources obtained from SiSEC2011

Signal	Language	Data name	Length [s]
Speech	English	dev1_female4_src_1	10.0
Speech	English	dev1_female4_src_2	10.0
Speech	Japanese	dev1_female4_src_3	10.0
Speech	Japanese	dev1_female4_src_4	10.0
Speech	English	dev1_male4_src_1	10.0
Speech	English	dev1_male4_src_2	10.0
Speech	Japanese	dev1_male4_src_3	10.0
Speech	Japanese	dev1_male4_src_4	10.0

DNN の学習では，バッチサイズを 128 とし，最適化アルゴリズムには Adam [20] を用いた．各ハイパーパラメータは学習率を 0.001，$\beta = 0.9$ に設定した．式 (3.3) 及び (3.5) における短時間区間長 τ は 20 とし，短時間区間のストライド幅は 4 とした．この時の入力層の次元は $20 \times 4 = 80$ であり，同一周波数に対する DNN の適用回数は $T = 37$ 回であった．評価指標には，次式で表される SDR [16] の改善量を用いた．

$$\hat{S}(l) = e_t(l) + e_i(l) + e_a(l) \tag{4.1}$$

$$\mathrm{SDR} = 10 \log \frac{\sum_{l=1}^{L} |e_t(l)|^2}{\sum_{l=1}^{L} |e_i(l) + e_a(l)|^2} \ [\mathrm{dB}] \tag{4.2}$$

ここで，$\hat{S}(l)$ は推定によって得られた時間領域の分離信号を示す．また，$e_t(l)$，$e_i(l)$ 及び $e_a(l)$ はそれぞれ時間領域の分離の目的とする成分，非目的（干渉）音源の残留成分及び BSS で生じたその他の人工的な歪み成分を示す．

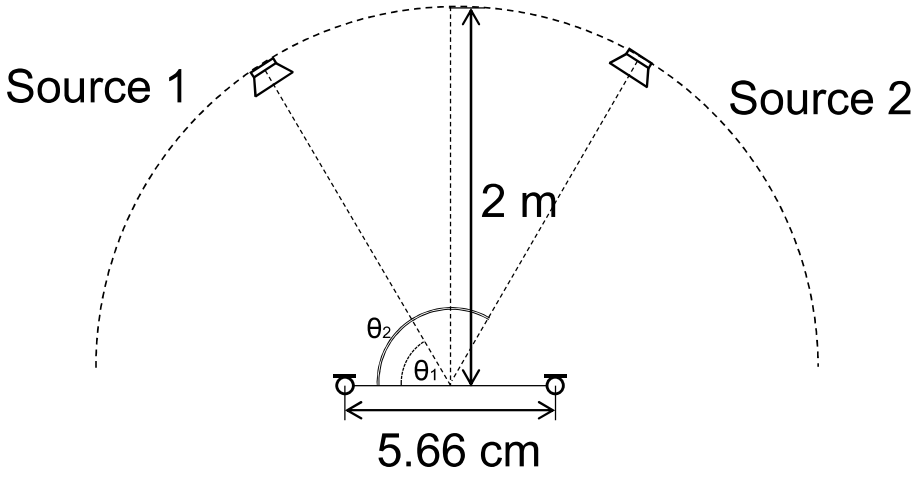


Fig. 4.1. Recording condition of JR2 impulse response.

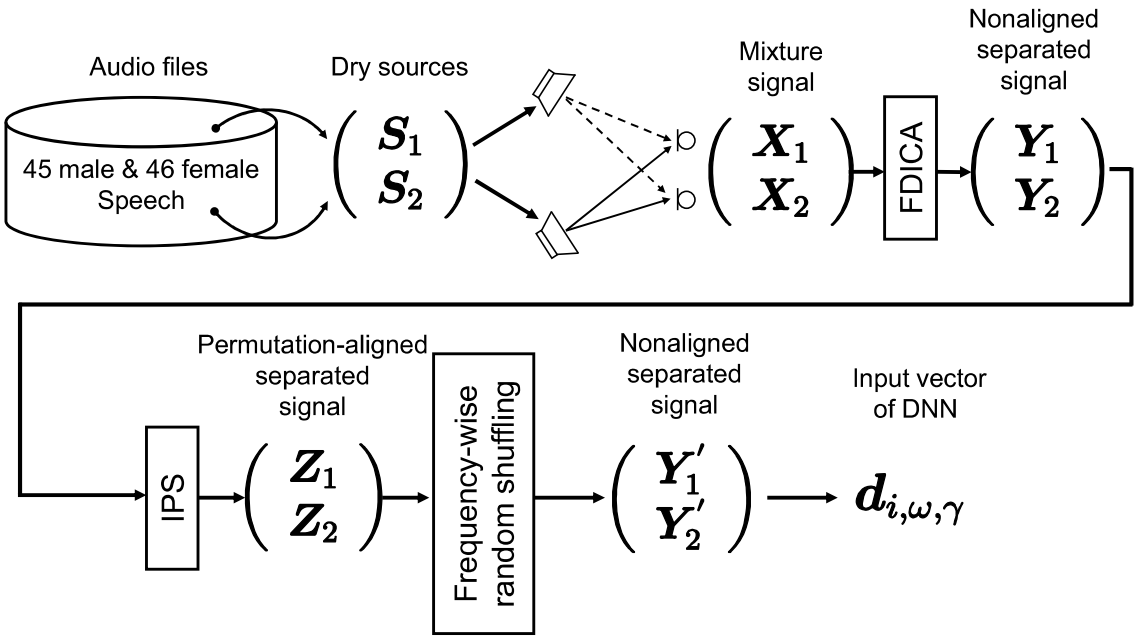


Fig. 4.2. Process flow of producing input vectors for DNN.

テストデータは，Table 4.2 に示すように SiSEC2011 [21] の日本語及び英語の音声信号（男性 4 名及び女性 4 名）8 種類を使用した．これらの音声信号に JR2 インパルス応答を畳み込み 2 チャネル 2 音源の残響付き混合信号を作成した．また，テストデータに畳み込むインパルス応答は，音源方位 $(\theta_1, \theta_2) = (60°, 120°)$，$(\theta_1, \theta_2) = (60°, 100°)$ 及び $(\theta_1, \theta_2) = (70°, 110°)$ の 3 種類を使用した．注意として，学習用データに畳み込んだインパルス応答は，$(\theta_1, \theta_2) = (60°, 120°)$ の 1 種類のみである．テストデータの数は，8 種類の音声から 2 つを選択する組み合わせであり，音源方位毎に 56 種類のデータを作成した．

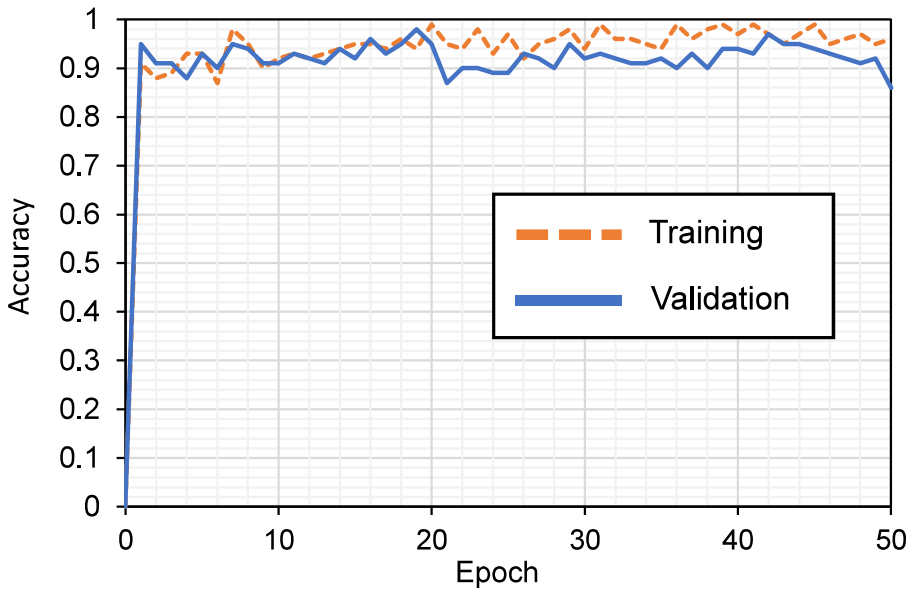


Fig. 4.3. Accuracy curves of DNN for training and validation datasets.

4.3 実験結果

Fig. 4.3 に，学習用データ及び検証用データそれぞれの DNN 正解率を示す．この結果から，DNN は周波数毎の正しいパーミュテーションを約 85% の精度で推定できていることが分かる．つまり，DNN が正しいパーミュテーション推定に失敗する確率は 15% である．しかし，3.6 節及び 3.7.2 項で示した通り，時間軸及び周波数軸に沿った多数決を取ることで，これらの予測誤差の悪影響を大幅に軽減することができる．

Figs. 4.4，4.5 及び 4.6 に，音源到来方向毎の SDR 改善量を示す．各箱ひげ図は，56 個（8 個の音声ファイルの全組み合わせ数）の分離実験結果から生成されている．各箱ひげ図における，四分位範囲（interquartile range: IQR），最大値 b_{max} 及び最小値 b_{min} は以下のように計算される．

$$\mathrm{IQR} = Q_3 - Q_1 \tag{4.3}$$

$$b_{max} = Q_3 + \mathrm{IQR} \times 1.5 \tag{4.4}$$

$$b_{min} = Q_1 - \mathrm{IQR} \times 1.5 \tag{4.5}$$

ここで，Q_3 及び Q_1 は，それぞれ第三四分位数及び第一四分位数を示す．$[b_{min}, b_{max}]$ の範囲外の値は，外れ値として扱った．

既に文献 [13] で報告されているように，どの音源到来方向でも ILRMA の分離性能は 4 dB 以下であるのに対し，IPS を用いた FDICA では約 10 dB 以上の改善を達成している．この結果から，高残響下にある音声混合信号の分離タスクに対して，ILRMA はしばしば分離に失敗している事がわかる．

提案パーミュテーション解決法を適用した FDICA は，いずれも平均的に ILRMA を上回る分離性能を示している．また，提案手法がうまく働いた場合は SDR 改善量が 13 dB に達成するなど，FDICA ベースの分離の上限性能に比較的近い性能も確認できた．しかし，提案 DNN の推定間違いの悪影響が，多数決でも軽減できないほど大きな場合は，並び替えに失敗することもある．この場合，0 dB 以下の SDR 改善量となることが多く，この場合は ILRMA のブロックパーミュテーションの様に中間の周波数でパーミュテーションが反転していることが考えられる．

Fig. 4.7 では，音源到来方向の違いによる提案手法の SDR 改善量の差を比較している．4.2 節でも説明したように学習データに使われている音声の到来方向は $(\theta_1, \theta_2) = (60°, 120°)$ のみであり，$(\theta_1, \theta_2) = (60°, 100°)$ 及び $(\theta_1, \theta_2) = (70°, 110°)$ の到来方向は，学習データに含まれていない．しかし，Fig. 4.7 から分かるように，学習データに存在しない音源到来方法であっても，パーミュテーション解決性能には大きな差が無い．この結果は，提案 DNN が空間情報を学習しているのではなく，文献 [4] のように, ある種の類似度を学習していると解釈できる．実際，DNN の学習データに到来方向のあらゆる組み合わせを準備することは現実的には不可能なことを考えると，音源到来方向に依存しない提案手法は，大きな利点であると考えられる．提案手法はいかなる音源到来方向であっても適用可能であり，一般的な FDICA のポスト処理として扱うことができる．

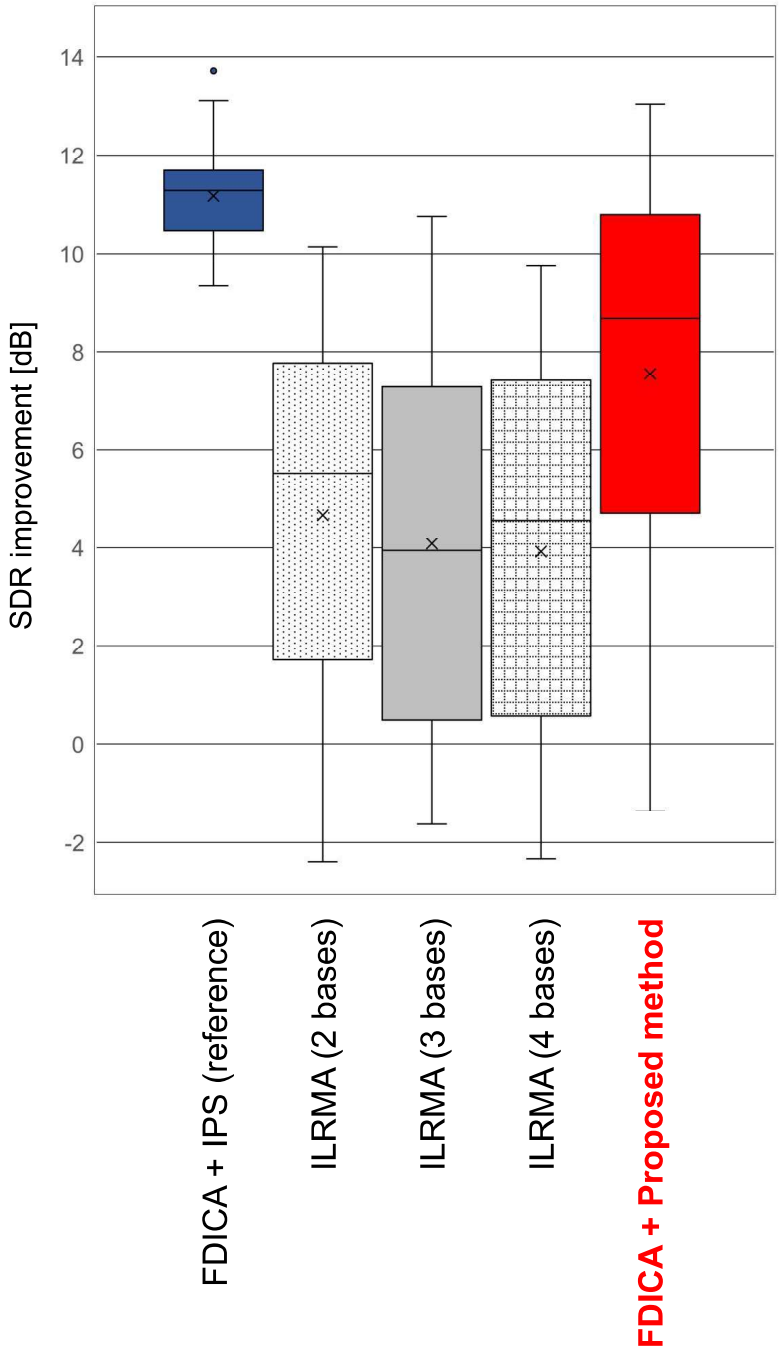


Fig. 4.4. SDR improvements of source directions $(\theta_1, \theta_2) = (60°, 120°)$.

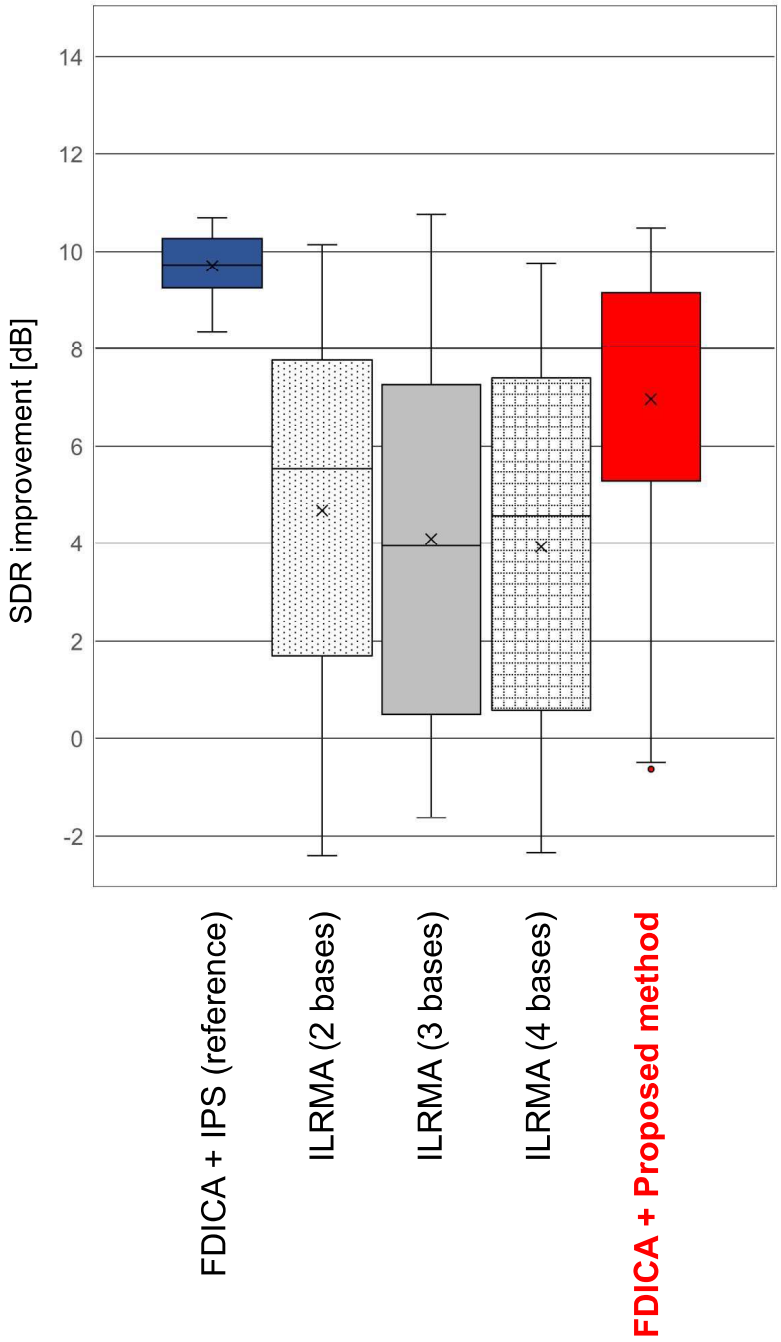


Fig. 4.5. SDR improvements of source directions $(\theta_1, \theta_2) = (60°, 100°)$.

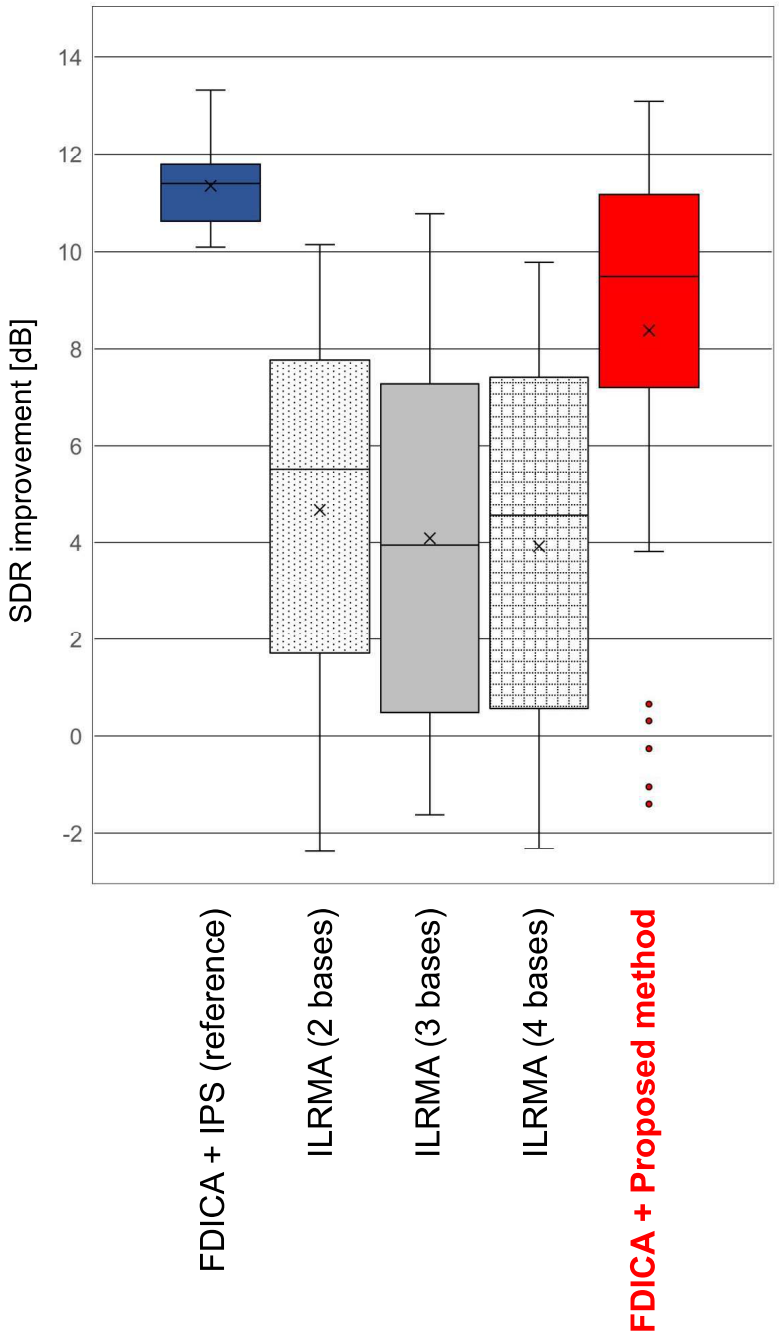


Fig. 4.6. SDR improvements of source directions $(\theta_1, \theta_2) = (70°, 110°)$.

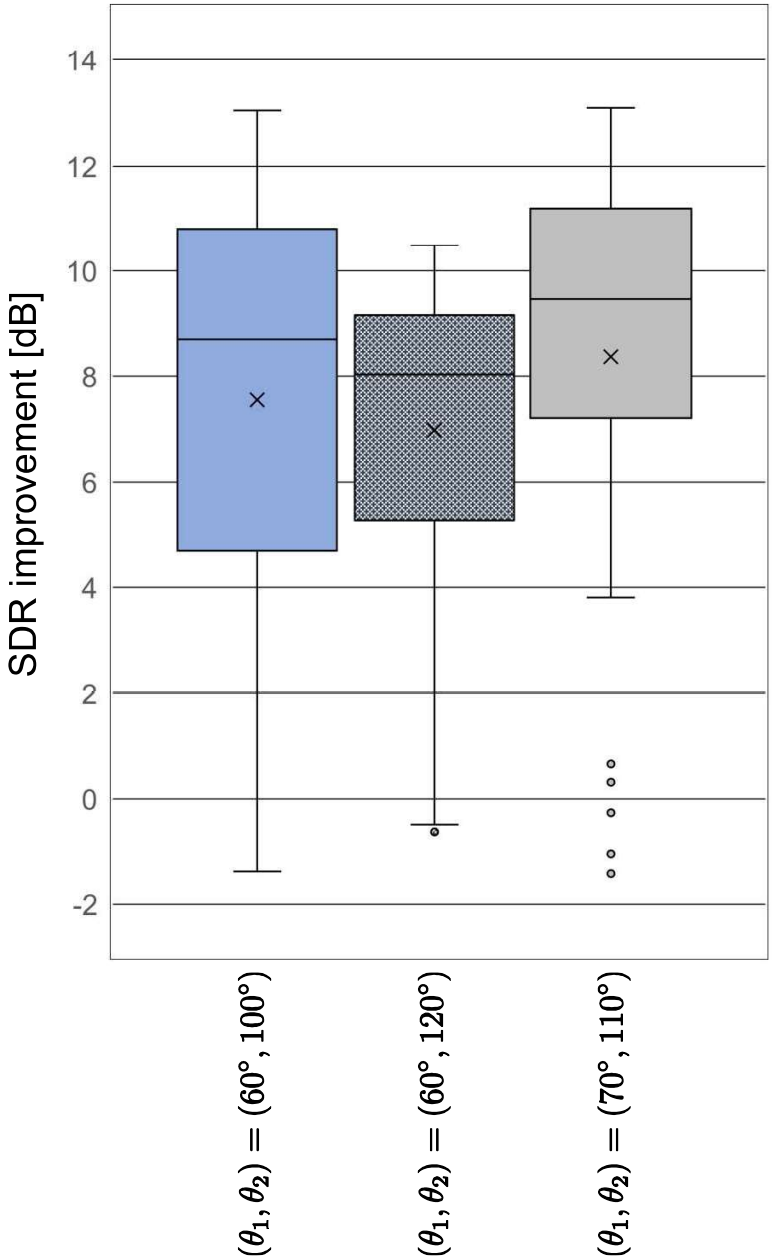


Fig. 4.7. SDR improvement of the proposed method for each direction.

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，高残響の音声混合信号に対して音源分離実験を行い，他手法と比較した．実験の結果，提案手法を適用した FDICA のスコアが平均的に ILRMA を上回る結果となった．また，提案手法は音源の到来方向に依存しないことから，FDICA の一般的なポスト処理として適用可能であることを示した．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．また DNN の推定間違いによる悪影響を軽減するために，信号スペクトルの時間軸及び周波数軸に沿った多数決処理を導入した．実験結果より，提案手法のパーミュテーション解決性能が有効であり，評価指標である SDR 改善量が ILRMA を上回ることを実験的に示した．

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，実行時間や計算量等はあまり考慮されていない．そのため，リアルタイムでの音源分離に適用する場合は，DNN モデル及びパーミュテーションベクトルの計算アルゴリズムを改良する必要がある．また，本論文では音源数が 2 音源のみの場合を扱ったが，より多くの音声を分離したいシーンも考えられる．しかし，単純な $N \geq 3$ の音源数への拡張では，DNN の適用回数が膨大になっていくことから，計算量の関係から現実的ではない．これは，提案手法では，入力された 2 本の周波数アクティベーションが同一音源か否かを推定していることが起因している．そのため，$N \geq 3$ の音源数に対応できるアルゴリズム構築及び DNN の拡張が求められる．拡張の一例として，同じ音源のチャネルインデクスを推定する DNN に変更することが有効であると考える．この拡張案の例を Fig. 5.1 に示す．この場合は，$\tilde{\boldsymbol{g}}_{f_3,2}$ の ch1 は $\tilde{\boldsymbol{r}}_{f_3}$ の ch2 と同じ音源である．同様に，$\tilde{\boldsymbol{g}}_{f_3,2}$ の ch2 は $\tilde{\boldsymbol{r}}_{f_3}$ の ch1 と同じであり，$\tilde{\boldsymbol{g}}_{f_3,2}$ の ch3 は $\tilde{\boldsymbol{r}}_{f_3}$ の ch3 と同じである．そのため，DNN の正解ラベルを（ch2, ch1, ch3）に設定し，DNN の学習を行う．このように，同じ音源であるチャネルインデクスを推定する DNN に拡張することで，DNN の適用回数は変わらずに（DNN 適用回数は多数決の回数）$N \geq 3$ の音源数に対応できると考える．

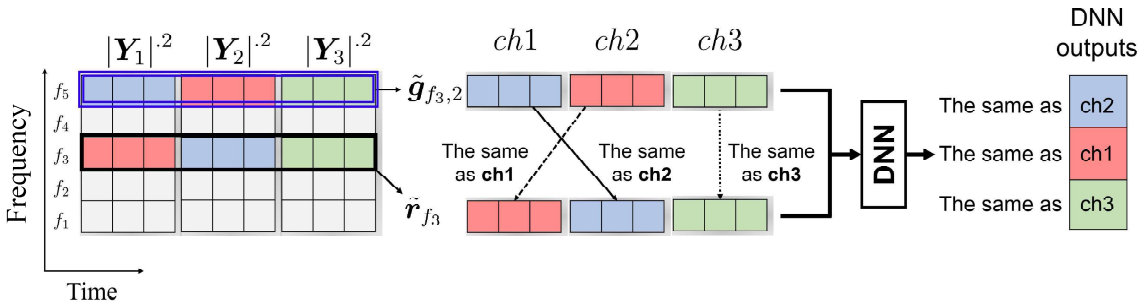


Fig. 5.1. Extension of proposed method.

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助教に心より感謝申し上げます．北村大地助教には，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．急な研究室変更のため，専攻科での 2 年間という短い期間でしたが，あの時の選択は間違ってなかったと実感しております．

本論の副査である村上幸一准教授・雛元洋一助教には，論文の構成や記述に関して大変有益な助言を頂き，大変お世話になりました．ここに厚く御礼申し上げます．

また，北村研究室同期の大島風雅氏，後輩の岩瀬佑太氏・大藪宗一郎氏・渡辺瑠伊氏には，ゼミや日頃のディスカッションのほか，2 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] N. Makishima , S. Mogami , N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. ASLP*, vol. 27, no. 10, pp. 1601–1615, 2019.

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

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

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

[21] 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.

発表文献一覧

査読付き国際会議

1. **Shuhei Yamaji** and Daichi Kitamura, "DNN-based permutation solver for frequency-domain independent component analysis in two-source mixture case," *Proceedings of Asia-pacific signal and information processing association annual summit and conference*, 2020.

国内学会

1. 山地修平, 北村大地, "局所時間周波数構造に基づく深層パーミュテーション解決法," **日本音響学会** *2020* **年春季研究発表会講演論文集**, pp. 317–320, 2020.
2. 山地修平, 北村大地, "局所時間周波数構造に基づく深層パーミュテーション解決法の実験的評価," **日本音響学会** *2020* **年秋季研究発表会講演論文集**, pp. 265–268, 2020.