特別研究論文

（査読済み）

研　究　題　目

深層パーミュテーション解決法に基づくブラインド音源分離

提出年月日		2025 年　1 月　27 日
氏	名	蓮池 郁也
主	査	北村 大地 准教授
副	査	柿元 健 准教授
副	査	重田 和弘 教授

香川高等専門学校
専攻科
創造工学専攻

NATIONAL INSTITUTE OF TECHNOLOGY KAGAWA COLLEGE
香川高専

Blind Source Separation Based on Deep Permutation Solver

Fumiya Hasuike

Advanced Course in Industrial and Systems Engineering
National Institute of Technology, Kagawa College

Abstract

Blind source separation (BSS) is a technique to estimate individual source signals from observed mixed signals without prior information. It is widely applied in fields such as speech recognition, audio signal processing, and noise reduction. Frequency-domain independent component analysis (FDICA), a key approach to BSS, estimates sources by assuming independence in each frequency bin. However, FDICA faces the permutation problem, where the order of signal components varies across frequency bins. To address this issue, various permutation solvers (PS) have been proposed, including methods based on source direction of arrival and frequency correlations. Since then, methods based on source models have been widely proposed as BSS to avoid permutation problem, but it is difficult to construct a universal source model that can be applied to a wide range of sources. Therefore, deep PS (DPS), which is based on deep neural networks (DNN), is being considered for PS with high generalization performance, rather than constructing a universal sound source model. However, conventional DPS has a complex algorithm and limited accuracy in source separation, as shown in the experimental results. Therefore, I propose a new DPS based on a bidirectional long short-term memory. The proposed method aims to solve the permutation problem by learning the relationship between the frequency directions of sources. The proposed DPS has the advantage of achieving generalization to various sources using only a few seconds of training data, while the use of DNN generally entails high training costs. To evaluate the performance of the proposed method, comparative experiments were conducted against conventional methods. The results demonstrated that the proposed DPS, trained with music signals, could partially solve the permutation problem in speech signals. On the other hand, the performance of the proposed method is still less than that of conventional methods based on sound source models. Nevertheless, it is meaningful that the proposed DPS was able to solve the permutation problem in speech signals using only two types of music signals, and further performance improvement is expected by introducing new machine learning models in the future.

Key Words: frequency-domain independent component analysis, deep neural networks, bidirectional long short-term memory, one-shot learning, permutation alignment

概要

ブラインド音源分離（blind source separation: BSS）は，観測信号から混合前の音源信号を事前情報なしに推定する技術であり，音声認識，音響信号処理，雑音除去など幅広い分野で応用されている．BSS の主要なアプローチである周波数領域独立成分分析（frequency-domain independent component analysis: FDICA）は，周波数ビンごとに独立性を仮定して音源を分離するが，その過程で信号成分の順序が周波数ごとに異なる「パーミュテーション問題」が生じる．このパーミュテーション問題を解決するために，音源の到来方向情報を用いたパーミュテーション解決法（permutation solver: PS）や周波数間相関に基づく PS 等が提案されている．その後，パーミュテーション問題をできるだけ回避する BSS として，音源モデルに基づく手法が広く検討されてきたが，幅広い音源に適合する万能な音源モデルの構築は困難である．そこで，万能な音源モデルではなく可能な限り汎化性能の高い PS の実現を目的として，深層ニューラルネットワーク（deep neural network: DNN）に基づく PS（deep PS: DPS）が検討されている．しかしながら，既存の DPS は，アルゴリズムが複雑であることに加えてパーミュテーション問題を完璧には解決できていない．そこで，本論文では，双方向長短期記憶ネットワークを基盤とする新しい DPS を提案する．提案手法は，音源の周波数方向における関係性を学習することで，パーミュテーション問題を解決することを目指している．また，一般的に DNN の活用には学習コストが大きい問題があるが，提案 DPS は数秒程度の学習データサンプルのみで種々の音源に適用できる汎化性能を獲得できる利点がある．提案手法の性能を評価するため，従来手法との比較実験を実施した．実験の結果，提案手法が省サンプルの音楽信号で学習した DPS を用いて，音声信号のパーミュテーション問題を良好な精度で解決できることが確認された．一方で，音源モデルに基づく従来手法と比較すると，性能面では依然として課題が残ることが明らかとなった．それでもなお，省サンプルの音楽信号を用いて音声信号のパーミュテーション問題を解決できた点には意義があり，今後，新たな機械学習モデルを導入することでさらなる性能向上が期待される．

目次

第 1 章

緒言

1.1　本研究の背景

音源分離は，観測された混合信号から，各音源の独立した信号を推定する技術である．この技術は，音響信号処理において重要な研究テーマであり，Fig. 1.1 に示すように，その応用範囲は広範にわたる．音源分離技術は，音声信号の処理をはじめとする多くの実用的なタスクに適用されている．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，スマートスピーカーや音声認識システムの普及に伴い，音声信号処理の性能向上が求められている．例えば，雑音や非目的話者の音声が混入した場合でも，目的話者の音声を正確に抽出する能力が製品の品質向上に直結する．また，ノイズキャンセリングイヤホンや補聴器における音声強調機能の実現にも音源分離技術が用いられている．これらの要因から，音源分離技術の需要はますます高まっている．

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

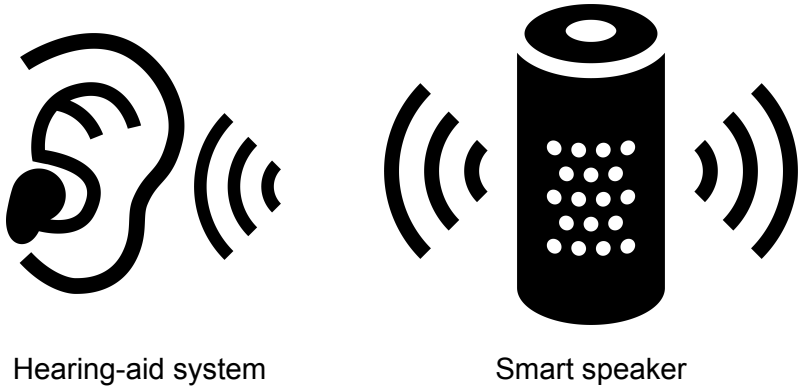

Fig.1.1. Examples of application using speech source separation.

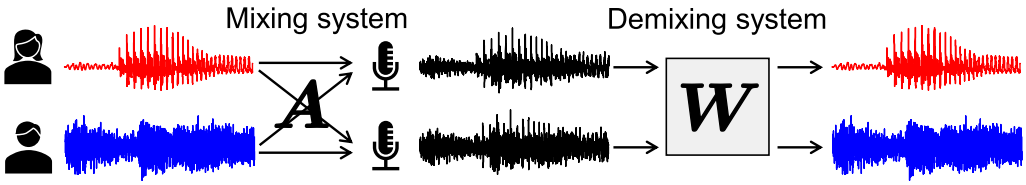

Fig.1.2. Overview of BSS in two-source case.

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間で不揃い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」と呼ばれている．この問題は，観測マイクロホンの数よりも音源数が多い劣決定条件下にも応用できる強力な BSS のフルランク共分散分析 [4] においても生じる典型的な問題である．このパーミュテーション問題に対して，これまでに，様々なパーミュテーション解決法（permutation solver: PS）が提案されてきた．具体的には，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 PS [5, 6, 7]，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て行う PS [8]，及びその両者を組み合わせた PS [9] が提案されている．また，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も提案されている．例えば，独立ベクトル分析（independent vector analysis: IVA）[10, 11, 12, 13] 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）[14] 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）[15, 16] 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．これらのブラインドな音源分離手法をもっ

てしても，特定のまとまった帯域で分離信号成分の順序を間違えるブロックパーミュテーション問題が生じることが報告されている [17, 18, 19]．このブロックパーミュテーション問題の解決法も研究されており，ユーザとのインタラクションを用いる手法 [18] や音源モデルを用いてコスト行列を設計しハンガリー法 [20] を適用する手法 [19, 21] が提案されている．

1.2 本研究の目的

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．近年では，パーミュテーション問題を解決するために，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 PS [6] をもとに，多様体最適化を用いて二重確率行列（doubly stochastic matrix: DSM）[22] で表現される周波数帯域間の相関を精密に最適化し，パーミュテーション問題の解決を高精度に実現する手法 [7] やサブバンドと呼ばれる局所帯域毎に従来の BSS 手法を逐次適用し，隣接帯域の分離結果を初期値として利用することで，周波数間の順序整合性を高める手法 [23] が提案されている．しかしながら，上記のいずれの手法を用いても，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．

そこで，様々な混合信号における頑健・高精度なパーミュテーション解決法を構築することを目的として，深層ニューラルネットワーク（deep neural networks: DNN）を用いた手法 [24] が提案されている．これは，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である．以後，DNN を用いてパーミュテーション問題を解決する手法を深層パーミュテーション解決法（deep PS: DPS）と呼ぶ．従来の DPS [24] では，高い分離性能を達成するために，複雑なモデル構造を必要としており，また，音源数が増加すると計算コストが増大するという課題がある．そこで，本論文では，新たな DPS について提案し，FDICA に適用して得られる分離信号に対して，提案 DPS を用いた際の分離性能について評価する．DPS は，少ない学習データで，汎化性能の高いモデルを構築する可能性を秘めており，万能な音源モデルの実現に向けた重要なアプローチである．汎化性能とは，学習データに過度に依存せず，未知のデータに対しても同様の性能を発揮できるモデルの能力を指す．特に，第 4 章で詳しく説明するが，音楽信号で学習した DPS を音楽信号だけでなく音声信号にも適用する実験を行うことで，幅広い音源環境に対応可能なモデルの構築を探求している．このアプローチは，学習コストの増大を回避するという観点から，実用化において重要な意義を持つ．

これまでに提案されてきた BSS とそれに伴う PS を Table1.1 に示す．本論文では，FDICA の後段に接続する PS の性能改善のみに焦点を当てており，分離信号成分の正しいパーミュテーションを予測する様に学習した DNN を用いてパーミュテーション問題を解決することを目的とする．そのため，FDICA を適用した際に得られる推定分離信号に含まれる分離誤差の改善は本論文の目的ではない．したがって，本論文で提案する DPS は，IVA や ILRMA の後段に接続することも可能な手法である．さらに，今後より高精度な BSS が提案された場合に

Table 1.1. Comparison of BSS and PS

Method	Year	PS	Principle of PS
FDICA [3]	1998	None (required)	None
FDICA+COR [6]	2007	Post-process	Based on frequency correlation in each source
FDICA+DOA [8]	2006	Post-process	Difference of DOAs in each source
FDICA+DPS [24]	2020	Post-process	Based on frequency correlation in each source trained by DNN
IVA [13]	2011	Unified	Assuming source model based on co-occurrence of all frequency components
ILRMA [15]	2016	Unified	Assuming source model based on low-rank time-frequency structure
Subband splitting ILRMA [23]	2024	Unified	Applying ILRMA to local frequency components with overlap

も，その後段に活用することが可能である．

1.3　本論文の構成

まず，第 2 章では，本論文の解決すべき課題であるパーミュテーション問題の説明に必要となる ICA の基本原理や音響信号の時間周波数領域への変換である短時間 Fourier 変換（short-time Fourier transform: STFT）に加え，パーミュテーション問題を可能な限り回避する BSS の IVA 及び ILRMA，そして既存の DPS について詳しく説明する．これらは，いずれも提案手法の説明に必要となる知識である．第 3 章では，本論文の提案手法である DPS 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．第 4 章では，実際に残響を含む混合信号を FDICA に適用して得られる分離信号に対して音源分離実験を行い，提案 DPS の性能の検証を行う．最後に第 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

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，BSS の基礎理論となる ICA について説明する．2.3 節では，音響信号処理でよく用いられる STFT について説明する．2.4 節では，時間周波数領域で周波数毎に ICA を適用する FDICA について説明する．2.5 節では，本論文の主題として，FDICA に付随するパーミュテーション問題の説明と，既存のパーミュテーション解決法について説明する．2.6 節では，パーミュテーション問題を可能な限り回避する BSS の IVA 及び ILRMA について詳細を述べる．2.7 節では，既存の DPS とその問題について説明する．2.8 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，本章では音源数及びマイクロホン数を，それぞれ N 及び M として説明する．但し，後述の通り，音源数とマイクロホン数は常に等しいという仮定が必要である．BSS の文脈では，このような「音源数がマイクロホン数以下」という条件を優決定条件と呼ぶ．

2.2.1　信号源の混合モデルと分離方法

今，N 個の音源を M 個のマイクロホンで観測するという状況を考える．信号源及び観測信号はそれぞれ次式のように示される．

$$\boldsymbol{s}(l) = \left[s_1(l),\, s_2(l),\, \cdots,\, s_N(l)\right]^{\mathrm{T}} \in \mathbb{R}^N \tag{2.1}$$

$$\boldsymbol{x}(l) = \left[x_1(l),\, x_2(l),\, \cdots,\, x_M(l)\right]^{\mathrm{T}} \in \mathbb{R}^M \tag{2.2}$$

ここで，$l = 1, 2, \cdots, L$，$n = 1, 2, \cdots, N$，及び $m = 1, 2, \cdots, M$ は，それぞれ離散時間，音源及びマイクロホンのインデクスを示す．また，$\cdot^{\mathrm{T}}$ はベクトルや行列の転置を表す．もし，複

数の音源信号の混合現象が，離散時間のサンプル毎の混合（瞬時混合）として表現できるならば，音源と観測信号の関係は次式でモデル化できる．

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

ここで，$\boldsymbol{A} \in \mathbb{R}^{M\times N}$ は $M \times N$ の混合行列であり，次式で示される．

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{bmatrix} \tag{2.4}$$

ここで，a_{mn} は音源 n からマイクロホン m への伝搬を表す係数である．音源やマイクロホンの位置が固定であると仮定すれば，a_{mn} は時間に依存しない．分離行列を $\boldsymbol{W} \in \mathbb{R}^{N\times M}$ と定義すると，分離信号 $\boldsymbol{y}(l) \in \mathbb{R}^{N}$ を次式で表せる．

$$\boldsymbol{y}(l) = \boldsymbol{W}\boldsymbol{x}(l) \tag{2.5}$$

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．

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

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を導入することで，分離行列 $\boldsymbol{W}$ 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重要な概念が，統計的独立性である．各音源信号 $s_n(l)$ は互いに無関係であると考えられるため，N 個の信号源 $\boldsymbol{s}(l)$ を確率変数としてみなすと，その同時分布は各単独の分布の積で表される．

$$p\big(s_1,\, s_2,\, \ldots,\, s_N\big) = \prod_{n=1}^{N} p\big(s_n\big) \tag{2.6}$$

一方で，理想的な分離行列 $\boldsymbol{W}$ が求まり，分離信号 $\boldsymbol{y}(l)$ を得た時にも，同様に次式が成立すると期待される．

$$p\big(y_1,\, y_2,\, \ldots,\, y_N\big) = \prod_{n=1}^{N} p\big(y_n\big) \tag{2.7}$$

したがって ICA による BSS は，式 (2.7) のように分離信号同士が統計的に独立となるような分離行列 $\boldsymbol{W}$ を推定する問題とみなすことができる．これを定式化すると，次式のような最適化問題に帰着できる．

$$\underset{\boldsymbol{W}}{\text{minimize}} \quad \mathfrak{I}(\boldsymbol{W}) \tag{2.8}$$

$$\mathfrak{I}(\boldsymbol{W}) = \mathfrak{D}_{\mathrm{KL}}\left(p(\boldsymbol{y}) \mid \prod_{n=1}^{N} p(y_n)\right) \tag{2.9}$$

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

$$\mathfrak{D}_{\mathrm{KL}}[p(s)|q(s)] = \int p(s)\log\frac{p(s)}{q(s)}ds \tag{2.10}$$

また式 (2.5) より，観測信号 $\boldsymbol{x}$ と分離信号 $\boldsymbol{y}$ との間には次式が成立する．

$$p(\boldsymbol{y}) = \frac{1}{\left|\det \boldsymbol{W}\right|}\, p(\boldsymbol{x}) \tag{2.11}$$

この関係を利用して最適化問題を解くことで，ICA による信号源分離が実現される．

2.2.3 ICA における任意性

前項より，分離信号 $\boldsymbol{y}(l) = \left[y_1(l), y_2(l), \cdots, y_N(l)\right]^{\mathrm{T}} \in \mathbb{R}^N$ の独立性を最大化する分離行列 $\boldsymbol{W}$ 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 (2.9) 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 $y_n(l)$ 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 (a) より，元々の信号源の順序が入れ替わる可能性がある．また，任意性 (b) より，分離信号のスケールが混合前の音源信号のスケールから変化してしまう可能性がある．なお，信号のスケールの任意性に関しては，プロジェクションバック（projection back: PB）法 [25] と呼ばれる解析的な補正方法が提案されており，2.4 節で説明する FDICA においても大きな問題にはならない．一方，順序の任意性は FDICA におけるパーミュテーション問題を引き起こす要因となる．

2.3 STFT

通常の BSS では，2.2 節で述べた時間領域の混合・分離モデルではなく，時間周波数領域での混合・分離モデルを仮定する．その動機や詳細は 2.4 節で述べる．本節では一般的な音響信号の時間周波数変換である STFT について，その詳細を示す．

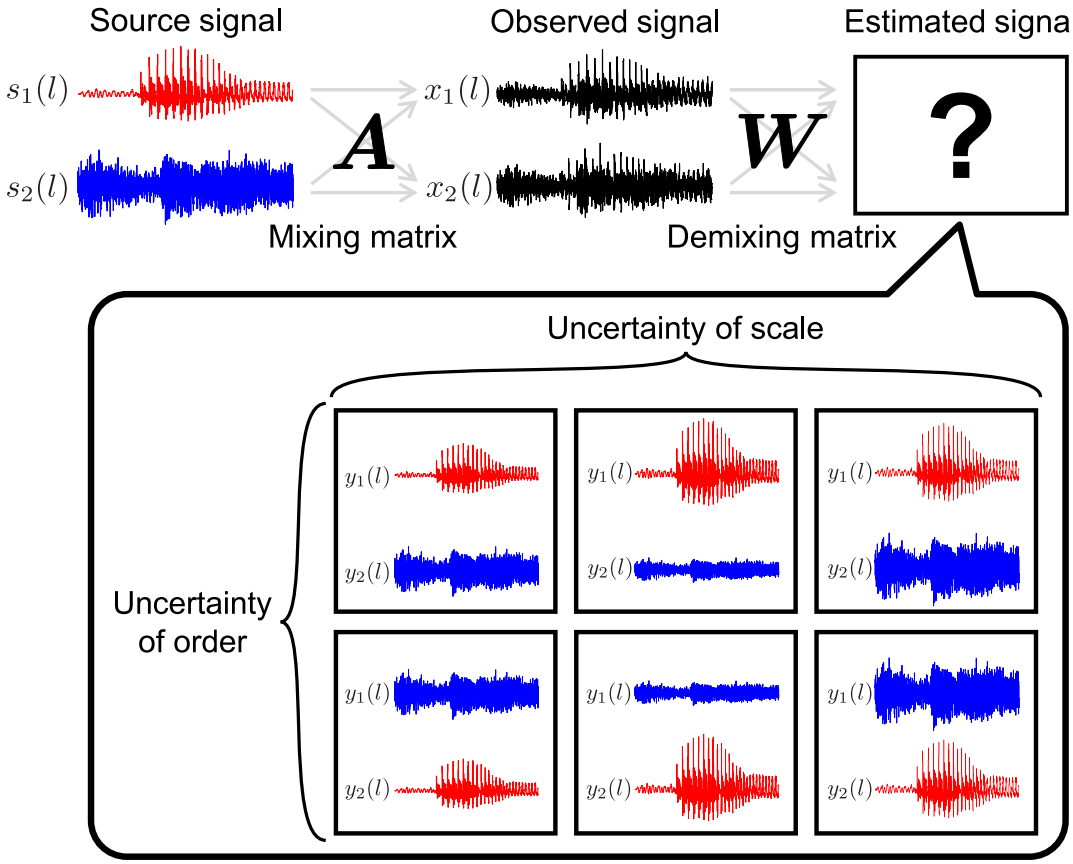

Fig.2.1. Uncertainty in ICA. ICA cannot determine order and scales of estimated signals ($N = 2$).

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．

$$\boldsymbol{x} = \left[x(1),\, x(2),\, \ldots,\, x(l),\, \ldots,\, x(L)\right]^{\mathrm{T}} \in \mathbb{R}^{L} \tag{2.12}$$

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 $\boldsymbol{x}$ 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．

$$\begin{aligned}\boldsymbol{x}^{(j)} &= \left[x\left((j-1)\tau+1\right), x\left((j-1)\tau+2\right), \cdots, x\left((j-1)\tau+Q\right)\right]^{\mathrm{T}} \\ &= \left[x^{(j)}(1), x^{(j)}(2), \cdots, x^{(j)}(q), \cdots, x^{(j)}(Q)\right]^{\mathrm{T}} \in \mathbb{R}^{Q}\end{aligned} \tag{2.13}$$

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．

$$J = \frac{L}{\tau} \tag{2.14}$$

ただし，時間領域の信号 $\boldsymbol{x}$ は式 (2.13) 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，窓関数 $\boldsymbol{\omega} = [\omega(1), \omega(2), \cdots, \omega(q), \cdots, \omega(Q)] \in \mathbb{R}^{Q}$

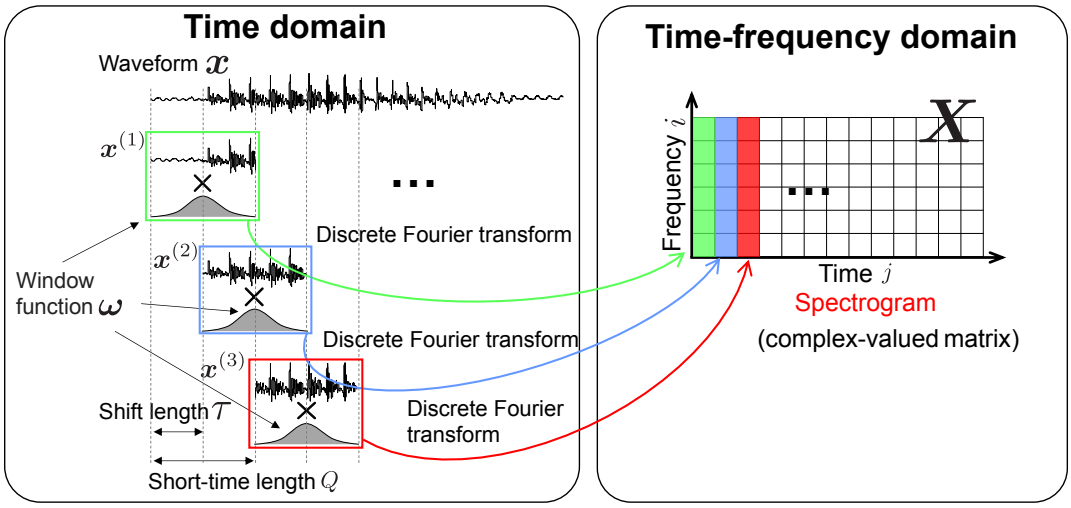

Fig.2.2. Mechanism of STFT. Each of windowed short-time signals are transformed to frequency domain by discrete Fourier transform.

を用いた信号 $\boldsymbol{x}$ の STFT を次式で表す．

$$\boldsymbol{X} = \mathrm{STFT}_{\boldsymbol{\omega}}(\boldsymbol{x}) \ \in \mathbb{C}^{I\times J} \tag{2.15}$$

ここで，$\boldsymbol{X}$ 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 $\boldsymbol{X}$ の (i,j) 番目の要素は次式で表される．

$$x_{ij} = \sum_{q=1}^{Q} \omega(q) x^{(j)}(q) \exp\left\{\frac{-\iota 2\pi(q-1)(i-1)}{F}\right\} \tag{2.16}$$

ここで F は $\lfloor \frac{F}{2} \rfloor + 1 = I$ を満たす整数（$\lfloor\cdot\rfloor$ は床関数），$i = 1, 2, \cdots, I$ は周波数ビンのインデクス，ι は虚数単位をそれぞれ示している．また，窓関数 $\boldsymbol{\omega}$ は短時間信号 $\boldsymbol{x}^{(j)}$ の両端の不連続性を解消するための解析窓関数である．このように STFT は，時間領域の信号を一定幅の短時間信号に分割して解析窓関数を乗じて離散フーリエ変換を適用し，スペクトログラムと呼ばれる複素時間周波数行列に変換する処理である．音源分離等の多くの音響信号処理では，このスペクトログラムを信号処理の対象とする．

2.4 FDICA

本節以降，音源数と観測チャネル数（マイクロホン数）をそれぞれ N 及び M とする．また，音源信号，観測信号，及び分離信号の時間周波数毎の成分をそれぞれ次式で表す．

$$\boldsymbol{s}_{ij} = [s_{ij1}, s_{ij2}, \cdots, s_{ijn}, \cdots, s_{ijN}]^{\mathrm{T}} \ \in \mathbb{C}^{N} \tag{2.17}$$

$$\boldsymbol{x}_{ij} = [x_{ij1}, x_{ij2}, \cdots, x_{ijm}, \cdots, x_{ijM}]^{\mathrm{T}} \ \in \mathbb{C}^{M} \tag{2.18}$$

$$\boldsymbol{z}_{ij} = [z_{ij1}, z_{ij2}, \cdots, z_{ijn}, \cdots, z_{ijN}]^{\mathrm{T}} \ \in \mathbb{C}^{N} \tag{2.19}$$

式 (2.17)–(2.19) 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．すなわち，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．

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

$$\tilde{\boldsymbol{x}}(l) = \sum_n \sum_{l'=0}^{R_T-1} \tilde{\boldsymbol{a}}_n(l')\tilde{s}_n(l-l') \tag{2.20}$$

ここで，$\tilde{\boldsymbol{x}}(l) = [\tilde{x}_1(l), \tilde{x}_2(l), \cdots, \tilde{x}_M(l)]^{\mathrm{T}}$ 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，$\tilde{\boldsymbol{a}}_n(l)$ は音源 n に対する畳み込み混合係数ベクトル（n 番目の音源から全マイクロホンまでのインパルス応答を時間 l 毎にまとめたもの）である．式 (2.20) のように混合される複数の音源を分離するためには，分離行列ではなく逆畳み込みフィルタを推定することが必要となる．一般的に逆畳み込みフィルタの推定非常に困難な問題となってしまうことから，時間領域での ICA による BSS は容易ではない．この問題を解決するために，各信号の STFT による時間周波数表現を用いて，式 (2.20) の時間領域における畳み込み混合を，時間周波数領域での周波数毎の瞬時混合に変換し，時間周波数領域で周波数毎に ICA を行う FDICA が提案された．

FDICA では，周波数毎の時不変な混合行列 $\boldsymbol{A}_i = [\boldsymbol{a}_{i1}\ \boldsymbol{a}_{i2}\ \cdots\ \boldsymbol{a}_{in}\ \cdots\ \boldsymbol{a}_{iN}] \in \mathbb{C}^{M\times N}$ を定義し，混合信号が次式で表現できると仮定する．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{2.21}$$

式 (2.21) をスペクトログラムを用いてモデル化すると，Fig. 2.3 のように表現される．この混合モデルは，観測信号収録時の残響長 R よりも STFT の短時間区間長 Q が十分長い場合に成立する．以後，決定的な系（$M = N$）を仮定すると，混合行列 $\boldsymbol{A}_i$ が正則であれば，周波数毎の分離行列 $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = [\boldsymbol{w}_{i1}\ \boldsymbol{w}_{i2}\ \cdots\ \boldsymbol{w}_{in}\ \cdots\ \boldsymbol{w}_{iN}]^{\mathrm{H}} \in \mathbb{C}^{N\times M}$ を用いて，分離信号を次式で表せる．

$$\boldsymbol{z}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{2.22}$$

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．式 (2.22) をスペクトログラムを用いてモデル化すると，Fig. 2.4 のように表現される．分離行列の行ベクトルである $\boldsymbol{w}_{in} \in \mathbb{C}^M$ は，i 番目の周波数ビンにおいて，観測信号から n 番目のみの音源が含まれる分離信号へ変換する分離フィルタである．このように FDICA では，観測信号 $\boldsymbol{x}_{ij}$ の各周波数ビンに対しそれ

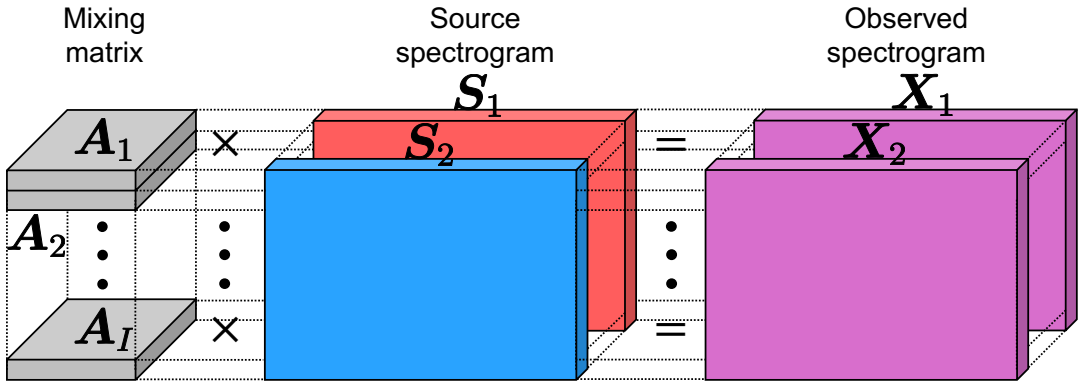

Fig.2.3. Mixing model in time-frequency domain ($N = M = 2$).

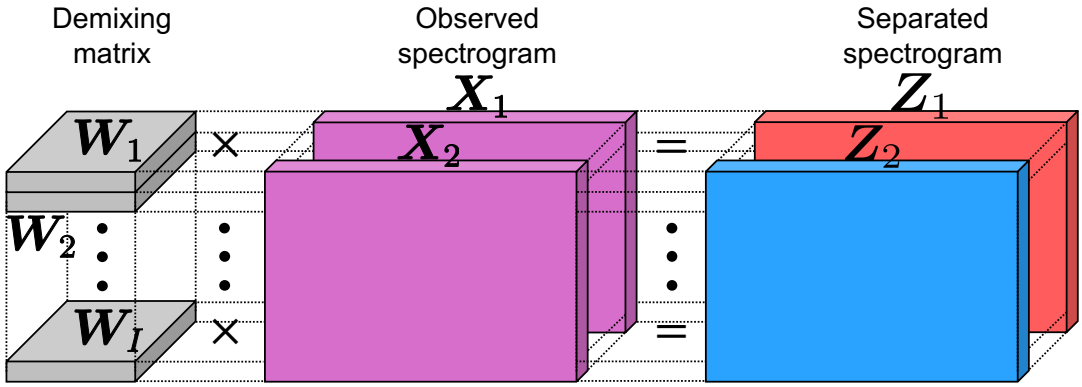

Fig.2.4. Demixing model in time-frequency domain ($N = M = 2$).

ぞれ独立に（複素数の）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が実現される．

2.5　パーミュテーション問題とその解決

FDICA 中で周波数毎に適用している ICA は，2.2.3 項で述べた通り，分離された推定信号の周波数毎のスケール及び順番に関しては不定である．従って，FDICA の推定分離行列を $\hat{\boldsymbol{W}}_i$ とすると，次式のような不定性が残る．

$$\hat{\boldsymbol{W}}_i = \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{W}_i \tag{2.23}$$

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 $\boldsymbol{w}_{in}$ 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．例えば，$N = M = 2$ の場合は

$$\boldsymbol{P}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{2.24}$$

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

$$\boldsymbol{P}_i = \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}, \begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}, \begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix}, \begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix}, \begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix}, \text{or } \begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix} \tag{2.25}$$

の 6 種類がパーミュテーション行列である．一方，$\boldsymbol{D}_i \in \mathbb{R}^{N\times N}$ は，$\boldsymbol{w}_{in}$ のスケールを変化させる可能性のある対角行列である．従って，FDICA で推定される分離信号

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

$$= [y_{ij1}, y_{ij2}, \cdots, y_{ijn}, \cdots, y_{ijN}]^{\mathrm{T}} \quad \in \mathbb{C}^N \tag{2.27}$$

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，時間領域での ICA の場合と同様に PB 法 [25] で解析的に復元可能である．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 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ぞれで N 個の音源の順番は $N!$ 種類あるため，全周波数のパーミュテーションは $(N!)^I$ 通り存在することになり，その内の正解（全周波数で同一の音源パーミュテーションとなるもの）は $N!$ 個である．この問題は，一般的にパーミュテーション問題と呼ばれる．パーミュテーション問題の概要を Fig. 2.5 に示す．ここで，FDICA で得られる（パーミュテーション問題が生じている状態の）推定信号 $\boldsymbol{y}_{ij}$ の n 番目のスペクトログラムを $\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$ と定義している．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，異なる音源の分離成分が 1 つの時間周波数内に混在していることが分かる．これがパーミュテーション問題であり，ICA の分離信号の順番に関する不定性に起因して発生している．そのため，FDICA にはポスト処理として，分離された音源の順番を全周波数ビンにわたって正しく並べ直す必要がある．

このパーミュテーション問題を解決する処理は次式で表される．

$$\boldsymbol{z}_{ij} = \boldsymbol{P}_i^{-1} \boldsymbol{D}_i^{-1} \boldsymbol{y}_{ij} \tag{2.28}$$

前述の通り，スケールの不定性を補正する $\boldsymbol{D}_i^{-1}$ は PB 法 [25] によって解析的に求められる．従って，パーミュテーション問題の解決とは，全周波数ビンにわたって $\boldsymbol{P}_i^{-1}$ を求める問題として解釈できる．このパーミュテーション問題を解決するために，これまでにも数々の PS 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 PS [5, 6] がある．これは，Fig. 2.5 に示す赤色と青色の分離信号のように，分離信号のパーミュテーションが正しければ，隣接した周波数アクティベーション間の相関が高くなりやすいという仮定の下で並べ替える手法である．このとき，離れた周波数においても，同じ音源のアクティベーション間の相関が高くなるように並び替えられている．他にも，マイクロホンの相対的な位置情報を既知として音源到来方位を計算し，パーミュテーション解決の手掛かりとする PS [8] 及び音源到来方位と周波数毎の時系列強度の両者を組み合わせた PS [9] も提案されている．

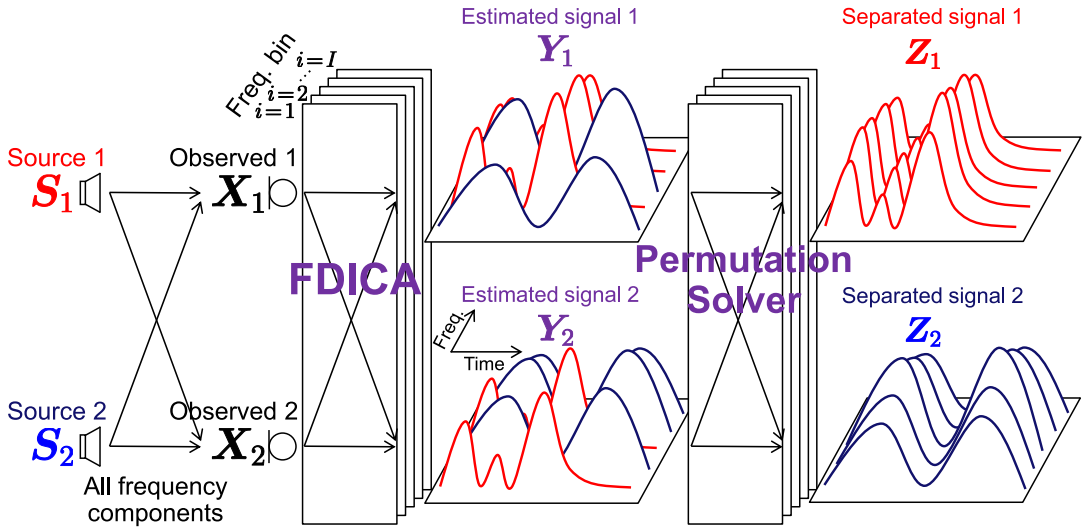

Fig.2.5. Permutation problem in FDICA ($N = 2$).

上記の PS に加え，パーミュテーション問題をより効率的かつ高精度に解決するための新たな手法として，マニフォールド最適化 [26] を用いたアプローチが提案されている [7]．この手法では，従来の手法が持つ制約を克服する工夫がなされている．文献 [6] では，隣接周波数間の相関を利用してパーミュテーションを整列する際に，周波数ごとのパワー比率を平均化して扱っていた．一方で，文献 [7] では，パワー比率の全ペアを直接考慮する厳密な目的関数を導入し，これにより周波数ビン間の関係性を詳細に評価することが可能となっている．さらに，この目的関数を解くために，パーミュテーション行列を DSM として緩和し，マニフォールド最適化を適用している．これにより，従来のような組み合わせ最適化ではなく，勾配法による効率的な解法が実現されている．具体的には，勾配を用いてパーミュテーション行列を徐々に最適化し，最終的にはハンガリー法 [20] を用いて DSM から厳密なパーミュテーション行列を取得する手法を採用している．

2.6 IVA 及び ILRMA

FDICA に対して音源の時間周波数成分の共起関係を新たに仮定して，パーミュテーション問題を可能な限り回避しつつ分離信号を推定する手法が登場している．例えば，IVA [12, 13] は，同一音源の周波数成分の共起を仮定しており，FDICA では周波数毎に独立性を最大化していたのに対し，IVA では全周波数成分をまとめてベクトル変数とし，ベクトル間の独立性を最大化するようなモデルとなっている．そのため， 実際に複数の周波数ビンで同時に共起する成分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．Fig. 2.6 に示すように，IVA は「同一音源であれば全周波数が共起する」という仮定を用いて，音源信号の時間周波数構造に関するモデルを構築する．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

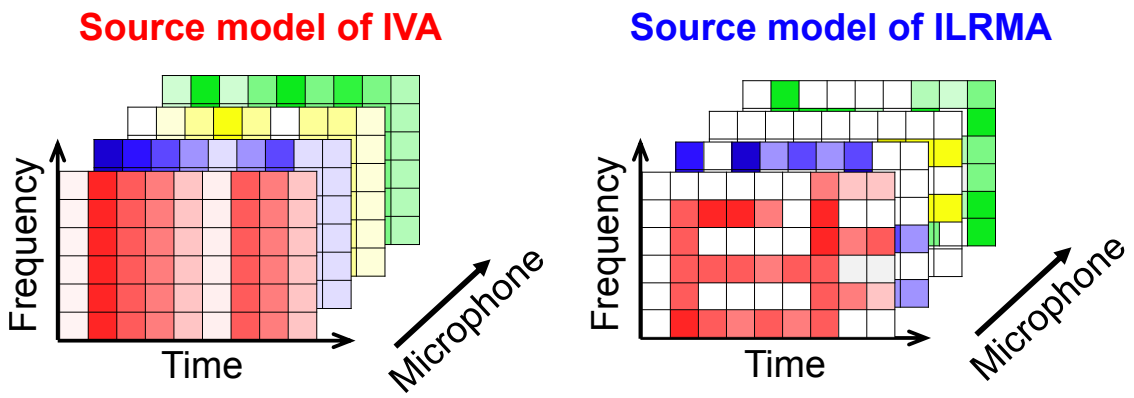

Fig.2.6. Source model of IVA and ILRMA ($N = 4$).

である程度パーミュテーション問題を回避できる．さらに，IVA の音源信号の時間周波数構造に関するモデル（以後，音源モデルと呼ぶ）をより詳細なモデルに発展させた BSS として，ILRMA [15, 16] が提案されている．ILRMA は，IVA で提案された音源モデルに NMF [14] 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おり，Fig. 2.6 に示すように，音源信号の時間周波数構造を低ランク行列で表現するような音源モデルを構築している．このような音源モデルは音声信号だけでなく音楽信号にもよく適合することから，ILRMA の登場によって多くの場合において IVA よりも高品質な BSS を達成することができるようになった．

しかし，声質の近い複数音声の混合や，音源数が $N \geq 4$ となる過酷な条件においては，IVA や ILRMA を用いてもしばしば分離に失敗してしまう．これは，各音源信号の時間周波数成分がダイナミックに変動することから，IVA や ILRMA が仮定する音源モデルが同一音源の時間周波数成分を正しく捉えられないことに起因していると思われる．例えば，IVA や ILRMA において，まとまった周波数帯域でパーミュテーションが入れ替わる問題（ブロックパーミュテーション問題）[17] が報告されている．Fig. 2.7 にブロックパーミュテーション問題の様子を示す．Fig. 2.7 では，4 kHz 以上の周波数帯がまとまって入れ替わった状態で分離信号が推定されてしまっている．このような事実からも，依然としてパーミュテーション問題の解決は不十分であり，更なる高精度なパーミュテーション解決法の模索が重要であることが分かる．Fig. 2.7 のような明らかなブロックパーミュテーションであれば，ユーザアノテーションにより修正するインタラクティブな BSS アルゴリズム [18] も適用可能であるが，多くの帯域にブロックパーミュテーション問題が発生する場合もあり，ユーザアノテーションの利用も難しい状況が存在する．

近年では，IVA や ILRMA におけるパーミュテーション問題の解決に対する新しいアプローチとして，周波数帯域を細分化する手法が提案されている [23]．この手法では，全ての周波数成分を重なりを持つ複数のサブバンドに分割し，それぞれのサブバンドに対して IVA や ILRMA といった BSS 手法を適用する．サブバンド内で分離が行われた結果を基に次のサブバンドの初期値を設定することで，サブバンド間のパーミュテーション整列が可能になる．

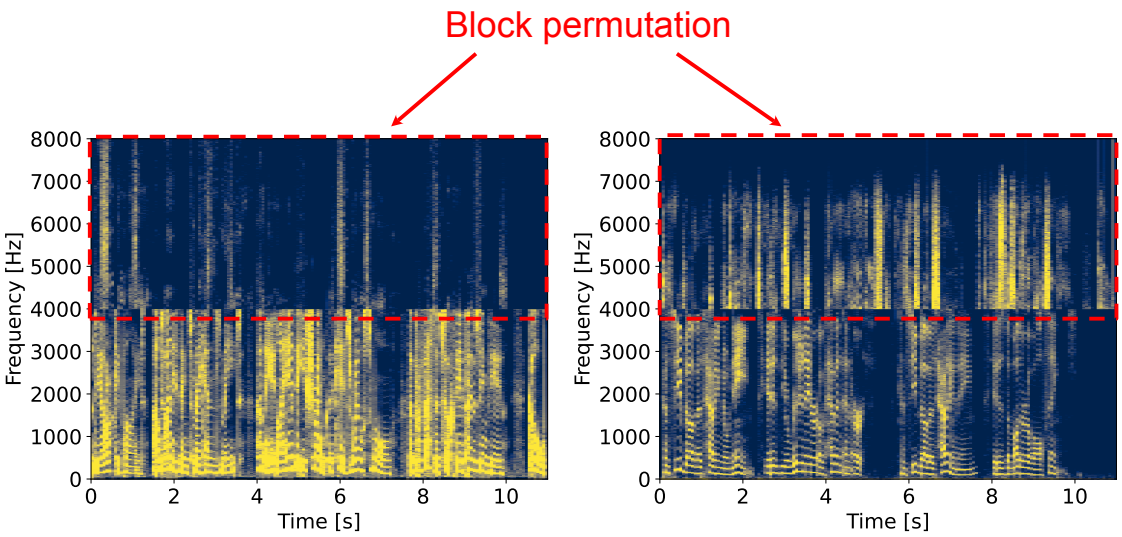

Fig.2.7. Example of block permutation problem.

このアプローチにより，従来の IVA や ILRMA では完全には解決できなかったブロックパーミュテーション問題の解決が実現され，計算コストを増加させることなく分離性能の向上が確認されている．このサブバンド分割手法は，BSS 手法の処理対象を狭い帯域に制限することで，分離アルゴリズムの効率を向上させる点が特徴的である．特に，重なりを持つサブバンド構造を利用することで，サブバンド間での分離結果の一貫性を確保し，結果としてパーミュテーション問題の解消に寄与することが示されている．しかし，それでもなお声質の近い複数音声の混合等では，パーミュテーション問題の解決が困難であることが報告されている．

2.7 DPS

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には，頑健性や汎化性に課題がある．これらの手法では，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適合する音源モデルを仮定することで高い性能を発揮する．しかし，この音源モデルが実際の音源信号に適合しない場合，分離性能が劣化し，ブロックパーミュテーション問題やその他の分離誤差が発生する可能性がある．加えて，幅広い音源信号に対応できる万能な音源モデルの構築は，困難である．

このような課題を解決するために，音源モデルに依存せず，可能な限り汎化性能の高いパーミュテーション解決モデルを学習するアプローチとして，DPS が提案されている [24]．文献 [24] では，観測信号に含まれる時間周波数成分を基に，同一音源に属する成分を予測する深層ニューラルネットワーク（DNN）を構築することで，パーミュテーション問題を解決する手法を提案している．具体的には，サブバンド内のパーミュテーション問題を解決する DNN をまず適用し，その後，サブバンド間のパーミュテーション問題をスティッチング処理 [27] によって統合する二段階構造を採用している．しかしながら，前段階で使用される DNN は，

参照周波数ビンと他の周波数ビンの成分が同一音源に属するかを二値分類する仕組みであるため，音源数が $N \geq 3$ の場合，スティッチング処理が非常に複雑となる．その結果，文献 [24] の手法は $N = 2$ の場合に限定され，3 音源以上への適用には限界がある．

それにもかかわらず，DNN を用いてパーミュテーション問題を解決する DPS のアプローチには重要な意義がある．DPS は従来の音源モデル仮定に依存せず，汎化性能が高いパーミュテーション解決法を構築することを可能にする．DPS は，万能な音源モデルの構築が困難な状況下で，パーミュテーション問題の新たな解決策として検討されている．そこで，本論文では，次章で述べる動機を基に，パーミュテーション問題を解決可能な DPS の構築を目指す．提案手法の有効性を検証するため，構築したモデルを用いた一連の実験を実施し，その結果について報告する．

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，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．2.5 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．2.6 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，BSS の IVA と ILRMA について説明した．そして，2.7 節では既存の DPS の概要と問題点について述べた．次章以降では，本論文で提案する新しい DPS の動機とアルゴリズムについて詳しく述べる．

第 3 章

提案手法

3.1 まえがき

前章では，音響信号の BSS において重要な FDICA のパーミュテーション問題について詳しく述べた．また，音源モデルに基づきパーミュテーション問題を回避する手法や，近年提案された DPS について説明した．

本章では，音源数 N が増加した場合でもアルゴリズムが複雑化せず，周波数ビン単位でのパーミュテーション問題を正確に解決する DPS を新たに提案する．まず 3.2 節では，BSS において DNN を用いてパーミュテーション問題の解決を目指す動機について述べる．3.3 節及び 3.4 節では，本論文で提案する DPS の DNN モデルの入出力及びモデル構造をそれぞれ説明する．3.5 節及び 3.6 節では，誤差逆伝播に用いる損失関数の取り方とパーミュテーション行列を正確に推定するモデルを学習するための入力データ及び正解データ（ラベル）の取得方法をそれぞれ説明する．最後に，3.7 節で本章のまとめを述べる．

3.2 動機

文献 [28] では，IVA や ILRMA に基づく BSS の STFT における最適な短時間区間長（窓長）Q について実験的に調査している．Fig. 3.1(b) は，文献 [28] 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 [28] を参照されたい．縦軸は信号対歪み比（source-to-distortion ratio: SDR）[29] の改善量であり，これはすなわち音源分離の性能を表している．この結果より，IVA 及び ILRMA では，残響時間が 470 ms という比較的残響の強い条件では，IVA も ILRMA も高精度な音源分離に失敗していることが分かる．一方で，FDICA に対して，音源信号 $\boldsymbol{s}_{ij}$ 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（すなわち FDICA の達成しうる限界性能）では 10 dB 以上の SDR の改善を達成している．この事実は，高残響下での音声信号の混合という難しい観測条件であっても，$\hat{\boldsymbol{W}}_i$ の推定自体（すなわち周波数ビン毎の BSS）は FDICA でも高精度に実現できていることを示している．すなわち，残る課題は推定信号 $\boldsymbol{y}_{ij}$ を正しい順番に並び変えるパーミュ

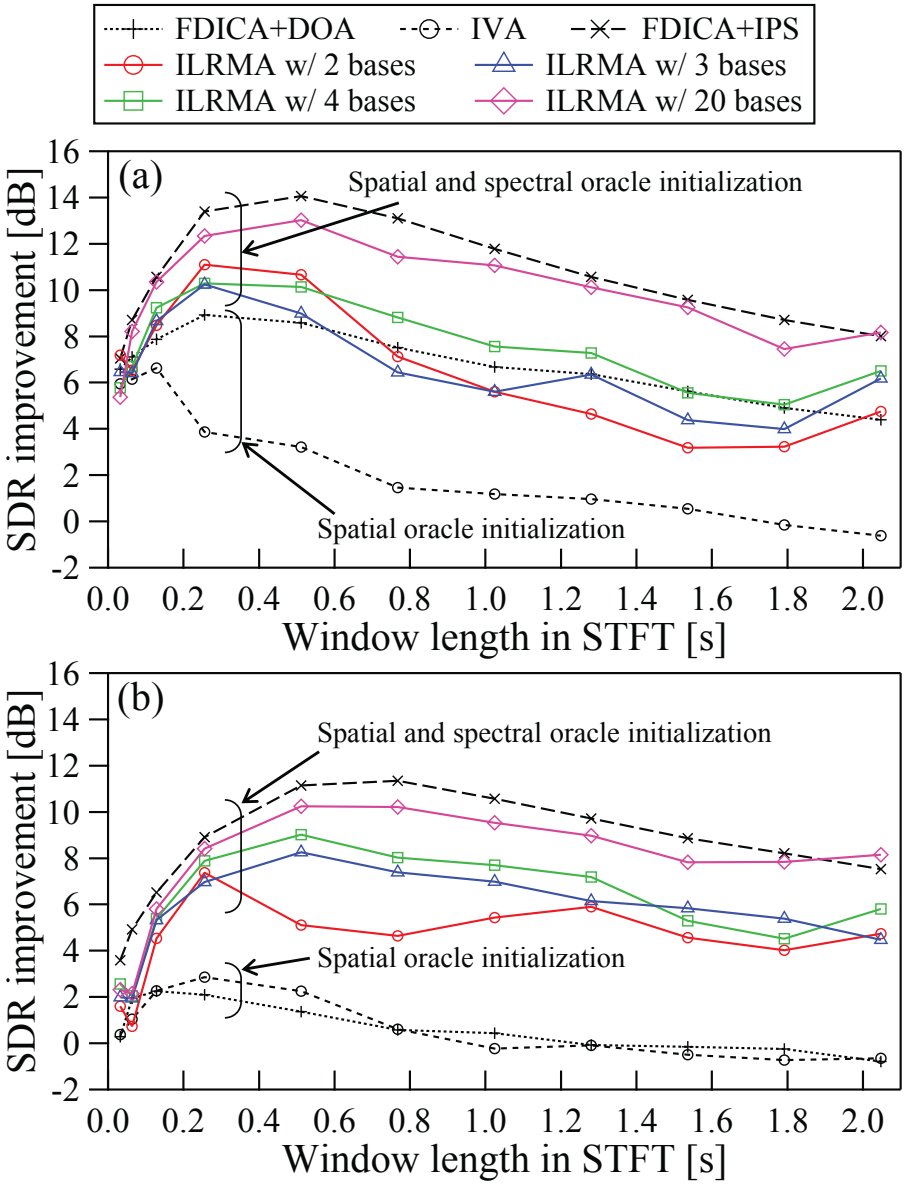

Fig.3.1. Average source separation results for speech signals using random initialization: (a) E2A (T_{60} = 300 ms) and (b) JR2 (T_{60} = 470 ms) impulse responses. For details of this figure, see [28].

テーション問題の解決（$\boldsymbol{P}_i^{-1}$ の推定）のみであることを示唆している．

この事実を動機として，DPS が提案されている．従来の DPS [24] では，サブバンド内のパーミュテーション問題を解決する際に，参照周波数ビンに対して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ミュテーション問題の解決（全サブバンドのスティッチング）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, 長･短期記憶（long-short term memory: LSTM）ユニット [30] を用いた双方向再帰型ニューラルネットワーク（bidirectional recurrent neural network using LSTM: BiLSTM）に基づく DPS を提案し，パーミュテーション問題の解決を目指す．BiLSTM は，入力データに対して順方向及び逆方向の両方から特徴量を抽出し，学習を行う点で特徴的な手法である．この特性により，時間的あるいは系列データにおける双方向の依存関係を効果的にモデル化することが可能である．また，BiLSTM の基盤となる LSTM ユニットは，入力系列内の長期的な依存関係を記憶する能力に優れており，勾配消失問題を軽減するゲート機構を備えているため，長い系列データを扱う際に強力な性能を発揮する．この特性は，周波数方向における各音源成分間の相関を効果的に捉えるために重要である．提案 DPS では，各音源間の周波数方向における関係性を明確に学習することを目的とし，各音源信号のスペクトログラムを周波数方向に結合した形で BiLSTM に入力するアプローチを採用する．DNN 構造の詳細は，3.4 節で述べる．提案 DPS は，音源数 N の増加に伴うアルゴリズムの複雑性を抑制することを目的として設計されている．しかしながら，本論文では，提案手法がパーミュテーション問題を解決可能であるかを検証するための調査に焦点を当て，音源数及びチャネル数が $N = M = 2$ の場合に限定して性能の評価を行う．なお，音源数 $N \geq 3$ 以上の条件における性能評価については，今後の課題として位置づける．

本論文で提案する DPS を BSS に適用する処理の概要は以下の通りである．

(a) パーミュテーション問題が未解決の状態である推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{N} \in \mathbb{C}^{I\times J*1}$に対し，それぞれの信号パワー比に基づく正規化 [6] を施す
(b) 正規化された振幅パワースペクトログラム $(\overline{\boldsymbol{Y}}_{n'})_{n'=1}^{N} \in \mathbb{C}^{I\times J}$ 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パワー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パワースペクトログラムを構成する
(c) 局所時間振幅パワースペクトログラム $(\check{\boldsymbol{Y}}_{jn'})_{n'=1}^{N} \in \mathbb{C}^{I\times(2\beta+1)}$ を時間方向に結合した行列 $\left[\check{\boldsymbol{Y}}_{j1} \ \cdots \ \check{\boldsymbol{Y}}_{jN}\right] \in [0,1]^{I\times N(2\beta+1)}$ を DNN に入力する．
(d) DNN は正規化局所時間振幅パワー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行列を出力する
(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行列 $\hat{\boldsymbol{L}}_j \in [0,1]^{I\times N!}$ を取得する
(f) 確率値行列 $\hat{\boldsymbol{L}}_j \in [0,1]^{I\times N!}$ をもとに，予測パーミュテーション行列 $\hat{\boldsymbol{P}}_{ij}^{-1} \in [0,1]^{N\times N}$ を構成する

*1 分離信号は，音源の順序が必ずしも n と一致しているとは限らないため，n と n' を使い分けている．

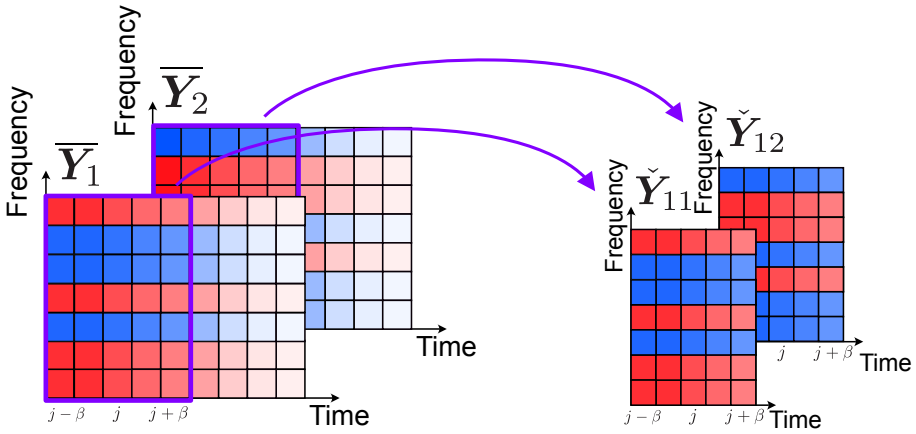

Fig.3.2. Extraction of local-time-frame power spectrogram ($N = 2$).

(g) 全時間フレームの確率値行列を用いて構成した，予測パーミュテーション行列に対して，時間方向に多数決処理を適用し，全時間フレーム共通のパーミュテーション行列 $\hat{\boldsymbol{P}}_i^{-1} \in \{0,1\}^{N\times N}$ を推定する

(h) $\hat{\boldsymbol{z}}_{ij} = \hat{\boldsymbol{P}}_i^{-1}\boldsymbol{y}_{ij}$ よりパーミュテーション問題が解決された分離信号を得る

上記の処理の詳細や DNN の学習方法については，次節以降で詳しく述べる．

3.3　DNN の入出力

FDICA からは，パーミュテーション問題が発生した状態の推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{N}$ が得られる．ここで，同一音源に属する成分の相関を強調するため，推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{N}$ をパワースペクトログラム $(|\boldsymbol{Y}_{n'}|^{.2})_{n'=1}^{N}$ の比率に変換する正規化 [6] を施す．この処理は次式で表される．

$$\overline{\boldsymbol{Y}}_{n'} = \frac{|\boldsymbol{Y}_{n'}|^{.2}}{\sum_{n'=1}^{N}|\boldsymbol{Y}_{n'}|^{.2}} \in [0,1]^{I\times J} \tag{3.1}$$

ここで，$|\cdot|^{.2}$ 及び括線はそれぞれ行列の要素毎の絶対値の 2 乗及び要素毎の割り算を示す．このような正規化は，文献 [6] 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，推定信号の正規化パワースペクトログラム $(\overline{\boldsymbol{Y}}_{n'})_{n'=1}^{N}$ 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パワースペクトログラムを抽出する．この処理は次式で表される．

$$\check{\boldsymbol{Y}}_{jn'} = [\overline{\boldsymbol{y}}_{(j-\beta)n'} \quad \cdots \quad \overline{\boldsymbol{y}}_{(j+\beta)n'}] \in [0,1]^{I\times(2\beta+1)} \tag{3.2}$$

ここで，$\overline{\boldsymbol{y}}_{jn'} \in [0,1]^{I}$ は $\overline{\boldsymbol{Y}}_{n'}$ の j 列目の列ベクトルを表す．また，β（0 以上の整数）は時間フレーム j の近傍時間フレームをどの程度 DNN に入力するかを決めるハイパーパラメータである．Fig. 3.3 に示すように，提案 DPS では $(\check{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ を時間方向に結合した行列 $\left[\check{\boldsymbol{Y}}_{j1} \ \cdots \ \check{\boldsymbol{Y}}_{jN}\right] \in [0,1]^{I\times N(2\beta+1)}$ を BiLSTM に入力する．

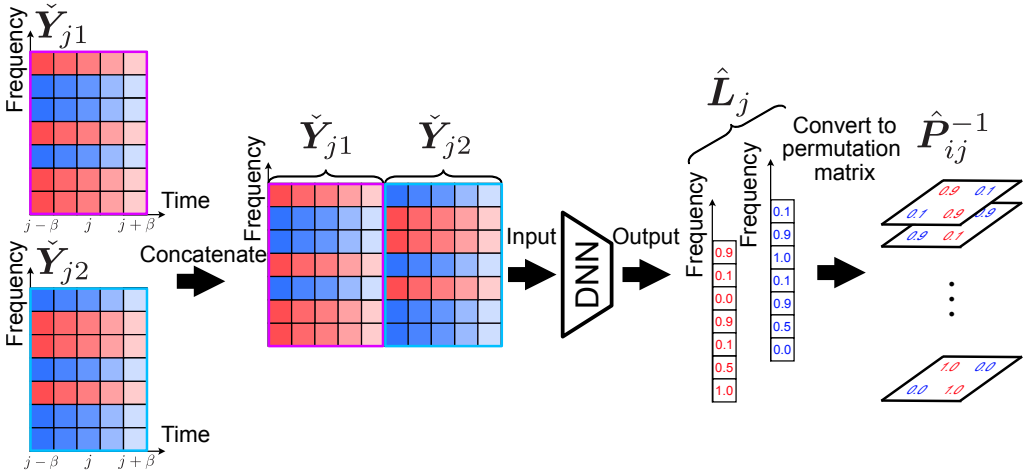

Fig.3.3. Estimation of permutation matrix ($N = 2$).

提案 DPS では，予測結果として行列 $\hat{\boldsymbol{L}}_j \in [0,1]^{I\times N!}$ を出力する．$\hat{\boldsymbol{L}}_j$ は，各周波数ビンにおけるパーミュテーション行列の予測確率値 $\hat{l}_{iqj} \geq 0$ を要素とする行列であり，$q = 1, 2, \cdots, N!$ は N 個の音源に対する $N!$ 通りの順列のインデックスを表す．さらに，$\hat{l}_{iqj}$ は確率値として，以下の制約を満たす．

$$\sum_q \hat{l}_{iqj} = 1, \quad \forall i, j \tag{3.3}$$

この制約は，softmax 関数を用いることで実現可能であり，DNN モデル内で処理される（詳細は次節で述べる）．

音源数 $N = 2$ の場合を例にとると，予測されるパーミュテーション行列 $\hat{\boldsymbol{P}}_{ij}^{-1}$ は，$\hat{l}_{i1j}$ 及び $\hat{l}_{i2j}$ を用いて次式のように表される．

$$\hat{\boldsymbol{P}}_{ij}^{-1} = \hat{l}_{i1j}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \hat{l}_{i2j}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \in [0,1]^{N\times N} \tag{3.4}$$

式 (3.4) において，$\hat{l}_{i1j}$ 及び $\hat{l}_{i2j}$ は，入力信号 $\check{\boldsymbol{Y}}_{j1}$ 及び $\check{\boldsymbol{Y}}_{j2}$ の各周波数成分が「1 番目の音源に属する確率」と「2 番目の音源に属する確率」をそれぞれ示している．

$\sum_q \hat{l}_{iqj} = 1$ を満たすことから，$\hat{\boldsymbol{P}}_{ij}^{-1}$ は DSM であることが分かる．また，Birkhoff–von Neumann の定理 [31]（付録 A 参照）を考慮すると，パーミュテーション問題が発生している入力データから DSM を予測する提案 DPS の DNN は，考えうるすべてのパーミュテーション行列に対する凸結合係数を推定していることになる．言い換えれば，DNN は考えうるパーミュテーション行列の中で，どの行列が正解であるかという確信度を確率値として予測していると解釈できる．

さらに，パーミュテーション問題の解は時間方向において不変である（式 (2.23) における $\boldsymbol{P}_i$ が時間フレーム j によらないことを意味する）．そのため，異なる局所時間パワースペクトログラムから得られる予測パーミュテーション行列 $(\hat{\boldsymbol{P}}_{ij}^{-1})_{j=1}^{J}$ について，多数決処理を適用することでより精度の高い予測結果 $\hat{\boldsymbol{P}}_i^{-1} \in \{0,1\}^{N\times N}$ を生成することが可能である．この操作

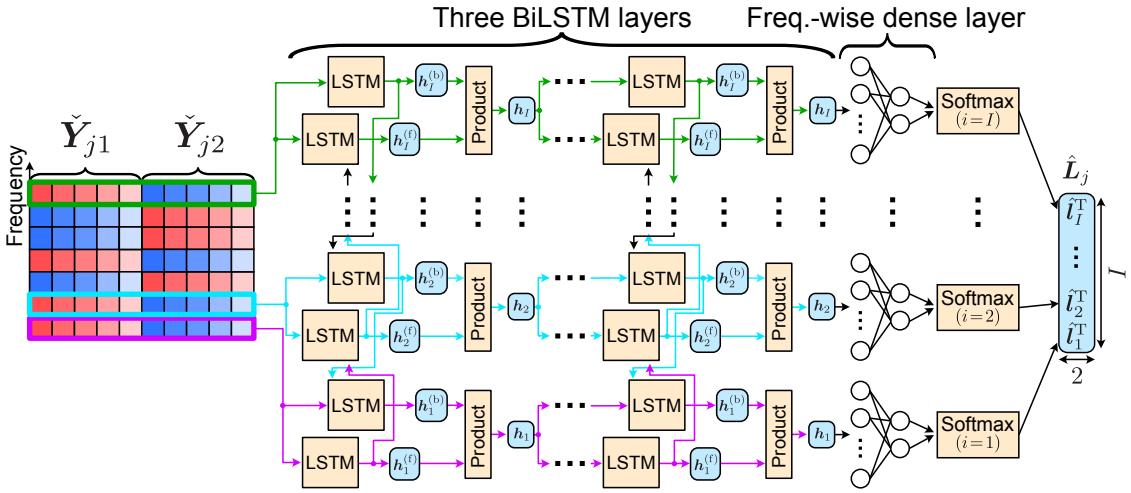

Fig.3.4. DNN architecture ($N = 2$).

により，パーミュテーション行列の推定誤差が低減され，音源分離精度が向上することが期待される．

3.4　DNN の構造

提案 DPS では，周波数ビン単位のパーミュテーション問題を解決する際に重要となる，各音源の周波数方向における関係性を学習するため，BiLSTM を周波数方向に適用する．BiLSTM は，時間や周波数など連続的な系列を持つ入力に対して，順方向及び逆方向の再帰性を同時に考慮した学習を可能とする DNN である．この特性により，周波数間の依存関係を効果的に学習することが可能である．

Fig. 3.4 に提案 DPS で用いる DNN の構造を示す．本手法では，まず BiLSTM を 3 層適用する．なお，ここで採用している構造は，1 つの BiLSTM 内で層を増やす「多層 BiLSTM（multilayer BiLSTM）」ではなく，BiLSTM 層そのものを複数積み重ねる「スタック型 BiLSTM（stacked BiLSTM）」である．各 BiLSTM 層では，周波数ビンの順方向の特徴量 $\boldsymbol{h}_i^{(\mathrm{f})} \in \mathbb{R}^{N(2\beta+1)}$ と逆方向の特徴量 $\boldsymbol{h}_i^{(\mathrm{b})} \in \mathbb{R}^{N(2\beta+1)}$ をそれぞれ出力する．これらの特徴量は，同一周波数ビンにおいて要素毎に積を取ることで結合され，次式で表されるベクトルが出力される．

$$\boldsymbol{h}_i = \boldsymbol{h}_i^{(\mathrm{f})} \odot \boldsymbol{h}_i^{(\mathrm{b})} \in \mathbb{R}^{N(2\beta+1)} \tag{3.5}$$

ここで，$\odot$ は要素毎の積を表す．この操作により，順方向と逆方向の特徴量が統合され，各周波数ビンにおける依存関係を考慮した特徴量が得られる．

続いて，3 層の BiLSTM の出力である特徴量 $\boldsymbol{h}_i$ を，$N(2\beta+1)$ から $N!$ 次元に圧縮するため，周波数毎に全結合層を適用し，さらに Softmax 関数を用いて次式のように正規化された

確率値ベクトルを生成する．

$$\hat{\boldsymbol{l}}_i = \mathrm{Softmax}(\mathrm{Dense}(\boldsymbol{h}_i)) \in [0,1]^{N!} \tag{3.6}$$

ここで，Dense(·) は全結合層を，Softmax(·) は Softmax 関数をそれぞれ表す．この処理により，各周波数ビン i に対して，$N!$ 個のパーミュテーション行列に対応する確率値を得ることができる．

最終的な DNN の出力である $\hat{\boldsymbol{L}}_j$ は，$\hat{\boldsymbol{l}}_i^{\mathrm{T}}$ を行ベクトルとして構成される行列であり，以下のように表される．

$$\hat{\boldsymbol{L}}_j = \begin{bmatrix} \hat{\boldsymbol{l}}_1^{\mathrm{T}} \\ \hat{\boldsymbol{l}}_2^{\mathrm{T}} \\ \vdots \\ \hat{\boldsymbol{l}}_I^{\mathrm{T}} \end{bmatrix} \in [0,1]^{I \times N!} \tag{3.7}$$

Softmax 関数の適用により，$\hat{\boldsymbol{L}}_j$ の要素は非負であり，さらに式 (3.3) 次の条件を満たす．すなわち，$\hat{\boldsymbol{L}}_j$ は各周波数ビン i において，パーミュテーション行列の予測確率値を表現していると解釈できる．

3.5　DNN 学習時の損失関数

提案 DPS の DNN 学習は，損失関数を定義し，その値を最小化するようにパラメータを更新する誤差逆伝播法に基づいて行われる．提案 DPS 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．具体的には，音源数が $N = 2$ の場合，$(\hat{l}_{i1j}, \hat{l}_{i2j})$ を予測する 2 クラス分類器として機能し，softmax 関数を用いて各クラスに対する確率値を出力する．

通常，多クラス分類器の損失関数には，カテゴリカル分布[*2]の負対数尤度関数であるカテゴリカル交差エントロピー（categorical cross entropy: CCE）が用いられる．これにより，DNN の学習を最尤推定の枠組みで行うことが可能である．しかしながら，提案 DPS の本来の目的は，全周波数ビンにおいて正確なパーミュテーション行列を予測することではなく，最終的に正確な分離信号 $(\boldsymbol{Z}_{n'})_{n'=1}^{N}$ を得ることである．

例えば，推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{N}$ のエネルギーがほとんどないような周波数ビンにおいて，誤ったパーミュテーションが予測されても，結果として得られる分離信号 $(\boldsymbol{Z}_{n'})_{n'=1}^{N}$ の音源分離精度にほとんど影響を与えない．CCE を損失関数として用いると，エネルギーが少ない（音源分離において重要度が低い）周波数ビンと，エネルギーが大きい（音源分離において重要度が高い）周波数ビンが等しい重要度で扱われることになる．このような場合，音源分離性能の向上が妨げられる可能性がある．

そこで，提案 DPS では，損失関数として，CCE を用いるのではなく，平均二乗誤差（mean squared error: MSE）を損失関数として用いる．具体的な損失関数の取り方を Fig. 3.5 に

[*2] 多項分布における試行回数を 1 回とした場合の分布．

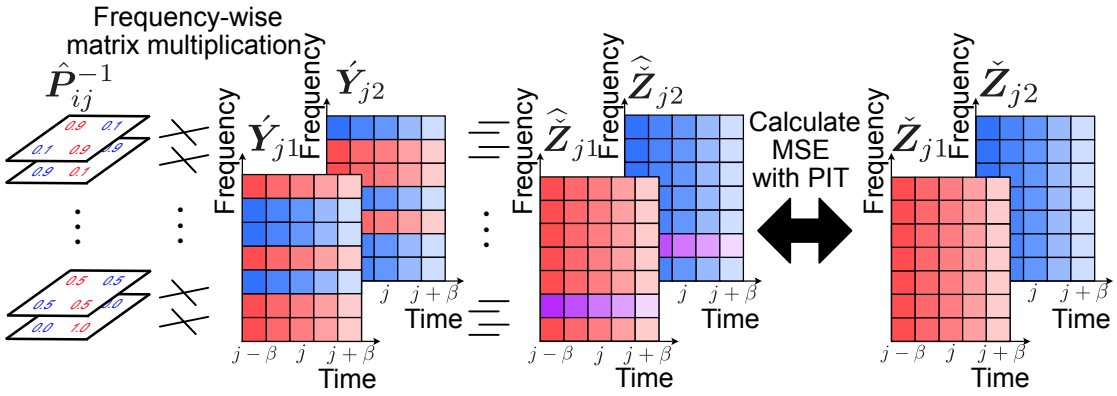

Fig.3.5. Loss function using MSE with PIT ($N = 2$).

示す．提案 DPS における DNN の出力から得られる予測パーミュテーション行列 $\hat{\boldsymbol{P}}_{ij}^{-1}$ と，FDICA による推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{N}$ の局所時間複素スペクトログラム $(\acute{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ との間で行列積を取ることで，予測分離信号 $(\hat{\tilde{\boldsymbol{Z}}}_{n'})_{n'=1}^{N}$ を得る．こうして得られる予測分離信号 $(\hat{\tilde{\boldsymbol{Z}}}_{n'})_{n'=1}^{N}$ に対して，正解ラベルとなる分離信号 $(\check{\boldsymbol{Z}}_{n'})_{n'=1}^{N}$ を準備する．正解ラベルは，分離信号 $(\boldsymbol{Z}_{n'})_{n'=1}^{N}$ の局所時間複素スペクトログラムで構成される．これらの信号間の誤差を評価するため，損失関数として MSE を以下の式により定義する．

$$C_q = \sum_{n'=1}^{N} \left\| \hat{\tilde{\boldsymbol{Z}}}_{jn'} - \check{\boldsymbol{Z}}_{j\mathcal{P}(q,n')} \right\|_2^2 \tag{3.8}$$

ここで，C_q は q 番目の順列における予測信号と正解信号の二乗誤差を示す．また，$\mathcal{P}(q, n')$ は $N!$ 個のすべての可能な順列のうち，q 番目の順列における n' 番目の音源インデックスを返す関数である．

提案 DPS では，順列に依存しない学習を実現するため，順序不変学習（permutation invariant training: PIT）[32] を導入し，損失関数 $\mathcal{L}$ を次式で定義する．

$$\mathcal{L} = \min_{q \in \{1,\ldots,N!\}} C_q \tag{3.9}$$

この損失関数設計の利点は，音源順序全体を厳密に固定する必要がなく，音源分離精度に直接影響するパーミュテーション問題の解決に焦点を当てられる点である．その結果，音源間の順序の不確実性を考慮しながら DNN を効果的に訓練することが可能となり，音源分離精度の向上が期待される．このアプローチは，複数音源を含む困難な状況でも柔軟に対応できる点で，従来手法に比べて優位性を有する．

3.6　学習済の DNN のテストデータへの適用

DNN の学習が完了した後，提案 DPS は，FDICA の推定信号 $(\boldsymbol{Y}_{n'})_{n'=1}^{N}$ に適用される．パーミュテーション問題は，時不変な分離行列 $\hat{\boldsymbol{W}}_i$ に起因して発生するため，正しい音源順序は時間フレーム方向において一定である（$\boldsymbol{P}_i$ は j に非依存）．この性質を利用し，Fig. 3.6 に

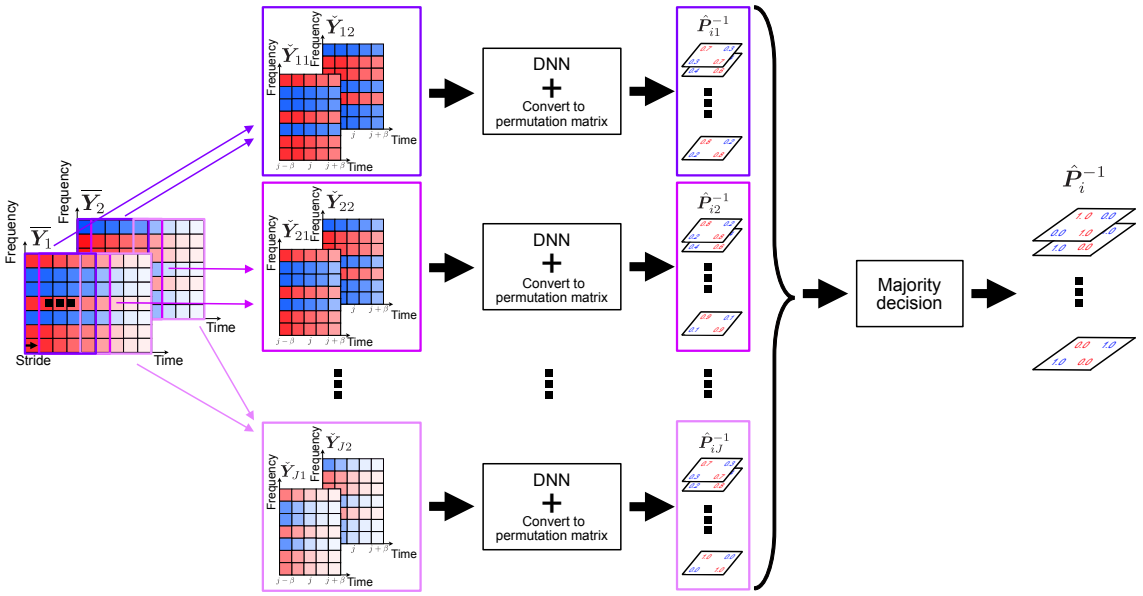

Fig.3.6. DNN predictions for all local-time-frame power spectrograms and their majority decision ($N = 2$).

示すように，テストデータへの適用時には，様々な時間フレーム j における局所時間パワースペクトログラム $(\check{\boldsymbol{Y}}_{jn'})_{n'=1}^{N}$ を DPS に入力し，出力されるパーミュテーション行列 $(\hat{\boldsymbol{P}}_{ij}^{-1})_{j=1}^{J}$ に対して多数決処理を行う．この処理により，全時間フレームにわたる一貫したパーミュテーション行列 $\hat{\boldsymbol{P}}_i^{-1}$ を以下の式で得る．

$$\hat{\boldsymbol{P}}_i^{-1} = \mathrm{round}\left(\frac{1}{J}\sum_{j=1}^{J}\hat{\boldsymbol{P}}_{ij}^{-1}\right) \in \{0,1\}^{N\times N} \tag{3.10}$$

ここで，round$(\cdot)$ は入力行列の要素ごとに四捨五入を適用する演算を表す．

多数決処理の利点は，個々の時間フレームにおける DNN の出力に含まれる誤差を統計的に緩和できる点にある．各時間フレーム j で出力される $\hat{\boldsymbol{P}}_{ij}^{-1}$ は，観測データに依存した局所的な特徴に基づいているため，一部の時間フレームで誤ったパーミュテーション行列が出力される可能性がある．しかし，多数決処理を行うことで，全時間フレームにわたる予測結果を統合し，時間的に安定した一貫性のあるパーミュテーション行列を導出することが可能となる．これにより，推定結果の頑健性が向上し，音源分離性能のさらなる改善が期待できる．

最終的な推定分離信号 $\hat{\boldsymbol{z}}_{ij}$ は，得られた一貫したパーミュテーション行列 $\hat{\boldsymbol{P}}_i^{-1}$ を用いて次式で計算される．

$$\hat{\boldsymbol{z}}_{ij} = \hat{\boldsymbol{P}}_i^{-1}\boldsymbol{y}_{ij} \tag{3.11}$$

この処理により，提案 DPS は時間フレーム間での一貫性を考慮した，高精度な音源分離を実現する．

3.7　本章のまとめ

本章では，FDICA のポスト処理として行う DPS について新たに提案した．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 の入力に同一音源に属する成分の相関を強調させるために，正規化を行った局所時間振幅パワースペクトログラムを用いることことと，DNN の出力としてパーミュテーション行列の係数となる確率値を得ることを説明した．3.4 節では，各音源の周波数方向における関係性を学習するため，BiLSTM を用いたモデルを構築したことについて説明した．3.5 節では，DNN の出力に従ってパーミュテーション行列を作成した後，FDICA を適用した信号を並び替えることで得られる予測分離信号と正解の分離信号との間で MSE を用いた損失関数を取ることを説明した．3.6 節では，テストデータに対して時間方向に多数決処理を行うことで，パーミュテーション問題の解決精度を向上させることを説明した．

第 4 章

実験

4.1　まえがき

本章では，前章で説明した DPS の有効性を評価するために行った一連の実験について述べる．提案 DPS は，FDICA による推定信号に対するパーミュテーション問題を解決することを目的としており，その性能を実際の音声信号及び音楽信号に適用することで評価を行った．

本章の構成は以下の通りである．4.2 節では，提案 DPS の性能を調査するために，残響や FDICA の推定信号に含まれる分離誤差を含まないクリーン信号を対象として行なった実験について説明する．4.3 節では，実際に FDICA の推定信号に対して提案 DPS を適用し，従来の BSS と比較した実験について説明する．最後に，4.4 節で本章のまとめを述べる．

4.2　提案 DPS の汎化性能に関する評価実験

本実験では，提案 DPS に対して 2 つの異なる条件下で実験を行い，性能を評価した．評価指標としては，SDR [29] の改善量を用いた．SDR 改善量は，音源分離性能を定量的に評価するための標準的な指標であり，推定信号が元の信号にどの程度近いかを示す．本実験では，SDR 改善量の比較を通じて，提案手法の有効性を検証した．以下の 4.2.1 項で，具体的な実験条件について説明する．

4.2.1　実験条件

本実験では，従来 DPS [24] 及び提案 DPS の汎化性能を評価するために，音声信号と音楽信号を用いた実験を実施した．音声信号のみを用いて学習した DNN モデルと，音楽信号のみを用いて学習した DNN モデルの 2 つを構築し，各モデルに対して in-domain（学習データとテストデータに同一の音源を用いる場合）及び out-of-domain（学習データとテストデータに異なる種類の音源を用いる場合）の条件で評価を行った．In-domain 及び out-of-domain の違いを Fig. 4.1 に示す．

音源信号 $(\boldsymbol{S}_1, \boldsymbol{S}_2)$ として，Table 4.1 に示す SiSEC2011 データセット [33] に含まれる男

Table 4.1. Speech and music sources obtained from SiSEC2011 [33]

Signal type	Source	Data name	Length
Speech	Male speech	`dev2_male4_inst_src_2.wav`	10.0 s
	Female speech	`dev3_female4_inst_src_2.wav`	10.0 s
Music	Drums	`dev1_wdrums_src_3.wav`	11.0 s
	Guitar	`dev1_wdrums_src_2.wav`	11.0 s

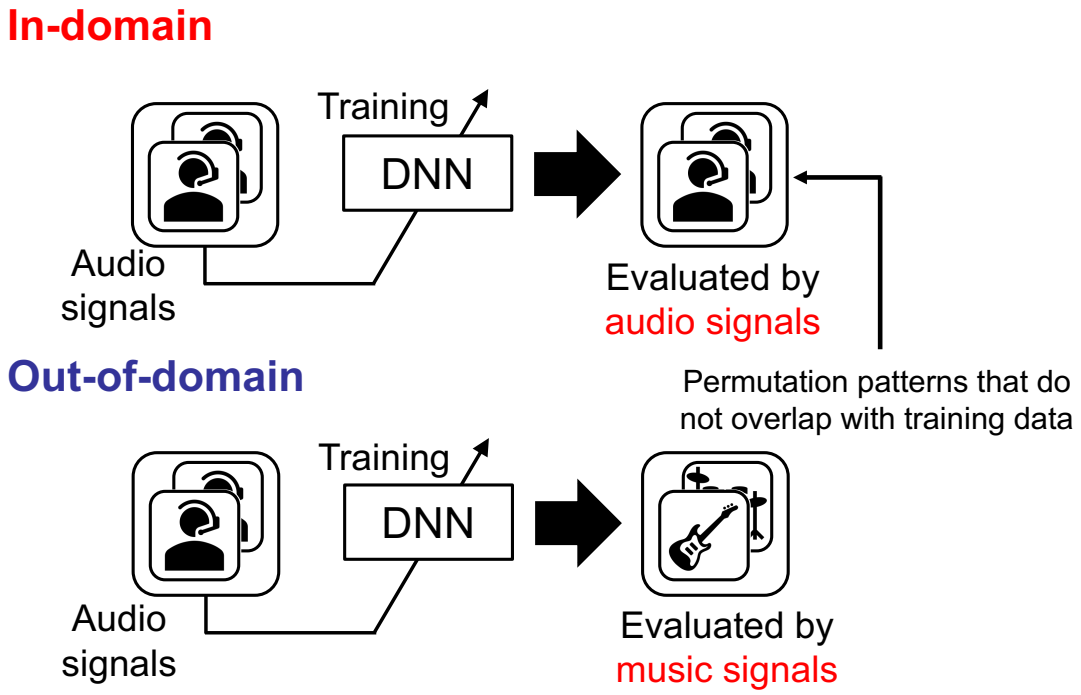

Fig.4.1. In-domain and out-of-domain experimental conditions.

女の音声信号，またはドラムとギターの音楽信号の 2 種類のペアを用いた．これらの信号のサンプリング周波数は 16 kHz に設定し，短時間フーリエ変換（STFT）の窓長は 2048 点（128 ms），シフト長は 1024 点（64 ms），窓関数には Hann 窓を用いた．

本実験では，Fig. 4.2 に示すように，各音源信号 $(\boldsymbol{S}_1, \boldsymbol{S}_2)$ の周波数ビン単位でランダムに成分を入れ替えることで，パーミュテーション問題を模擬した推定信号 $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ を生成した．学習データには，150 パターンのランダム入れ替えを用いて作成した $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ を使用し，テストデータには学習データには含まれない 10 パターンのランダム入れ替えを用いて作成した $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ を使用した．

DNN の条件については，最適化手法として Adam を用い，ミニバッチサイズを 8，エポック数を 500 に設定した．BiLSTM の局所時間長は $\beta = 13$ とし，隠れ層の次元は各 BiLSTM 層の出力と同じ $N(2\beta + 1)$ に設定した．

DNN の最適化法には Adam [34] を用いる．Adam の重みの最適化アルゴリズムを次式に

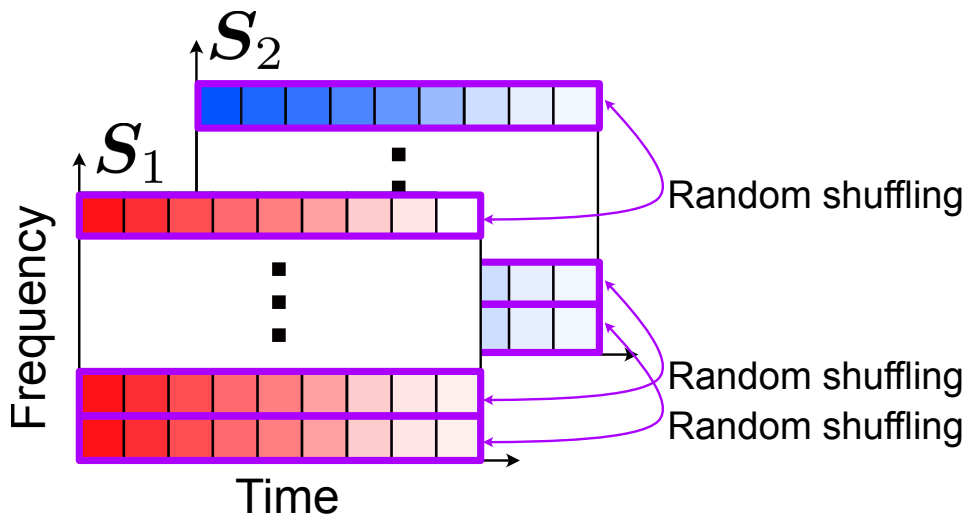

Fig.4.2. Creating signals that imitate permutation problems.

示す．

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

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

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

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

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

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

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

ここで，$\mathcal{L}(\mathbf{w})$ 及び $\mathbf{w}$ はそれぞれ損失関数及び DNN の最適化変数をまとめたベクトルであり，$\mathbf{m}_t$ 及び $\mathbf{v}_t$ はいずれも過去の勾配変化を表すモーメンタムと呼ばれる量である．また，上付き文字の t は最適化（変数更新）の反復回数を表す．$\mathbf{m}_t$ 及び $\mathbf{v}_t$ が考慮されていることにより，変数更新における振動を抑えながら高速かつ安定な最適化が可能となる．式 (4.2)–(4.6) 中のハイパーパラメータはそれぞれ標準的な設定値である $\varepsilon = 1.0\times10^{-8}$，$\rho_1 = 0.9$，$\rho_2 = 0.999$，及び学習率 $\eta = 0.001$ に設定した．

評価指標である SDR は，音源分離の度合と分離音の歪みの少なさの両方を加味した客観評価尺度である．今，音源分離の目的音源信号を $\mathrm{s}(l)$，目的音以外の音源（干渉音源）信号を $\mathrm{n}(l)$ とすると，これらが混合した信号 $\mathrm{x}(l)$ は次式となる．

$$\mathrm{x}(l) = \mathrm{s}(l) + \mathrm{n}(l) \tag{4.8}$$

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

Table 4.2. SDRs [dB] for music test data using DPS trained with music signals (in-domain evaluation)

Test data pattern	Observed signal	Conventional DPS	Proposed DPS
1	-0.95	2.95	**64.75**
2	2.00	2.95	**64.75**
3	0.55	2.95	**155.00**
4	1.25	2.95	**64.75**
5	-1.00	2.95	**66.65**
6	-1.00	2.95	**61.15**
7	-0.85	2.95	**66.65**
8	-0.15	2.95	**64.75**
9	0.60	2.95	**64.75**
10	-0.35	2.95	**61.15**

れる．

$$\hat{\mathrm{s}}(l) = \mathrm{s}_{\mathrm{target}}(l) + \mathrm{e}_{\mathrm{interf}}(l) + \mathrm{e}_{\mathrm{artif}}(l) \tag{4.9}$$

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

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

4.2.2　実験結果

Tables 4.2 及び 4.3 は，それぞれ音楽信号及び音声信号の in-domain のテストデータに対する SDR を示している．In-domain 評価では，学習データとテストデータで同じ音源信号 $(\boldsymbol{S}_1, \boldsymbol{S}_2)$ を使用しているため，高い性能が得られる一方で，過学習の可能性がある点に留意する．結果として，従来 DPS では観測信号の SDR 値から一定量の改善が見られたが，音楽信号で平均約 3 dB，音声信号で平均約 10 dB 程度に留まった．一方，提案 DPS では全てのテストデータで 50 dB 以上の SDR 改善が確認され，パーミュテーション問題を適切に解決できていることが分かった．ここで，SDR は一般的に知られる信号対雑音比（signal-to-noise ratio: SNR）と異なり，式 (4.10) に示すように，分母に干渉音源成分に加え，音源分離処理によって生じる歪み成分も含む指標であるため，SNR よりも厳しい基準である．提案 DPS が達成した 50 dB 以上の SDR 改善量は，SNR 換算で同程度以上に相当し，ハイエンドの電話

Table 4.3. SDRs [dB] for speech test data using DPS trained with speech signals (in-domain evaluation)

Test data pattern	Observed signal	Conventional DPS	Proposed DPS
1	-6.25	3.60	**44.5**
2	-6.85	4.65	**44.5**
3	-5.40	3.60	**44.5**
4	-6.45	3.55	**44.5**
5	-6.60	4.70	**44.5**
6	-6.45	4.65	**44.5**
7	-6.35	3.60	**44.5**
8	-5.50	4.65	**44.5**
9	-5.85	3.60	**44.5**
10	-5.55	4.65	**44.5**

Table 4.4. SDRs [dB] for speech test data using DPS trained with music signals (out-of-domain evaluation)

Test data pattern	Observed signal	Conventional DPS	Proposed DPS
1	-6.25	-8.00	**33.55**
2	-6.85	-5.85	**22.85**
3	-5.40	-7.20	**33.85**
4	-6.45	-7.60	**23.50**
5	-6.60	-7.40	**22.00**
6	-6.45	-7.25	**24.05**
7	-6.35	-1.40	**23.60**
8	-5.50	-7.65	**26.65**
9	-5.85	-6.40	**25.15**
10	-5.55	-7.90	**24.05**

システムにも利用できるほどの値である．実際に，人間の耳で確認するレベルでは完全に分離されたと言えるほどの高い性能である．

Tables 4.4 及び 4.5 は，それぞれ音楽信号及び音声信号の out-of-domain のテストデータに対する SDR を示している．これらの結果は，「音声信号で学習した DPS が音楽信号のパーミュテーション問題を解決できるか」及び「音楽信号で学習した DPS が音声信号のパーミュ

Table 4.5. SDRs [dB] for music test data using DPS trained with speech signals (out-of-domain evaluation)

Test data pattern	Observed signal	Conventional DPS	Proposed DPS
1	-0.95	**5.05**	3.35
2	2.00	**5.05**	1.75
3	0.55	**5.05**	3.35
4	1.25	**11.35**	3.35
5	-1.00	**11.35**	3.35
6	-1.00	**11.35**	3.35
7	-0.85	**11.35**	3.35
8	-0.15	**5.05**	3.35
9	0.60	**5.05**	3.35
10	-0.35	**5.05**	1.75

テーション問題を解決できるか」を評価したものであり，各モデルの汎化性能を示している．従来 DPS は，音声信号で学習した場合には一定の改善が得られたが，音楽信号で学習した場合にはほとんど改善が見られなかった．一方，提案 DPS では，音楽信号で学習したモデルが音声信号のパーミュテーション問題を解決し，平均 24 dB 以上の SDR 改善が得られた．また，音声信号で学習したモデルが音楽信号に適用された場合には従来 DPS と比較すると，やや性能が低下した．本実験で得られた分離信号のスペクトログラムの結果については付録 B を参照されたい．以上の結果から，提案 DPS は，特に音楽信号で学習したモデルにおいては，高い汎化性能を持ち，高精度でパーミュテーション問題を解決できることがわかった．

4.3　提案 DPS を BSS に応用した際の評価実験

本実験では，提案 DPS を FDICA によって得られる推定信号に適用し，パーミュテーション問題の解決性能を評価した．評価指標としては，SDR [29] の改善量を用いた．

4.3.1　実験条件

提案 DPS を実際の FDICA に応用した際の性能を評価するために，複数の比較手法と組み合わせた実験を実施した．具体的には，PS を用いない FDICA（PS: none），到来方向（direction of arrivals: DOA）情報を用いた PS [8] による FDICA（PS: DOA），提案手法（PS: DPS），IVA [13]，及び IPS を用いた FDICA（PS: IPS）との比較を行った．FDICA (PS: IPS) は FDICA に基づく音源分離の上限性能を示すものである．

提案 DPS の学習データには，Table 4.6 に示す SiSEC2011 のドラムとギターの音楽信号を

Table 4.6. Dry sources obtained from SiSEC2011 [33]

Source	Data name	Length
Drums	`dev1_wdrums_src_3.wav`	11.0 s
Guitar	`dev1_wdrums_src_2.wav`	11.0 s

用いた．これらの音楽信号に対して，音響シミュレーションライブラリである pyroomacoustics [35] を使用して，100 部屋分の残響環境をシミュレートした．pyroomacoustics は，室内音響環境を仮想的に再現するための Python ライブラリであり，部屋の形状，音源位置，マイクロホン配置，壁面の反射係数，及び残響時間（T_{60}）をパラメータとして設定することで，現実的な観測信号を合成することが可能である．このライブラリは，室内残響のシミュレーションや音響信号処理の研究において広く使用されている．

本実験では，部屋のサイズを横幅 5〜12 m，奥行き 5〜10 m，及び高さ 3〜5 m の範囲で一様分布に従ってランダムに設定した．壁面の反射係数を調整し，T_{60} は 220 ms 程度となるように設定した．2 個のマイクロホンを使用し，マイクロホン間隔は 5 cm，音源とマイクロホンの配置については高さを 1.5 m に固定し，横幅及び奥行きについては部屋の範囲内で一様分布に従う乱数で設定した．ただし，音源とマイクロホンがなす角度は必ず 30° 以上になるように調整した．

実際の音源分離タスクにおいては，FDICA で推定される分離行列 $\hat{\boldsymbol{W}}_i$ に推定誤差が含まれる．このため，周波数毎に異なる分離誤差を含む信号に対して，正確にパーミュテーション問題を解決する必要が生じる．これまでに説明してきた提案 DPS を FDICA の分離誤差を含む信号に適用した場合，FDICA の推定信号に含まれる分離誤差の影響によりパーミュテーション問題の解決性能が低下することが予想される．この課題を解決するために，FDICA における分離誤差を考慮した学習データを用意し，DPS の学習に用いるモデルを構築した．

推定誤差を模倣するために, 推定誤差量の相対的な割合を表す時間周波数行列 $\boldsymbol{R} \in [0, \alpha]^{I \times J}$ を作成する．$\boldsymbol{R}$ は次式で定義される．

$$\boldsymbol{R} = \begin{bmatrix} r_1 & r_1 & \cdots & r_1 \\ r_2 & r_2 & \cdots & r_2 \\ \vdots & \vdots & \ddots & \vdots \\ r_I & r_I & \cdots & r_I \end{bmatrix} \in [0, \alpha]^{I \times J} \tag{4.11}$$

ここで，$r_1, r_2, \cdots, r_I$ は区間 $[0, \alpha]$ に従う一様分布からランダムに生成された値であり，周波数ビン i 毎に異なる乱数が割り当てられる．

完全分離信号 $(\boldsymbol{Z}_{n'})_{n'=1}^{N}$ を用いて，FDICA の推定分離誤差を模倣した振幅スペクトログラム $|\grave{\boldsymbol{Z}}_{n'}|$ は以下の式で表される．

$$|\grave{\boldsymbol{Z}}_{n'}| = \boldsymbol{R} \odot \left(\sum_{\tilde{n} \neq n'}^{N} |\boldsymbol{Z}_{\tilde{n}}| \right) + (\boldsymbol{1} - \boldsymbol{R}) \odot |\boldsymbol{Z}_{n'}| \in \mathbb{R}^{I \times J} \tag{4.12}$$

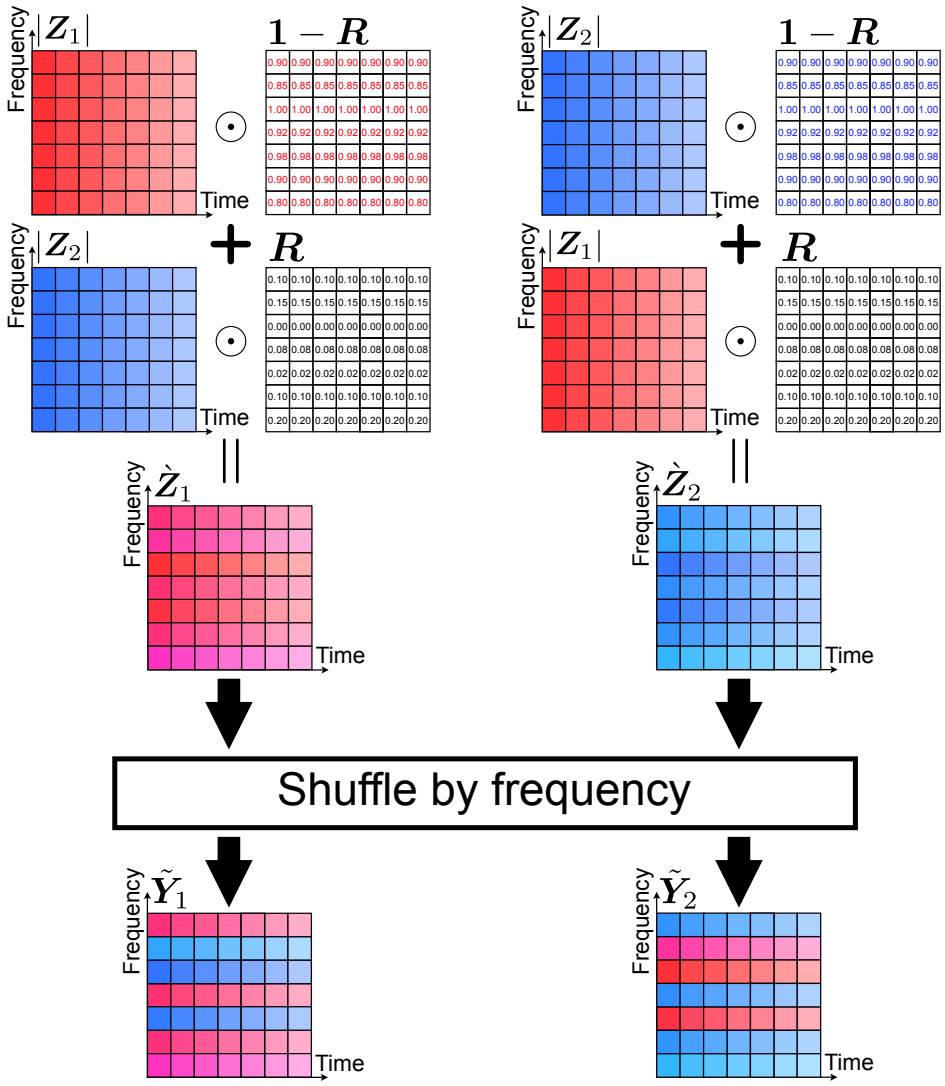

Fig.4.3. Simulation of estimation error in FDICA ($N = 2$).

ここで，$\mathbf{1}$ は全ての要素が 1 であるサイズ $I \times J$ の行列，$\odot$ は要素ごとの積を示す．

式 (4.12) の処理により，FDICA の推定誤差を模倣した振幅スペクトログラムが生成される．この模倣信号は，各周波数成分に区間 $[0, \alpha]$ の割合で他の音源の成分を含ませることができ，FDICA で生じる分離誤差の影響を再現している．その後，FDICA の推定分離誤差を模倣した振幅スペクトログラム $(|\hat{\boldsymbol{Z}}_{n'}|)_{n'=1}^{N}$ の各周波数成分において順番を不揃いにすることで，パーミュテーション問題を含む振幅スペクトログラム $(|\tilde{\boldsymbol{Y}}_{n'}|)_{n'=1}^{N}$ が得られる．FDICA の推定分離誤差を模倣した学習データを作成する一連の流れを Fig. 4.3 に示す．これにより，実際の観測環境における FDICA で生じる推定分離誤差とパーミュテーション問題を含むデータが得られるため，提案 DPS を FDICA に適用した際の性能向上が期待される．

DNN 学習データでは，FDICA の分離誤差及びパーミュテーション問題を模倣するために，各サンプルに対して異なる $\boldsymbol{R}$ を用いて Fig. 4.3 の処理を行った．テストデータには，JVS

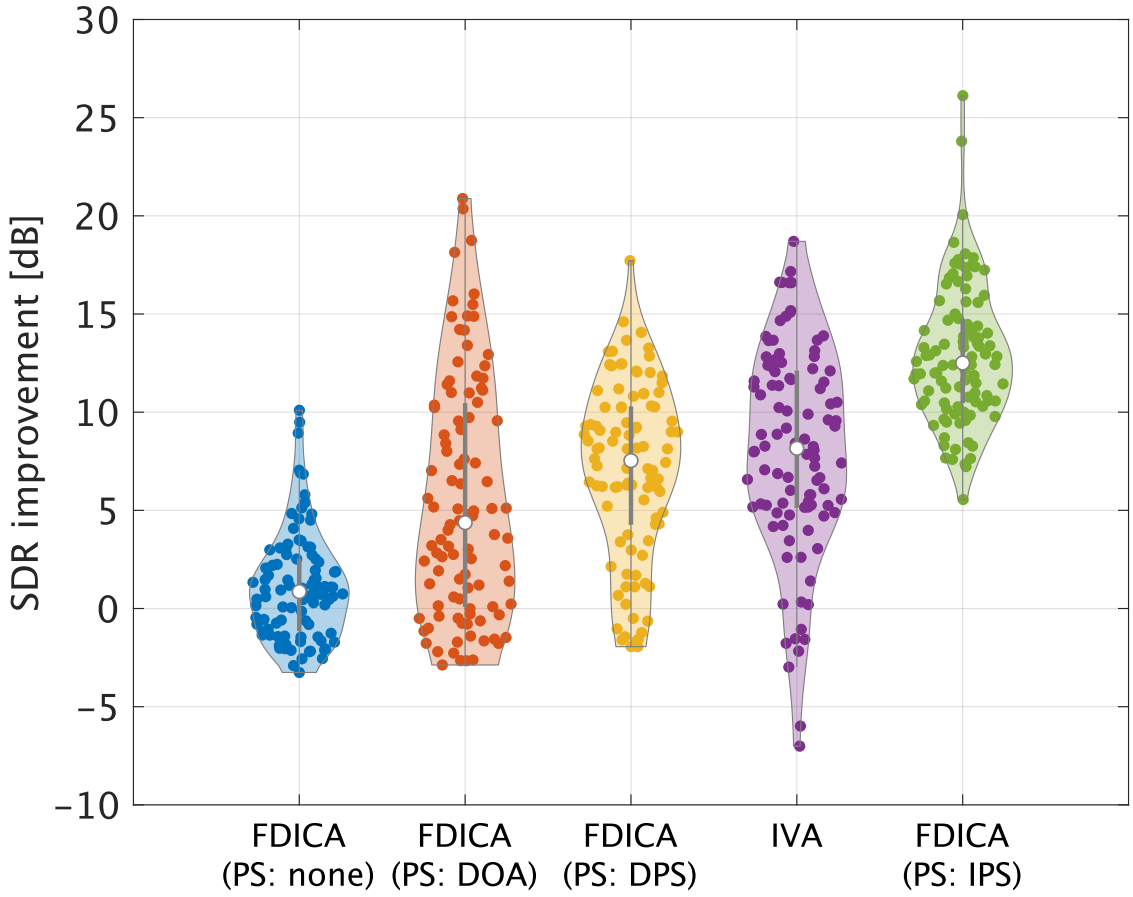

Fig.4.4. Violin plot of SDR improvements.

コーパス [36] に含まれる男女の 100 セット分の音声信号（nonpara30）を用いた．これらの音声信号に対しても pyroomacoustics を用いて，学習データと同一条件でインパルス応答を生成し，観測信号を作成した．サンプリング周波数，STFT の条件，及び DNN の条件は 4.2.1 節と同様である．提案 DPS においては，式 (4.11) における α を 0.2 に設定した．

4.3.2　実験結果

Fig. 4.4 に，提案手法を含む各手法のテストデータ 100 セットにおける SDR 改善量の分布を示すバイオリン図を示す．バイオリン図の色付き点は，各テストデータにおける 2 種類の男女音声信号の平均 SDR 改善量を表しており，中央の白点は中央値，グレーの縦棒は四分位範囲，曲線はカーネル密度推定による分布を示している．

提案手法である FDICA (PS: DPS) の SDR 改善量の中央値は約 7.5 dB であり，テストデータにおいて FDICA によるパーミュテーション問題をある程度解決できることが確認された．また，FDICA (PS: DPS) の SDR 改善量の最小値は-2.2 dB 程度であり，各手法の中で比較的ばらつきが少ないという特徴が見られる．

一方，FDICA (PS: IPS) は真の音源信号を用いた理想的なパーミュテーション解決法であ

り，FDICA におけるパーミュテーション問題解決性能の上限を示している．これに対して，FDICA (PS: DPS) の性能はやや劣る結果となり，完全なパーミュテーション問題解決には至っていない．FDICA（PS: DOA）は，ばらつきが大きく安定性に欠ける結果となった．一方で，IVA は SDR 改善量の中央値が 8.2 dB と比較的高い性能を示しつつも，ばらつきがやや大きい．

これらの結果より，提案手法である FDICA (PS: DPS) は，Table 4.6 に示す音楽信号を用いてワンショット学習を行ったモデルであっても，異なる音源種類である男女の音声信号に対するパーミュテーション問題を解決する能力を有していることが示唆された．さらに，提案手法によって得られた SDR 改善量はばらつきが小さく，安定した性能を示した．これらの特性は，提案手法が音源の種類や学習データの制約に依存せず，高い汎化性能を持つことを示している．しかしながら，実環境における提案手法の応用にはまだ性能が不足しているため，今後の改善が求められる．

4.4　まとめ

本章では，提案 DPS の性能評価を目的とした一連の実験について述べた．まず，4.2 節では，提案 DPS の評価結果を示した．音声信号及び音楽信号に対する in-domain 及び out-of-domain の実験において，提案 DPS は従来手法と比較して，パーミュテーション問題をより高精度に解決できることを示した．特に in-domain の実験では，ほぼすべてのテストデータにおいて 50 dB 以上の SDR 改善量が得られた．50dB 以上の SDR 改善量は，人間の耳で確認するレベルでは，完璧に近い精度で分離されているように聞こえるが，これは残響や雑音のない信号を用いた条件下であり，実環境を模擬したものではなく，提案 DPS が正しく動作するかを確認する目的で行われたものである．一方，out-of-domain の条件下では，音楽信号で学習したモデルが音声信号のパーミュテーション問題を一定程度解決できる結果を示し，少量の学習データで汎化性能を持つモデルの構築が可能であることを示唆した．

次に，4.3 節では，提案 DPS を実際の FDICA に適用した際の評価結果を示した．FDICA の推定誤差を考慮した提案手法は，音声信号におけるパーミュテーション問題を一定程度解決できることが確認された．特に，提案 DPS を用いた BSS の SDR 改善量の中央値は 7.5 dB 程度であり，従来の PS に比べて安定した性能を示した．この結果より，提案 DPS は良好な精度で分離できているが，IVA の SDR 改善量の中央値と比較すると，提案 DPS の性能が劣っていることもあり，実環境への応用には依然として課題が残る．しかしながら，テストデータにおける SDR 改善量の分布を Fig. 4.4 に示したように，提案 DPS はばらつきが少なく，従来手法に対して堅牢性が向上していることが確認された．また，FDICA（PS: IPS）と比較すると，提案 DPS の性能は理想的なパーミュテーション解決には到達していないが，提案 DPS は省サンプルの音楽信号で学習した DNN モデルが音声信号に対して，パーミュテーション問題解決性能を持つことが示された．

以上の結果より，提案手法は，省サンプルの学習データで，パーミュテーション問題を効果

的に解決するための有望なアプローチであることが示された．しかしながら，実環境における提案手法を応用にはまだ性能が不足しているため，新たな DNN モデル構造（Transformer ベースのモデル）を使用する等の今後の改善が求められる．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 におけるパーミュテーション問題を解決するために，新たな DPS を提案し，その有効性を検証した．提案 DPS では，音源数の増加に伴うアルゴリズムの複雑性を抑えつつ，パーミュテーション問題を解決可能とした．また，FDICA の推定分離信号に含まれる誤差を考慮した学習データを用いることで，実環境下における頑健性の向上も図った．

提案 DPS において，正規化された局所時間振幅スペクトログラムを DNN の入力として用い，DNN 出力の確率値を基に周波数ビン毎のパーミュテーション行列を予測する．さらに，パーミュテーション問題は時間方向に対して一貫性を保持しているため，時間方向に対して多数決処理を導入し，予測パーミュテーション行列の精度を向上させた．損失関数には MSE を用い，正解分離信号との間で誤差逆伝播を行うことで，DNN を学習した．

実験結果では，提案手法が音源の種類や学習データの制約に依存せず，高い汎化性能を持つことを示した．特に，音楽信号や音声信号を用いた in-domain 及び out-of-domain の評価において，提案 DPS が従来 DPS を上回る SDR 改善量を達成した．また，FDICA に基づく推定誤差を考慮した提案 DPS は，少量の学習データで DNN を構築しているにも関わらず，ある程度パーミュテーション問題を解決することを示した．

最後に，本研究の課題と今後の展望について述べる．提案手法は音源数 $N = 2$ の場合に限定して評価を行ったが，音源数 $N \geq 3$ の条件下での性能検証が残されている．また，DNN の構造についても，現行の BiLSTM に代わる新たなモデル構造（例えば，Transformer ベースのモデル）を採用することで，さらなる性能向上が期待される．

本研究は，深層学習を活用したパーミュテーション問題解決における新たな可能性を示したものであり，音源分離の精度向上や実環境での応用に向けた基盤を提供するものである．本研究が今後の音源分離技術の発展に貢献することを期待する．

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，本研究を遂行するにあたり，日々多忙な中，終始熱心にご指導くださった指導教員の北村大地准教授に，心の底から感謝申し上げます．北村准教授には，研究の進め方だけでなく，問題解決へのアプローチ，学問への真摯な姿勢，そして一人の研究者としての在り方まで，人生においてかけがえのない学びを与えていただきました．北村准教授と過ごした時間を振り返るたび，その広い見識と深い洞察力，そして常に学生の成長を第一に考えてくださる温かさに触れ，自分がどれほど恵まれた環境にいるのかを実感しています．特に，研究が停滞しているときや挫折しそうなとき，北村准教授が掛けてくださった一言一言が，困難を乗り越える原動力となりました．また，学問だけでなく人間としての在り方をも学ばせていただいたことに，心から感謝しています．

また，専攻科 1 年次までずっと，北村研究室で共に時間を過ごした川口氏，溝渕氏，村田氏には，研究生活だけでなく，日々の生活の中で多くの助言を頂きました．特に研究における姿勢や困難を乗り越える精神を学ぶ機会を与えてくださり，心より感謝いたします．あの何気ない日々の会話や笑い合った瞬間の数々が，私の心に深く刻まれています．

また，研究室の同期である綾野氏には，研究の議論を通じて刺激を受け，私自身の成長を支えていただきました．年齢は 1 つ違いますが，それでもフレンドリーに接してくれたことが，私にとってどれほど心の支えとなっていたか，言葉に尽くせません．そして，後輩である加藤氏，鈴木氏，和気氏，小川氏，谷野宮氏，夏山氏には，研究室生活を温かく彩ってくれたことに感謝しています．皆さんと共に過ごした日々は数少ないですが，日々の会話や活動が，研究に専念できる環境を作り出してくれました．本当に感謝しています．

最後になりますが，私の人生において最も大切な存在である両親には，言葉では表せないほどの感謝の気持ちでいっぱいです．どんな時も私の決断を尊重し，心からの応援を続けてくれたお二人のおかげで，私はここまで進むことができました．時に辛いことがあっても，両親が支えてくれるという安心感が，私の原動力でした．この場を借りて，人生の中で何よりもお二人に感謝を述べたいと思います．いつも温かく見守り，そして支えてくださって本当にありがとうございます．

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. 12, pp. 1–14, 2019.

[2] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[4] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[5] 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.

[6] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," *in Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 3247–3250, 2007.

[7] S. Emura, "Permutation-alignment method using manifold optimization for frequency-domain blind source separation," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 601–605, 2024.

[8] 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 Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[9] 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 Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[10] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," *in Proceedings of International Conference on*

Independent Component Analysis and Signal Separation (ICA), pp. 601–608, 2006.

[11] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: an extension of ICA to multivariate components," *in Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA)*, pp. 165–172, 2006.

[12] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[13] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," *in Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 189–192, 2011.

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

[15] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[16] 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.

[17] Y. Liang, S. M. Naqvi, and J. A. Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electronics Letters*, vol. 48, no. 8, pp. 460–462, 2012.

[18] 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.

[19] L. Li, H. Kameoka, and S. Seki, "HBP: An efficient block permutation solver using Hungarian algorithm and spectrogram inpainting for multichannel audio source separation," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 516–520, 2022.

[20] K, Harold W., "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.

[21] 山地修平, 中嶋大志, 若林佑幸, 小野順貴, "ハンガリー法を用いたパーミュテーション解法に基づくブラインド音源分離," 日本音響学会 *2021* 年秋季研究発表会講演論文集, pp. 305–306, 2021.

[22] A. Horn, "Doubly stochastic matrices and the diagonal of a rotation matrix," *American Journal of Mathematics*, vol. 76, no. 3, pp. 620–630, 1954.

[23] K. Matsumoto and K. Yatabe, "Subband splitting: simple, efficient and effective tech-

nique for solving block permutation problem in determined blind source separation," *IEEE Signal Processing Letters*, 2024 (in press).

[24] S. Yamaji and D. Kitamura, "DNN-based permutation solver for frequency-domain independent component analysis in two-source mixture case," *in Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 781–787, 2020.

[25] 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.

[26] P. A. Absil, R. Mahony, and R. Sepulchre, "Optimization algorithms on matrix manifolds," *Princeton University Press*, 2008.

[27] 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.

[28] D. Kitamura, N. Ono, and H. Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," *in Proceedings of European Signal Processing Conference (EUSIPCO)*, pp. 1210–1214, 2017.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[30] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[31] G. Birkhoff, "Three observations on linear algebra," *Universidad Nacional de Tucuman Revista*, vol. 5, pp. 147–151, 1946.

[32] D. Yu, M. Kolbak, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2017.

[33] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011): audio source separation," *in Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 414–422, 2012.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, 1412.6980, 2014.

[35] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: a python package for audio room simulation and array processing algorithms," *in Proceedings of IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 351–355, 2018.

[36] 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.

発表文献一覧

査読付き国際会議

1. **Fumiya Hasuike**, Daichi Kitamura, and Rui Watanabe, “DNN-based frequency-domain permutation solver for multichannel audio source separation,” in Proceedings of *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 872–877, 2022.

国内学会

1. **蓮池郁也**, 渡辺瑠伊, 北村大地, “深層ニューラルネットワークに基づくパーミュテーション解決法の基礎的検討,” 信学技報, EA2022-13, vol. 122, no. 20, pp. 62–67, 2022.
2. **蓮池郁也**, 北村大地, 渡辺瑠伊, “深層パーミュテーション解決法の汎化性能に関する実験的評価,” 日本音響学会 2022 年秋季研究発表会講演論文集, pp. 351–354, 2022.
3. **蓮池郁也**, 北村大地, 渡辺瑠伊, 川口翔也, “周波数双方再帰に基づく深層パーミュテーション解決法,” 電子情報通信学会 第 37 回信号処理シンポジウム, A13–2, pp. 308–313, 2022.
4. **蓮池郁也**, 北村大地, “深層パーミュテーション解決法に基づくブラインド音源分離の性能評価,” 日本音響学会 2024 年秋季研究発表会講演論文集, 1-R-25, pp. 235–238, 2024.

付録 A

Birkhoff–von Neumann の定理

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

$$\boldsymbol{D} = \sum_{i=1}^{N!} \sigma_i \boldsymbol{P}_i \tag{A.1}$$

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

付録 B

提案 DPS の実験結果

4.2 項で，提案 DPS の性能を評価した際のスペクトログラムの結果を以下に掲載する．

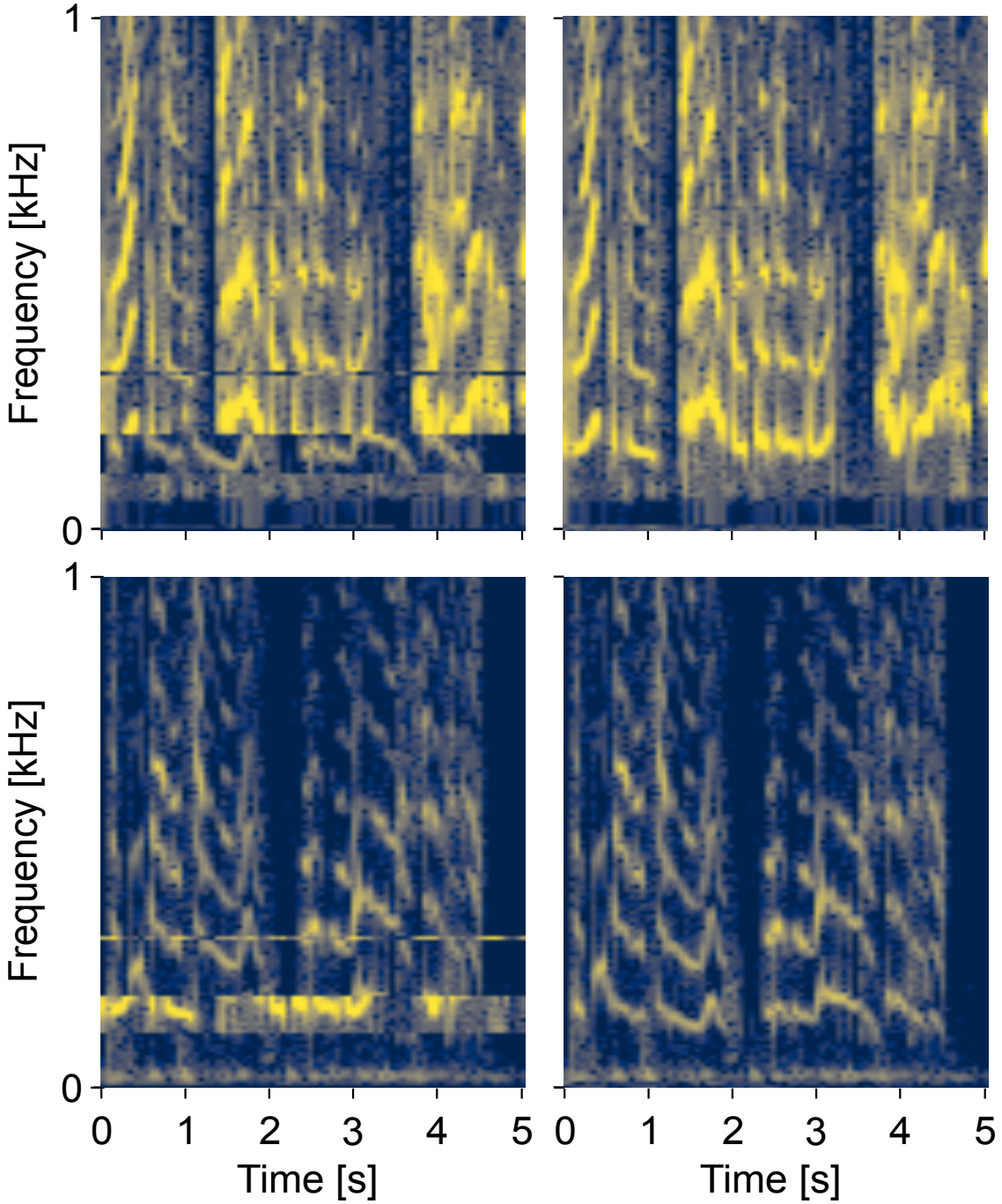

Fig.B.1. Spectrograms (in-domain evaluation) for speech test data using DPS trained with speech signals, conventional DPS (left), proposed DPS (right).

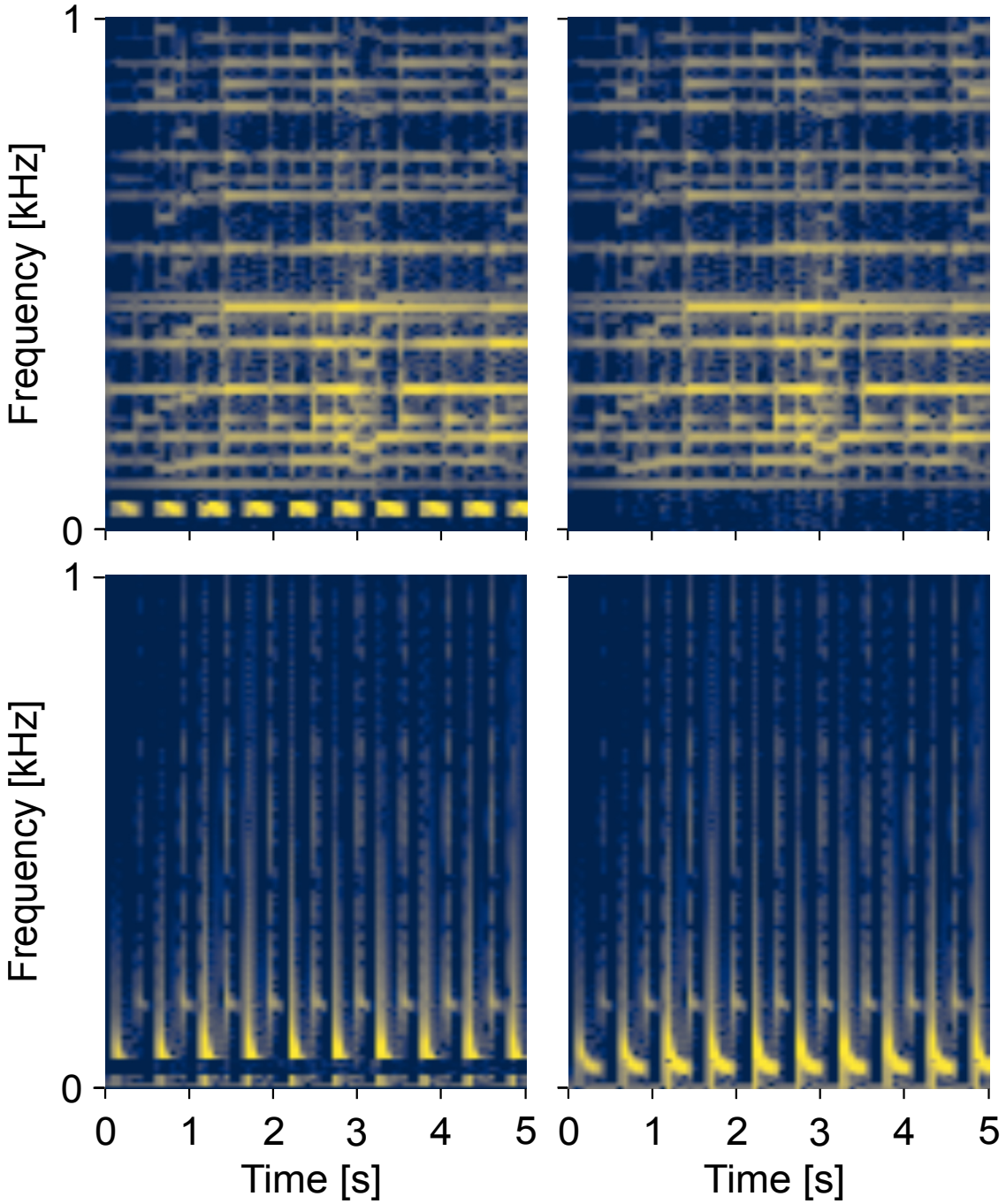

Fig.B.2. Spectrograms (in-domain evaluation) for music test data using DPS trained with music signals, conventional DPS (left), proposed DPS (right).

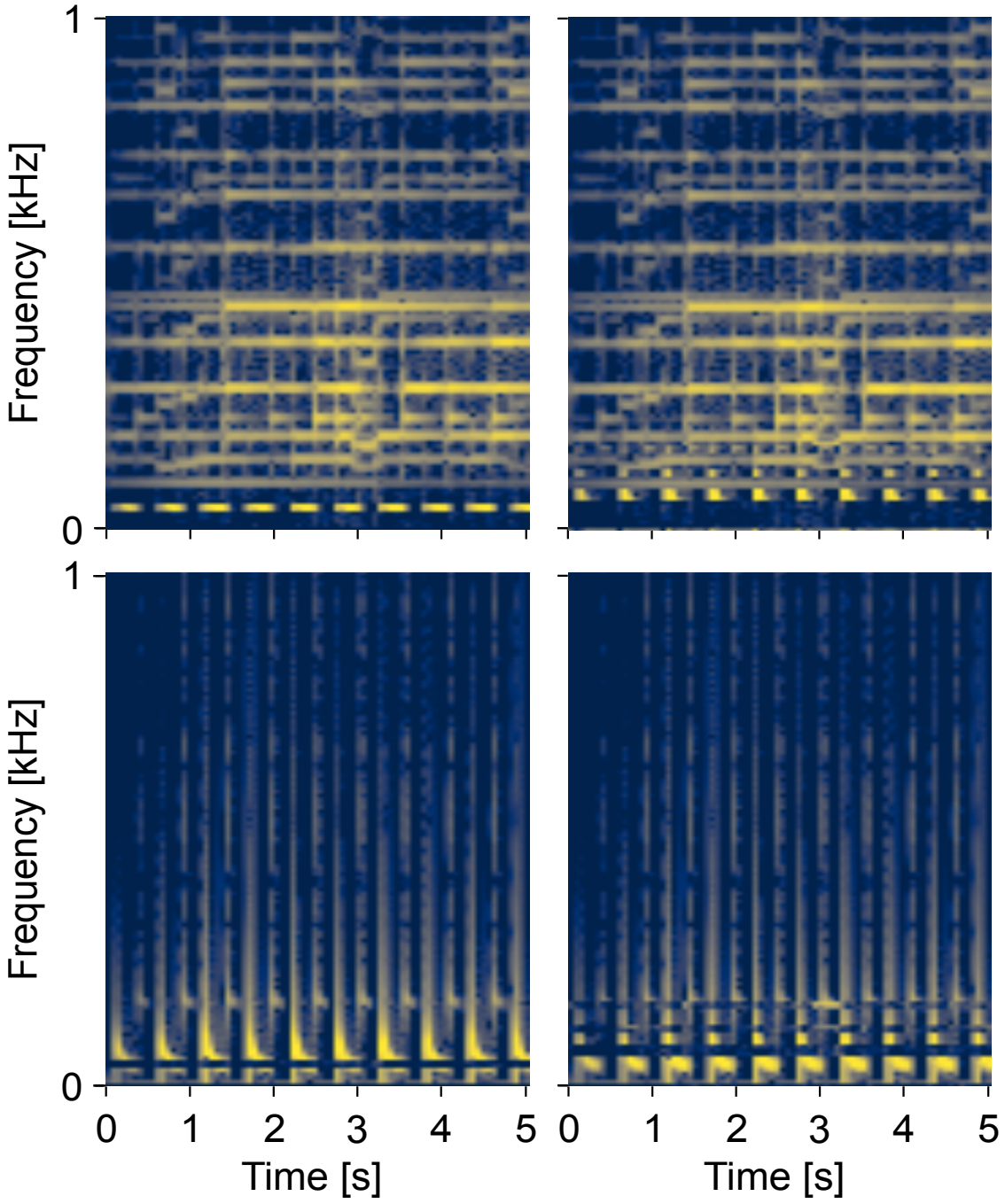

Fig.B.3. Spectrograms (out-of- domain evaluation) for music test data using DPS trained with speech signals, conventional DPS (left), proposed DPS (right).

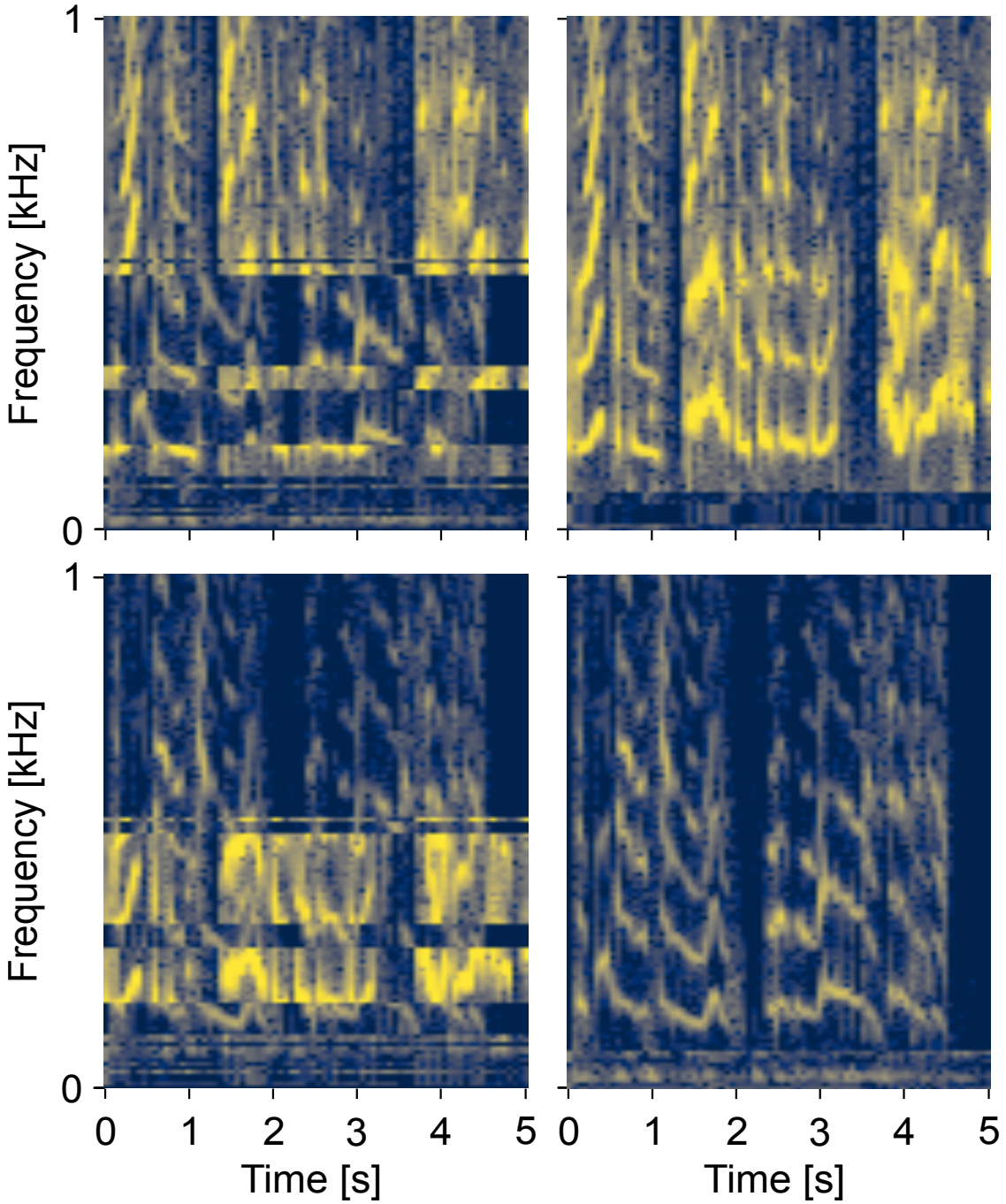

Fig.B.4. Spectrograms (out-of- domain evaluation) for speech test data using DPS trained with music signals, conventional DPS (left), proposed DPS (right).