特別研究論文

（査読済み）

研　究　題　目

独立低ランク行列分析を用いたインタラクティブ音源分離システム

提出年月日		2021 年　2 月　1 日
氏	名	大島 風雅
主	査	北村 大地
副	査	重田 和弘
副	査	柿元 健

香川高等専門学校
専攻科
創造工学専攻

Interactive Audio Source Separation System Using Independent Low-Rank Matrix Analysis

Fuga Oshima

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

Abstract

Audio source separation is a technique for separating original sources from an observed mixture signal. For example, it can be used for a pre-process of an automatic speech recognition system and remastering of existing music signals by users. In the former ap- plication, audio source separation is utilized for separating a target speech signal and the other speech or background noise. Also, in the latter application, vocals and individual musical instruments are estimated by the technique. In particular, audio source separation that does not utilize any prior information about mixing system is called blind source separation (BSS). BSS has mainly been developed based on the algorithm called independent component analysis (ICA). The state-of-the-art BSS algorithm is called independent low-rank matrix analysis (ILRMA), which can achieve high quality audio source separation in (over-)determined situations. However, the separation performance is of- ten degraded because of the dependence of initial values for optimization parameters and the difference of optimization speeds of source and spatial models. This is due to the occurrence of a mismatch of estimated source permutations in certain ranges (blocks) of frequency bands, which is called the block permutation problem. In this thesis, I investigate the appropriate balance of optimization speeds between source and spatial models and clarify the relationship between the behavior of the cost function and the separation performance in ILRMA. The experimental results show that the convergence behaviors of the cost function value are significantly different in the cases of high- and low-quality separation in ILRMA. This fact implies that we can expect the failure or success of separation during the iterative optimization. On the basis of this findings, I extend the interactive audio source separation system, which utilizes user annotations to achieve more robust and precise audio source separation.

Key Words: audio source separation, independent low-rank matrix analysis, user interaction

（和訳）

音源分離とは，複数の音源が混合している観測信号から混合前の音源を推定する技術である．この技術は，例えば音声認識の前段処理や既存の音楽信号のリマスタリング等に用いられる．前者では，目的話者の音声信号とその他の音声や背景雑音等の分離に音源分離が活用される．また後者においては，ボーカルや個々の楽器音信号が音源分離によって推定される．特に，音源の混合系に関する事前情報を全く用いずに音源を分離する技術はブラインド音源分離（blind source separation: BSS）と呼ばれる．BSS は主に独立成分分析と呼ばれる数理理 論に基づき，これまで発展してきた歴史を持つ．最先端の BSS アルゴリズムは独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）と呼ばれ，優決定な観測条件において高性能な音源分離を実現する．しかしながら，ILRMA の音源分離性能はパラメタの初期値依存性や音源・空間モデルの最適化速度の違いに依存して低下してしまう場合がある．これはブロックパーミュテーション問題と呼ばれる，まとまった周波数帯でのパーミュテーション不整合が発生するためである．本論文では，ILRMA における音源モデル及び空間モデルの目的関数の収束の様子と音源分離性能の関係性を明らかにする．実験結果より，ILRMA の分離性能が良いときと悪いときの目的関数値の収束の様子は明らかに異なっていることを確認した．この事実は，ILRMA の反復最適化の途中であっても，分離の成否をある程度判断できることを示している．さらに，得られた知見に基づいて，ユーザアノテーションを活用するインタラクティブ音源分離システムに拡張することで，より頑健かつ高精度な音源分離システムを提案する．

目次

第 1 章

緒言

1.1 本論文の背景

音源分離とは，複数の音響信号が混ざった観測信号から混合前の信号を推定する技術である．この技術は，Fig. 1.1 に示す例のように，様々な場面で用いられている．例えば，スマートフォンやスマートスピーカーの音声認識機能において入力音声の前段処理として用いられている．これは端末に入力された信号から外部雑音や残響音を取り除いた目的信号のみを抽出することで認識性能を高めるためである．また，楽曲を各楽器ごとに分離することで，一度マスタリングされた楽曲のリマスタリングが可能となり，ユーザによる音楽の再編集を実現することで音楽文化の興隆につながっている．この音楽音源分離技術も既にいくつかの形で実用化が行われている．例として iZotope の RX8 [1] や Deezer の Spleeter [2] などは深層学習を利用して 2 チャネルの混合信号をボーカル，ドラム及びベースの各楽器信号へ分離することができる．特に Spleeter は GitHub で公開されており [3]，簡単に利用することができる．

音源分離手法は主に音源数及びチャネル数の関係によって手法が大別される．Fig. 1.2 にその概要を示す．まず，劣決定条件（音源数 $>$ チャネル数）の場合について述べる．一般的に mp3 形式や wav 形式の音楽信号はチャネル数が 1ch（モノラル）もしくは 2ch（ステレオ）であり，そこに含まれる音源の数は 2 つより大きいことの方が多いため，通常の音楽音源分離は劣決定条件の下で行われる．しかし，劣決定条件の音源分離を音源の性質や定位，残響長などの事前情報が無い状態で実現するのは困難である．そこで，劣決定条件での音源分離は楽器信号特有の時間周波数構造を手掛かりとして，非負値行列因子分解（nonnegative matrix factorization: NMF）[4] を用いた音源分離手法 [5, 6, 7] が提案されている．一方で，決定的条件（音源数 $=$ チャネル数）もしくは優決定条件（音源数 $<$ チャネル数）では，事前情報が無い状態でも音響信号の統計的性質を利用することで分離を推定するブラインド音源分離（blind source separation: BSS）[8] の手法が盛んに研究されている．独立成分分析（independent component analysis: ICA）[9] は BSS の最も一般的な手法で，様々な手法が提案されている．本研究で取り扱う独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[10, 11] もまた ICA を発展させた手法である．

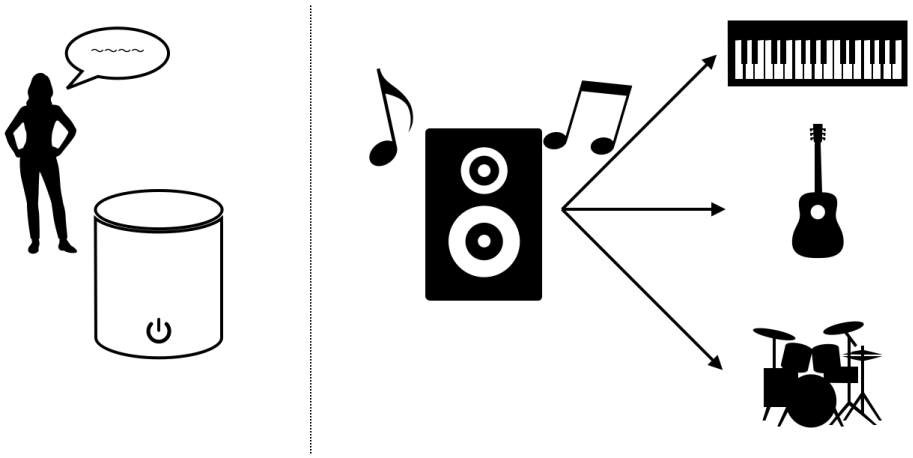

Fig. 1.1. Typical application examples of audio source separation technique.

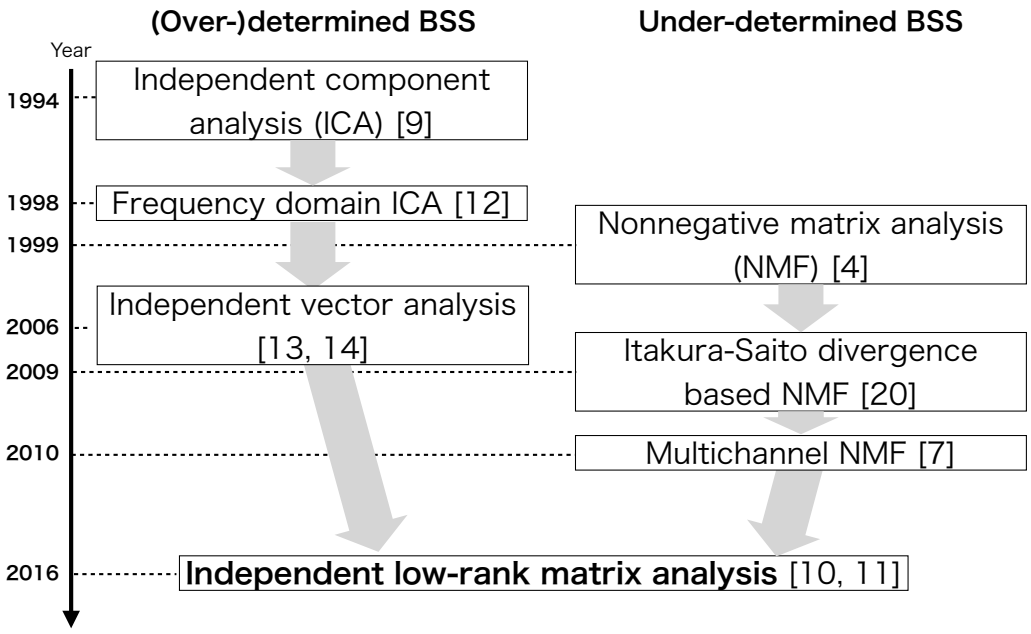

Fig. 1.2. History of audio source separation techniques.

このように盛んに音源分離手法が研究され，人間の聴覚に対して違和感のない高品質な音源分離が実現している．一方で，偶発的に分離性能が低下することがある．これは ICA に基づく手法が反復最適化アルゴリズムであり，その最適化パラメタの初期値が乱数で与えられるために発生するものである．実際の応用では，分離結果がどの程度の性能であったかを知るには主観的に評価するしかないため，パラメタ初期値依存性による分離性能の変動は大きな問題である．そのため，最適化パラメタがどのように初期化されても常に高品質な音源分離が実現されるような初期値頑健性を持つ音源分離が必要とされている．

1.2 本論文の目的

本論文では，最新の BSS 手法である ILRMA を用いた初期値頑健性を持つ音源分離の実現を目指す．ILRMA は反復最適化アルゴリズムである周波数領域 ICA（frequency-domain ICA: FDICA）[12] 及び独立ベクトル分析（independent vector analysis: IVA）[13, 14] といった BSS 手法に NMF による音源モデルを導入し，観測信号を分離するフィルタ（空間モデル）と音源モデルを交互に反復最適化することでより高性能な音源分離を実現している．しかし音源モデルの初期値の乱数によって分離性能が低下してしまうことがある．これは交互に反復最適化する二つのモデルパラメタの最適化速度のバランスが悪いために局所最適解に陥りやすいためである．

本論文では，より良い分離が達成される最適化速度を実験により調査するとともに，二つのモデルの目的関数の挙動と分離性能の相関関係を調査するものである．また，それにより得られた知見を Fig. 1.3 のように既存のインタラクティブ音源分離システムに導入する．既存のインタラクティブ音源分離システムには，システムを初めて使うユーザには分離失敗を判定し，アノテーションを与えるのが難しいという問題がある．しかし，本研究で得られた知見を活かすことでユーザの音源分離性能の判断基準が増え，アノテーションを誤って与えないようなシステムへと改善できる．

1.3 本論文の構成

2 章では，既存のブラインド音源分離の手法及びインタラクティブ音源分離システムの解説を行う．3 章では本論文で取り扱う手法のパラメタと性能の相関関係を実験により調査する．4 章では 3 章で得られた知見を活かして既存のインタラクティブ音源分離システムの拡張を行

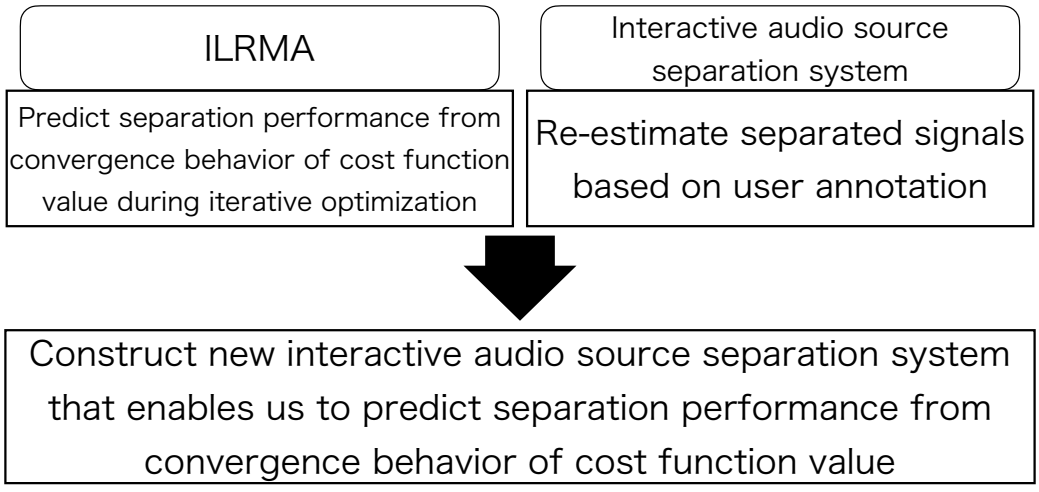

Fig. 1.3. Extension of interactive audio source separation system.

う．5 章では全ての章を総括した結言を述べる．

第 2 章

既存手法

2.1 まえがき

本章では，BSS の手法について説明する．まず，2.2 節では，短時間フーリエ変換について説明する．次に 2.3 節では，時間領域 BSS の問題点とパーミュテーション問題について説明する．2.4 節では，周波数領域 BSS の定式化を行う．2.5 節では，FDICA で発生する問題の解決のために提案された IVA，2.6 節では行列分解の手法の一つである NMF，2.7 節では，本研究で扱った ILRMA についてそれぞれ説明する．2.8 節では，反復更新式を一般化させた ILRMA について説明する．また，2.9 節では ILRMA を用いたインタラクティブ音源分離システムについて，2.10 節ではシステムのパラメタの処理方法について説明する．

2.2 短時間フーリエ変換

短時間フーリエ変換（short-time Fourier transform: 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 番目の短時間区間信号（時間フレーム）は次式で表される．

$$\begin{aligned}\boldsymbol{z}^{[j]} &= [z[(j-1)\tau+1], z[(j-1)\tau+2], \cdots, z[(j-1)\tau+Q]]^{\mathrm{T}} \\ &= [z^{[j]}[1], z^{[j]}[2], \cdots, z^{[j]}[q], \cdots, z^{[j]}[Q]]^{\mathrm{T}} \in \mathbb{R}^{Q}\end{aligned} \tag{2.1}$$

ここで，$l = 1, 2, \cdots, L$，$j = 1, 2, \cdots, J$ 及び $q = 1, 2, \cdots, Q$ はそれぞれ離散時間のインデクス，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，時間フレーム数 J は次式によって与えられる．

$$J = \frac{L}{\tau} \tag{2.2}$$

ただし，信号長 L はセグメント数 J が整数となるように各時間フレームの信号の両端にゼロを挿入する処理（ゼロパディング）が施される．そして，信号 $\boldsymbol{z} = [z[1], z[2], \cdots, z[L]]^{\mathrm{T}} \in \mathbb{R}^{L}$

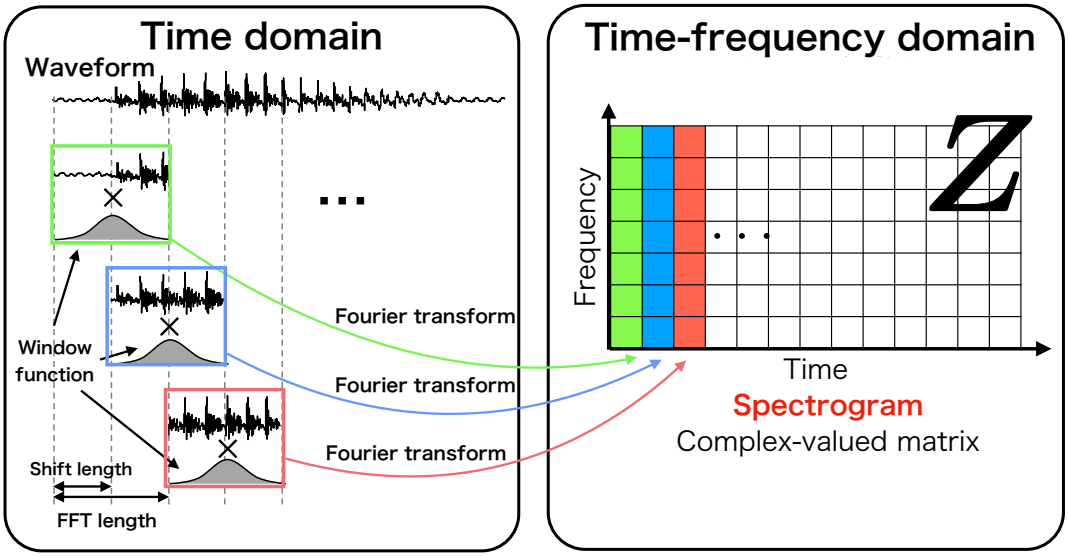

Fig. 2.1. Mechanism of STFT.

の STFT は次式のように表される．

$$\boldsymbol{Z} = \mathrm{STFT}_{\boldsymbol{\omega}}(\boldsymbol{z}) \in \mathbb{C}^{I \times J} \tag{2.3}$$

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．

$$z_{ij} = \sum_{q=1}^{Q} \omega[q] z^{[j]}[q] \exp\left\{ \frac{-\imath 2\pi (q-1)(i-1)}{F} \right\} \tag{2.4}$$

ここで $i = 1, 2, \cdots, I$ は周波数ビンのインデクスを，F は $\lfloor \frac{F}{2} \rfloor + 1 = I$ を満たす整数（$\lfloor \cdot \rfloor$ は床関数）を，$\imath$ は虚数単位を，$\boldsymbol{\omega}$ は分析窓関数を示している．また，逆 STFT は合成窓関数 $\tilde{\boldsymbol{\omega}}$ を用いて定義され，$\mathrm{ISTFT}_{\tilde{\boldsymbol{\omega}}}(\cdot)$ と記述される．

このように，時間領域の信号に対して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．

2.3　ICA と FDICA

ICA は優決定条件（マイクロホン数 $>$ チャネル数）もしくは決定条件（マイクロホン数 $=$ チャネル数）における BSS として最も一般的な手法で，音源間の独立性が最大になるような分離行列を時間領域で推定する手法である．このような時間領域の BSS は時間領域の信号と混合行列の積で表される．しかし，実際の音響信号の混合は残響の影響で畳み込み混合になるため，単純な ICA で分離信号を推定することはできない．そこで，時間領域の観測信号に STFT を施して時間周波数領域の信号に変換し，各周波数において独立に ICA を適用する FDICA [12] が提案されている．時間周波数領域では，残響長に対して STFT の窓関数長が十

分に大きい場合に，時間領域での畳み込み混合を時間周波数領域での瞬時混合（行列積）で表現できるため，FDICA は残響を含む混合信号の音源分離が可能となる．次節にて FDICA における各信号の定式化とパーミュテーション問題について説明する．

2.4 周波数領域における BSS（FDICA）の定式化

周波数領域の BSS の定式化を行う．複数の音源が混合している観測信号中の音源数及びチャネル数をそれぞれ N 及び M と定義する．Fig. 2.2 に $N = M = 2$ の場合の BSS の概略を示す．混合前の音源，観測信号及び分離信号に対して STFT したものをそれぞれ

$$\boldsymbol{s}_{ij} = (s_{ij,1}, s_{ij,2}, \cdots, s_{ij,n}, \cdots, s_{ij,N})^{\mathrm{T}} \in \mathbb{C}^N \tag{2.5}$$

$$\boldsymbol{x}_{ij} = (x_{ij,1}, x_{ij,2}, \cdots, x_{ij,m}, \cdots, x_{ij,M})^{\mathrm{T}} \in \mathbb{C}^M \tag{2.6}$$

$$\boldsymbol{y}_{ij} = (y_{ij,1}, y_{ij,2}, \cdots, y_{ij,n}, \cdots, y_{ij,N})^{\mathrm{T}} \in \mathbb{C}^N \tag{2.7}$$

と定義する．ここで，$n = 1, 2, \cdots, N$ 及び $m = 1, 2, \cdots, M$ はそれぞれ音源及びチャネルのインデクスを表す．式 (2.5) から式 (2.7) と周波数ごとの時不変な（時間フレーム j に依存しない）混合行列 $\boldsymbol{A}_i \in \mathbb{C}^{M \times N}$ を用いると，混合時の残響長が STFT の短時間信号長（窓長）L よりも十分に短い場合，観測信号は次式で表せる．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{2.8}$$

ここで，$M = N$ かつ $\boldsymbol{A}_i$ がフルランクの場合は，分離行列 $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}\ \boldsymbol{w}_{i2}\ \cdots\ \boldsymbol{w}_{iN})^{\mathrm{H}} \in \mathbb{C}^{N \times M}$ が存在し，分離信号は次式で表せる．

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{2.9}$$

ここで，$^{\mathrm{H}}$ は行列またはベクトルのエルミート転置を表す．BSS は混合行列 $\boldsymbol{A}_i$ が未知の状態で分離行列 $\boldsymbol{W}_i$ を全ての周波数 $i = 1, 2, \cdots, I$ において推定する問題である．

ICA を周波数領域に拡張した FDICA では，周波数領域の分離行列 $\boldsymbol{W}_i$ に次式の任意性が存在する．

$$\boldsymbol{W}_i = \boldsymbol{D}_i \boldsymbol{P}_i \boldsymbol{W}_i \tag{2.10}$$

ここで，$\boldsymbol{D}_i \in \mathbb{C}^{N \times N}$ 及び $\boldsymbol{P}_i \in \{0, 1\}^{N \times N}$ はそれぞれ任意の対角行列及びパーミュテーション行列である．つまり，分離行列のスケールとパーミュテーションに任意性が存在するため，周波数ごとに分離信号のスケールと順序がばらばらとなる可能性がある．スケールの問題はプロジェクションバック法 [15] を用いて解決できる．しかし，順番の問題はパーミュテーション問題と呼ばれ，解決が困難である．パーミュテーション問題の概要を Fig. 2.3 に示す．FDICA は周波数ビンごとに分離信号を推定するが，分離した信号がどのような順番で推定されるかは不定であり，分離行列の初期値に依存して決まる．よって，周波数ごとの分離が完璧に達成されていたとしても，分離信号が任意のパーミュテーション行列によって交換されるため，全ての周波数ビンにおいてパーミュテーション行列が一致していなければ音源分離は達成されない．パーミュテーション問題は様々な解決法が提案されており，後述の IVA や ILRMA はパーミュテーション問題を回避する手法である．

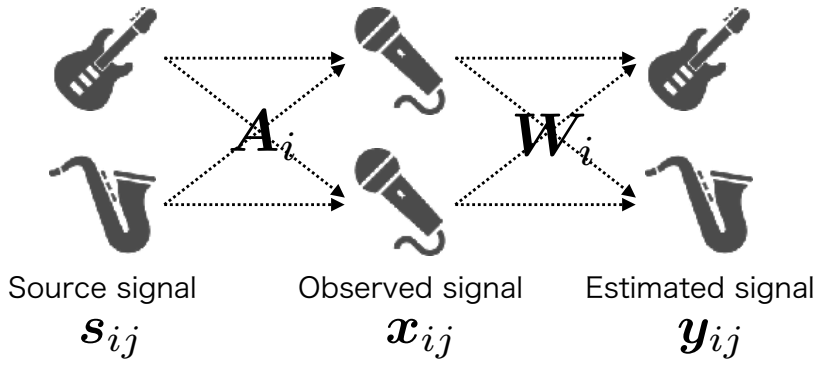

Fig. 2.2. Outline of BSS, where $N = M = 2$.

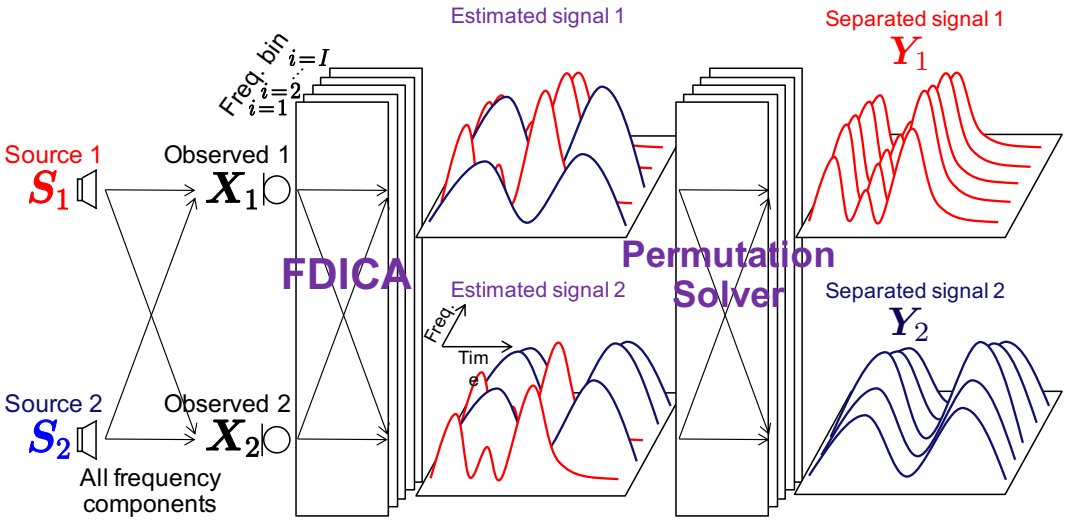

Fig. 2.3. Permutation problem in FDICA.

2.5　IVA

IVA は FDICA の多変量拡張であり，分離行列 $\boldsymbol{W}_i$ の推定と同時にパーミュテーション問題をある程度回避できる BSS である．ここで音源，観測信号及び分離信号それぞれについて，全ての周波数ビンに関する成分をまとめたベクトルを

$$\overline{\boldsymbol{s}}_{j,n} = (s_{1j,n}, s_{2j,n}, \cdots, s_{ij,n}, \cdots, s_{Ij,n})^{\mathrm{T}} \in \mathbb{C}^I \tag{2.11}$$

$$\overline{\boldsymbol{x}}_{j,m} = (x_{1j,m}, x_{2j,m}, \cdots, x_{ij,m}, \cdots, x_{Ij,m})^{\mathrm{T}} \in \mathbb{C}^I \tag{2.12}$$

$$\overline{\boldsymbol{y}}_{j,n} = (y_{1j,n}, y_{2j,n}, \cdots, y_{ij,n}, \cdots, y_{Ij,n})^{\mathrm{T}} \in \mathbb{C}^I \tag{2.13}$$

と定義する．Fig. 2.4 に $M = N = 2$ の場合の混合分離モデルを示す．IVA も FDICA のように周波数ごとに独立した分離行列 $\boldsymbol{W}_i$ を推定する．ただし，推定の過程で全周波数を含む I 次元分布を生成モデルとして仮定し，I 次元ベクトル内の高次相関を仮定している．この IVA の生成モデルには，Fig. 2.5 で示すような球状対称ラプラス分布が用いられ，式 (2.14) で表

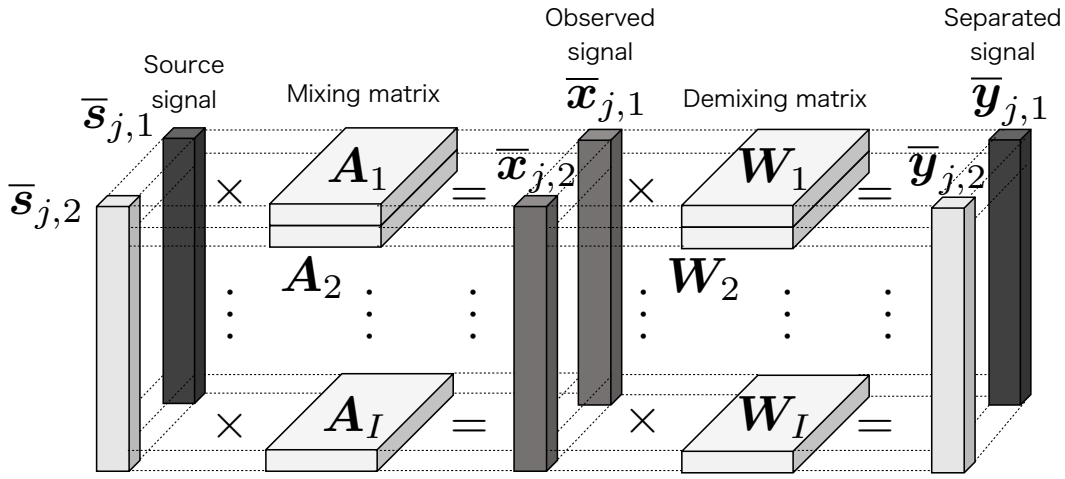

Fig. 2.4. Mixing and demixing model in IVA, where $N = M = 2$.

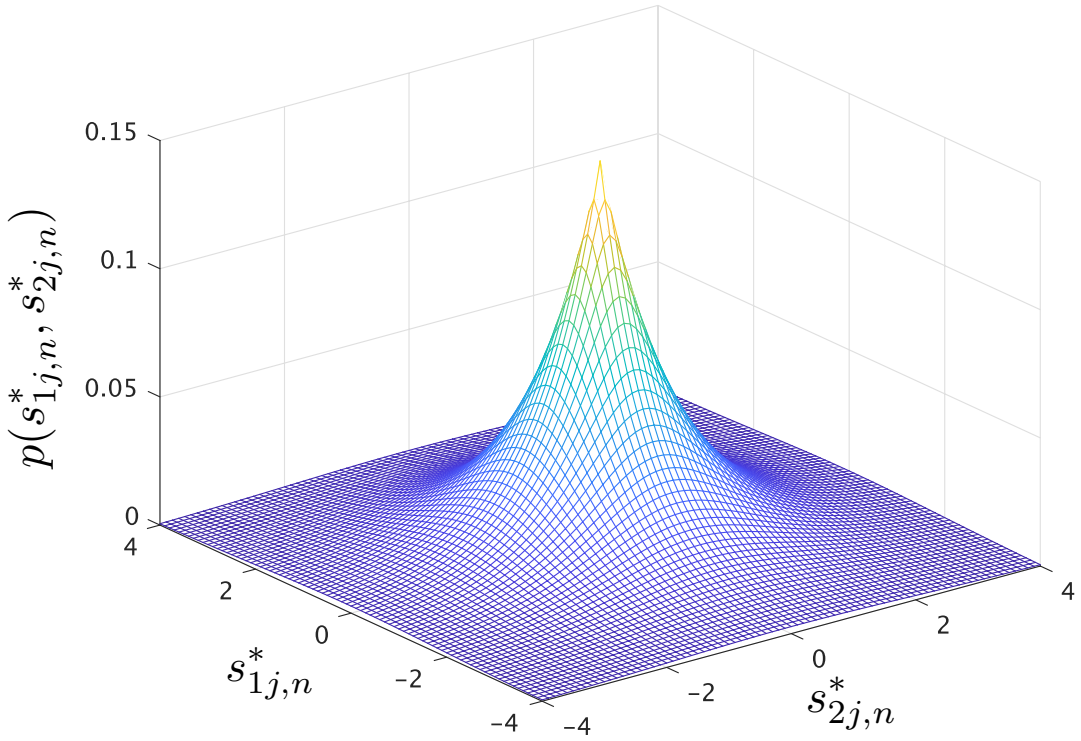

Fig. 2.5. Zero-mean and spherically symmetric Laplace distribution, where $I = 2$ and $s^*_{ij,n}$ can be considered as either real or imaginary part of $s_{ij,n}$.

される．

$$p(\overline{\boldsymbol{y}}_{jn}) = \frac{1}{\pi \prod_i \sigma_{i,n}} \exp\left(-\sqrt{\sum_i \left|\frac{y_{ij,n}}{\sigma_{i,n}}\right|^2}\right) \tag{2.14}$$

ここで，$\sigma_{i,n}$ はスケールパラメタである．式 (2.14) の多次元ラプラス分布は球対称性を持つため，同一ベクトル内の成分が高次相関を持つ．したがって，IVA は同時に生起する周波数成

分を一つの音源としてまとめるという傾向がある．つまり，信号の基本周波数とその倍音成分が同一音源として扱われやすいと言える．このような「同一音源の周波数成分の共起性を仮定した統計的音源モデル」により，FDICA で発生するパーミュテーション問題をある程度回避することができる．音源周波数ベクトル間の独立性 $p(\overline{\boldsymbol{y}}_{j,1}, \overline{\boldsymbol{y}}_{j,2}, \cdots, \overline{\boldsymbol{y}}_{j,N}) = \prod_n p(\overline{\boldsymbol{y}}_{i,n})$ を仮定すると，IVA の観測信号に対する負対数尤度関数は次式で得られる．

$$\mathcal{L} = -2J \sum_i \log |\det \boldsymbol{W}_i| + \sum_{j,n} G(\overline{\boldsymbol{y}}_{j,n}) \tag{2.15}$$

ここで，$G(\overline{\boldsymbol{y}}_{j,n})$ はコントラスト関数と呼ばれ，$G(\overline{\boldsymbol{y}}_{j,n}) = -\log p(\overline{\boldsymbol{y}}_{j,n})$ で定義される．IVA の最適化は補助関数法 [22] を用いた手法によって高速かつ安定に行える [14, 16].

2.6　NMF

NMF [4] とは行列分解の方法の一つである．NMF が他の行列分解の方法である LU 分解や固有値分解と異なるのは非負値行列を対象としている点である．また，NMF は対象とする非負値行列の低ランク性を仮定して分解することで，行列の潜在パタンを抽出できるアルゴリズムである．そして，NMF は劣決定音源分離に適用することが可能である [5, 6, 7]．単一チャネルの音響信号を STFT することで得られるパワースペクトログラムの NMF による分解は次式で表される．

$$|\boldsymbol{Z}|^{\cdot 2} = \boldsymbol{T}\boldsymbol{V} \tag{2.16}$$

ここで，$|\cdot|$ は要素ごとの絶対値を，ドット付きの指数は要素ごとの累乗を示す．よって，$|\boldsymbol{Z}|^{\cdot 2}$ はパワースペクトログラムを表す．また，$\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{I \times K}$ を基底行列，$\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{K \times J}$ をアクティベーション行列という．K は NMF の分解において手動で与えるパラメタであり，基底行列 $\boldsymbol{T}$ の列ベクトルの本数（基底ベクトル数）である．従って，行列積 $\boldsymbol{T}\boldsymbol{V}$ のランクは K と一致し，通常は $K \ll \min(I, J)$ となるように設定される．つまり，Fig. 2.6 に示すように単一チャネルの音響信号を対象とした NMF では，パワースペクトログラムに対して低ランク近似を行うことで，$\boldsymbol{T}$ が音響信号中の頻出スペクトルパタンとなり，$\boldsymbol{V}$ が各スペクトルパタンの発音タイミング（アクティベーション）となるような分解が可能である．また，基底行列 $\boldsymbol{T}$ とアクティベーション行列 $\boldsymbol{V}$ は次式の最小化問題の解として推定される．

$$\begin{aligned} \min_{\boldsymbol{T},\boldsymbol{V}} \mathcal{D}(|\boldsymbol{Z}|^{\cdot 2} | \boldsymbol{T}\boldsymbol{V}) \ \ &\text{s.t. } t_{ik}, v_{kj} \geq 0 \\ &\forall i = 1, 2, ..., I, \ j = 1, 2, ..., J, \ k = 1, 2, ..., K \end{aligned} \tag{2.17}$$

ここで，t_{ik} 及び v_{kj} は $\boldsymbol{T}$ 及び $\boldsymbol{V}$ の要素である．また，$\mathcal{D}(|\boldsymbol{Z}|^{\cdot 2} | \boldsymbol{T}\boldsymbol{V})$ は 2 つの行列（$|\boldsymbol{Z}|^{\cdot 2}$ 及び $\boldsymbol{T}\boldsymbol{V}$）間の類似度を測る関数である．行列の類似度を図る関数には，次式で表される

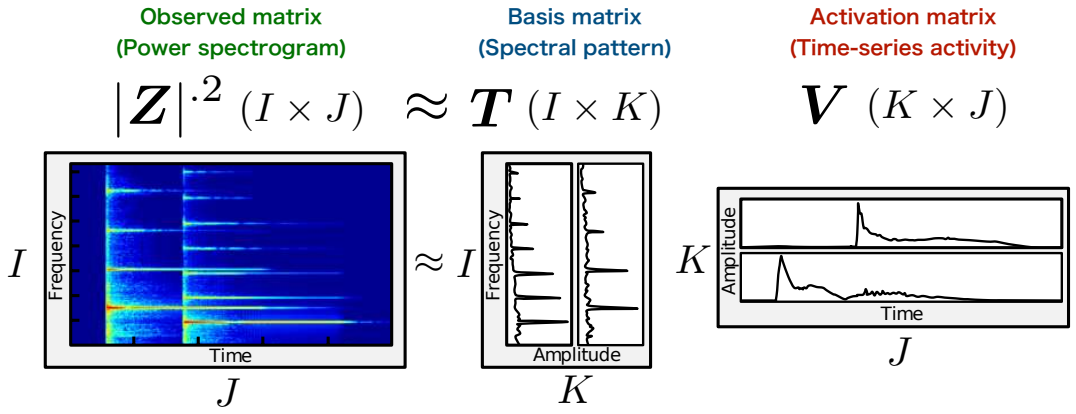

Fig. 2.6. NMF decomposition of power spectrogram, where $|\cdot|$ and dotted exponent for matrices denote entrywise absolute value and entrywise exponent, respectively.

β-divergence [17, 18] がよく利用される [19].

$$d_\beta(a|b) = \begin{cases} \dfrac{a}{b} - \log\dfrac{a}{b} - 1 & (\beta = 0) \\ a\log\dfrac{a}{b} + b - a & (\beta = 1) \\ \dfrac{a^\beta}{\beta(\beta-1)} + \dfrac{b^\beta}{\beta} - \dfrac{ab^{\beta-1}}{\beta-1} & (\text{otherwise}) \end{cases} \tag{2.18}$$

β-divergence の中でも特に $\beta = 2, 1, 0$ の場合はそれぞれ二乗 Euclid 距離，一般化 Kullback–Leibler ダイバージェンス及び Itakura–Saito ダイバージェンスと呼ばれている．これらのうち，ダイバージェンスと名のつくものは以下に示す距離の公理

1. 非負性：$\mathcal{D}(a|b) \geq 0 \quad \forall a, b \in \mathbb{R}$
2. 同一性：$\mathcal{D}(a|b) = 0 \Leftrightarrow a = b \quad \forall a, b \in \mathbb{R}$
3. 対称性：$\mathcal{D}(a|b) = \mathcal{D}(a|b) \quad \forall a, b \in \mathbb{R}$
4. 三角不等式：$\mathcal{D}(a|b) + \mathcal{D}(b|c) \geq \mathcal{D}(a|c) \quad \forall a, b, c \in \mathbb{R}$

のうち，対称性と三角不等式を満たさない．

本論文では Itakura–Saito ダイバージェンスに基づく NMF（Itakura–Saito-divergence-based NMF: ISNMF）[20] について述べる．$\boldsymbol{Z}$ の要素である複素スペクトル z_{ij} が以下の確率モデルに従って生成されていると仮定する．

$$z_{ij} = \sum_k c_{ij,k} \tag{2.19}$$

$$c_{ij,k} \sim \mathcal{N}_{\mathbb{C}}(0, t_{ik}v_{lj}) \tag{2.20}$$

ここで，$c_{ij,k} \in \mathbb{C}$ は全ての i，j 及び k に関して互いに独立と仮定する．また，c を複素数の確率変数としたとき，$\mathcal{N}_{\mathbb{C}}(\mu, \sigma^2)$ は一次元複素ガウス分布を表し，その確率密度関数は次式で

与えられる．

$$p_c(c|\mu,\sigma^2) = \frac{1}{\pi\sigma^2}\exp\left\{-\frac{|c-\mu|^2}{\sigma^2}\right\} \tag{2.21}$$

ここで，μ 及び σ^2 はそれぞれ平均及び分散を示す．また，c_1 と c_2 が独立である場合，ゼロ平均の一次元複素ガウス分布において以下の加法性が成り立つ．

$$c_1 \sim \mathcal{N}_{\mathbb{C}}(0,\sigma_1^2) \text{ かつ } c_2 \sim \mathcal{N}_{\mathbb{C}}(0,\sigma_2^2) \implies c_1 + c_2 \sim \mathcal{N}_{\mathbb{C}}(0,\sigma_1^2+\sigma_2^2) \tag{2.22}$$

よって，

$$\sum_k c_{ij,k} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k t_{ik}v_{kj}\right) \tag{2.23}$$

を用いて次式が成り立つ．

$$z_{ij} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k t_{ik}v_{kj}\right) \tag{2.24}$$

これが ISNMF の生成モデルで，Fig. 2.7 のような球対称複素ガウス分布である．ここで，観測信号 z_{ij} が与えられた場合における t_{ik} 及び v_{kj} の最尤推定問題を考える．このとき，尤度関数は

$$\mathcal{L}(\boldsymbol{T},\boldsymbol{V}) = \prod_{i,j}\frac{1}{\pi\sum_k t_{ik}v_{kj}}\exp\left(-\frac{|z_{ij}|^2}{\sum_k t_{ik}v_{kj}}\right) \tag{2.25}$$

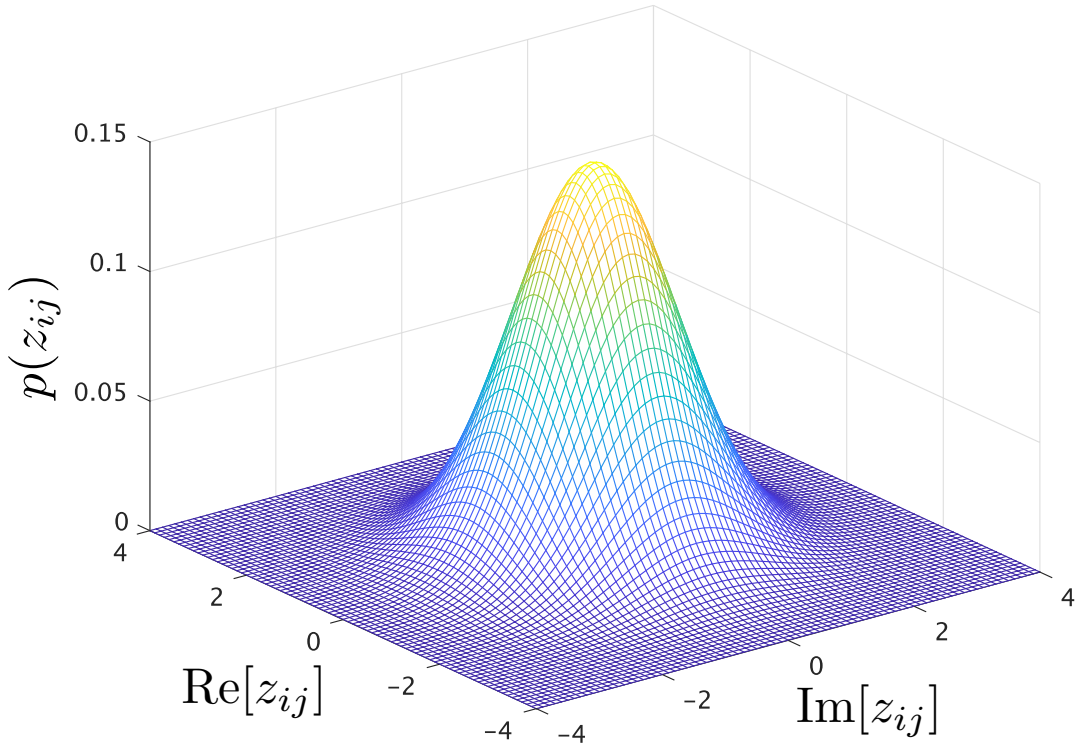

Fig. 2.7. Circularly symmetric complex Gauss distribution.

となり，負対数尤度は

$$-\log \mathcal{L}(\boldsymbol{T}, \boldsymbol{V}) = \sum_{i,j} \left(\frac{|z_{ij}|^2}{\sum_k t_{ik} v_{kj}} + \log \sum_k t_{ik} v_{kj} + \log \pi \right) \tag{2.26}$$

で表される．これは観測信号のパワースペクトログラム $|z_{ij}|^2$ に対する ISNMF の目的関数 (式（2.18) の $\beta = 0$ の場合）と定数部分を除いて一致するので，尤度関数は以下のように書き換えられる．

$$-\log \mathcal{L}(\boldsymbol{T}, \boldsymbol{V}) = d_{\mathrm{IS}} \left(|z_{ij}|^2 | \sum_k t_{ik} v_{kj} \right) + \text{const.} \tag{2.27}$$

ここで，$d_{\mathrm{IS}}(\cdot|\cdot)$ は 2 つの行列間の Itakura-Saito ダイバージェンスを示す．つまり，ISNMF を観測信号のパワースペクトログラム $|\boldsymbol{Z}|^{\cdot 2}$ に適用したとき，複素スペクトル z_{ij} が式 (2.24) で表される生成モデルに従い，全時間周波数グリッドに関して互いに独立であると仮定されている．また，ISNMF の $\boldsymbol{T}$ 及び $\boldsymbol{V}$ の最適化のための反復更新式は式 (2.28) 及び式 (2.29) で表される [21].

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_j |z_{ij}|^2 v_{kj} \left(\sum_{k'} t_{ik'} v_{k'j} \right)^{-2}}{\sum_j v_{kj} \left(\sum_{k'} t_{ik'} v_{k'j} \right)^{-1}}} \tag{2.28}$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_i |z_{ij}|^2 t_{ik} \left(\sum_{k'} t_{ik'} v_{k'j} \right)^{-2}}{\sum_i t_{ik} \left(\sum_{k'} t_{ik'} v_{k'j} \right)^{-1}}} \tag{2.29}$$

この更新式は乗算型反復更新式と呼ばれ，目的関数が単調非増加であることが保証されている．

2.7 ILRMA

ILRMA [10, 11] は IVA の音源モデルを NMF に拡張した手法である．IVA は同一音源の周波数成分の共起性からパーミュテーション問題を回避しつつ，分離行列 $\boldsymbol{W}_i$ を推定できる．しかし，全周波数に対して一様な強度の変化しか表せない．つまり，楽器信号のような基本周波数に対して倍音のみが大きい成分を持ち，それ以外の周波数では小さい成分を持つ，という明確な周波数構造を持つ信号の分離には適していない．それに対して，2.6 節で解説した NMF は，同一音源の時間周波数構造を少数（K 個）のスペクトルパタンとそのアクティベーションで表すことができ，これは音声や楽器信号の時間周波数構造のモデル化に良く適合している．ILRMA の反復最適化の概要を Fig. 2.8 に示す．ここで $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ は n 番目の音源のパワースペクトログラムをモデル化する基底行列及びアクティベーション行列を示す．ILRMA は IVA に対応する分離行列 $\boldsymbol{W}_i$ の反復最適化と ISNMF の低ランクモデリングに対応する $\boldsymbol{T}_n \boldsymbol{V}_n$ の反復最適化が交互に行われる．具体的には，分離行列 $\boldsymbol{W}_i$ により推定された分離信号を NMF によって非負低ランク行列でモデル化し，得られた $\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の各時間周波

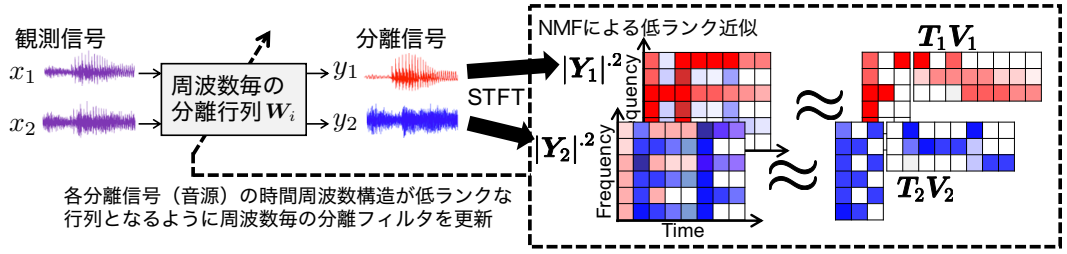

Fig. 2.8. Outline of ILRMA.

数成分を式 (2.24) における分散（各音源の生成モデルの推定パラメタ）として用いて分離行列を再度推定する，というプロセスが反復的に行われる．ILRMA の生成モデルは ISNMF と同様に次式の複素ガウス分布で表される．

$$y_{ij,n} = \sum_k c_{ij,k,n} \tag{2.30}$$

$$c_{ij,k,n} = \mathcal{N}_{\mathbb{C}}(0, t_{ik,n} v_{kj,n}) \tag{2.31}$$

ここで，$t_{ik,n}$ 及び $v_{kj,n}$ は n 番目の音源に関する基底行列 $\boldsymbol{T}_n$ 及びアクティベーション行列 $\boldsymbol{V}_n$ の非負要素であり，$k = 1, 2, \cdots, K$ は基底インデクスである．また，$c_{ij,k,n} \in \mathbb{C}$ は互いに独立であると仮定する．このとき，観測 $x_{ij,n}$ が与えられた場合において $\boldsymbol{W}_i$，$\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ を最尤推定する問題を考える．ISNMF のときと同様に

$$\sum_k c_{ij,k,n} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k t_{ik,n} v_{kj,n}\right) \tag{2.32}$$

より，

$$y_{ij,n} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k t_{ik,n} v_{kj,n}\right) \tag{2.33}$$

が成り立つので，この生成モデルに基づく観測信号の負対数尤度は次式で表される．

$$\mathcal{L}(\mathsf{W}, \mathsf{T}, \mathsf{V}) = -2J \sum_i \log |\det \boldsymbol{W}_i| + \sum_{i,j,n} \left(\frac{|\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}|^2}{\sum_k t_{ik,n} v_{kj,n}} + \log \sum_k t_{ik,n} v_{kj,n} \right) \tag{2.34}$$

ここで，$\mathsf{W} = \{\boldsymbol{W}_i\}_{i=1}^I$，$\mathsf{T} = \{\boldsymbol{T}_n\}_{n=1}^N$ 及び $\mathsf{V} = \{\boldsymbol{V}_n\}_{n=1}^N$ は最適化パラメタの集合である．式 (2.34) を見ると，第一項と第二項は式 (2.15) で表される IVA の尤度関数に対応し，第二項と第三項は式 (2.27) の ISNMF の尤度関数に対応していることがわかる．

分離行列 $\boldsymbol{W}_i$ の関する最適化は，分離ベクトル $\boldsymbol{w}_{i,n}$ の更新を反復射影法（iterative projection: IP）[14] を用いることで次式で行われる．

$$\boldsymbol{U}_{i,n} = \frac{1}{J} \sum_j \frac{1}{\sum_l t_{ik,n} v_{kj,n}} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{H}} \tag{2.35}$$

$$\boldsymbol{w}_{i,n} \leftarrow (\boldsymbol{W}_i \boldsymbol{U}_{i,n})^{-1} \boldsymbol{e}_n \tag{2.36}$$

$$\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n} (\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{U}_{i,n} \boldsymbol{w}_{i,n})^{-\frac{1}{2}} \tag{2.37}$$

ここで，$\boldsymbol{e}_n \in \mathbb{R}^N_{\{0,1\}}$ は n 番目の要素が 1，他要素が 0 のベクトルである．

NMF による低ランクモデリングのパラメタ $\boldsymbol{T}_n\boldsymbol{V}_n$ の最適化は式 (2.28) 及び式 (2.29) の乗算型反復更新式を用いて次式で表される．

$$t_{ik,n} \leftarrow t_{ik,n}\sqrt{\frac{\sum_j |\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2 v_{kj,n}(\sum_{k'} t_{ik',n}v_{k'j,n})^{-2}}{\sum_j v_{kj,n}(\sum_{k'} t_{ik',n}v_{k'j,n})^{-1}}} \tag{2.38}$$

$$v_{kj,n} \leftarrow v_{kj,n}\sqrt{\frac{\sum_i |\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2 t_{ik,n}(\sum_{k'} t_{ik',n}v_{k'j,n})^{-2}}{\sum_i t_{ik,n}(\sum_{k'} t_{ik',n}v_{k'j,n})^{-1}}} \tag{2.39}$$

これらの更新式も式 (2.28) 及び式 (2.29) と同様に目的関数式 (2.34) の値が単調非増加であることが保証されている．

2.8　更新式を一般化させた独立低ランク行列分析

ILRMA の反復更新式は，前節で述べたように分離行列（以後，空間モデルと呼ぶ）$\boldsymbol{W}_i$ は式 (2.35) から式 (2.37) で，低ランクモデリングによるパラメタ $\boldsymbol{T}_n\boldsymbol{V}_n$（以後，音源モデルと呼ぶ）は式 (2.38) 及び式 (2.39) で交互に反復最適化される．これらの空間モデルと音源モデルの収束速度は大幅に異なることが経験的に知られている．それに加えて ILRMA の目的関数式 (2.34) は非凸関数である．これらの理由から，$\boldsymbol{W}_i$，$\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ に与える乱数初期値に対して，空間モデルと音源モデルの最適化速度のバランスが悪い場合，音源分離が正しく達成されない局所最適解に陥ることがある．そこで，文献 [23] では補助関数法の一つである majorization-equalization 法 [24] を音源モデルの更新式に用いることで最適化速度を制御できる，音源モデルの一般化された更新式が提案されている．式 (2.40) 及び式 (2.41) に音源モデルの一般化反復更新式を示す．

$$t_{ik,n} \leftarrow t_{ik,n}\left[\frac{\sum_j |\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2 v_{kj,n}(\sum_{k'} t_{ik',n}v_{k'j,n})^{-2}}{\sum_j v_{kj,n}(\sum_{k'} t_{ik',n}v_{k'j,n})^{-1}}\right]^p \tag{2.40}$$

$$v_{kj,n} \leftarrow v_{kj,n}\left[\frac{\sum_i |\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2 t_{ik,n}(\sum_{k'} t_{ik',n}v_{k'j,n})^{-2}}{\sum_i t_{ik,n}(\sum_{k'} t_{ik',n}v_{k'j,n})^{-1}}\right]^p \tag{2.41}$$

ここで，p は最適化速度を表すパラメタで，$0 < p \leq 1$ の範囲で単調非増加であることが保証されている．また，$p = 0.5$ とき，従来の更新式である式 (2.38) 及び式 (2.39) に一致する．

2.9　インタラクティブ音源分離システム

前節で述べたように，ILRMA は最適化の過程で局所最適化に陥ることがある．これはブロックパーミュテーションという現象が発生するためである．Fig. 2.9 にブロックパーミュテーション問題の例を示す．ブロックパーミュテーション問題とは，2.3 節で述べたようなパーミュテーション問題がまとまった周波数帯で発生することを言う．つまり，ある周波数帯では特定の音源の分離音であるにも関わらず，その他の周波数帯では別の音源の分離音が出力

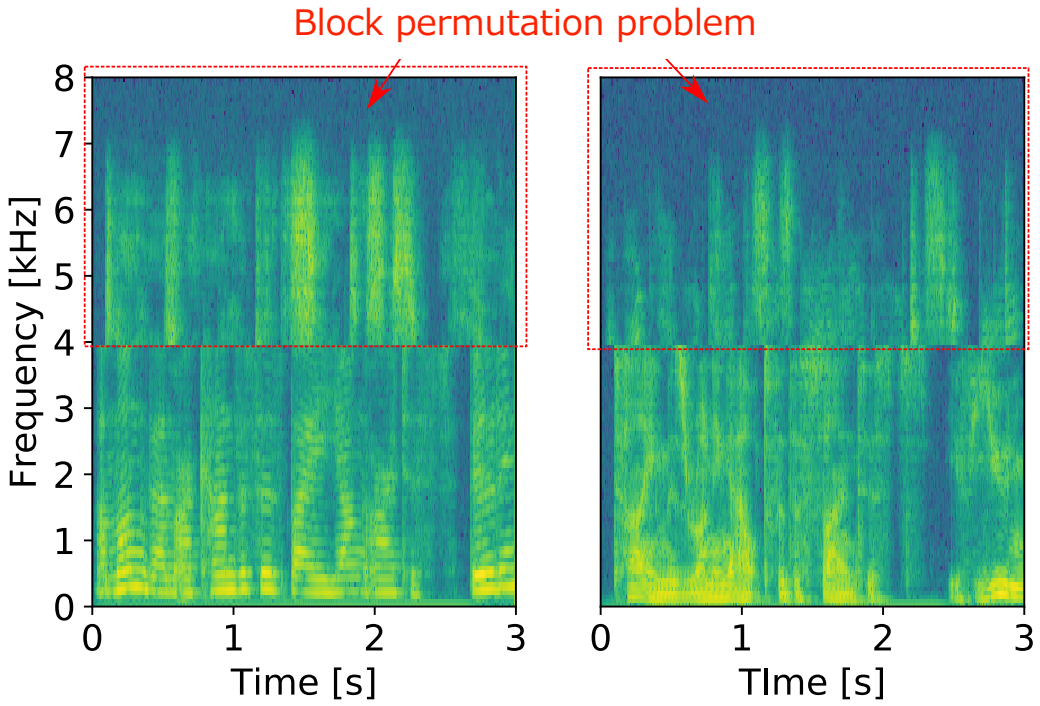

Fig. 2.9. Example of block permutation problem.

されている，という状態になっている．この現象は特に音声信号の分離の際に頻発する．そこで文献 [25] では，ユーザに分離音のスペクトログラムを提示し，ブロックパーミュテーション問題が発生していれば，その周波数帯をアノテーションとしてアルゴリズムに与えて再分離する，というインタラクティブ音源分離システムが提案されている．このシステムの概要を Fig. 2.10 に示す．本システムはシステムの柔軟性からサーバ・クライアントモデルのようにフロントエンドとバックエンドを分割する方法がとられている．Fig. 2.11 にシステムのユーザインターフェース（user interface: UI）を示す．このシステムはブロックパーミュテーションを直接する指定するだけでなく空間モデルの反復更新の際に重要な情報となる「ある特定の音源のみが沈黙している時間」を指定する方法でもアノテーションを与えることができる．

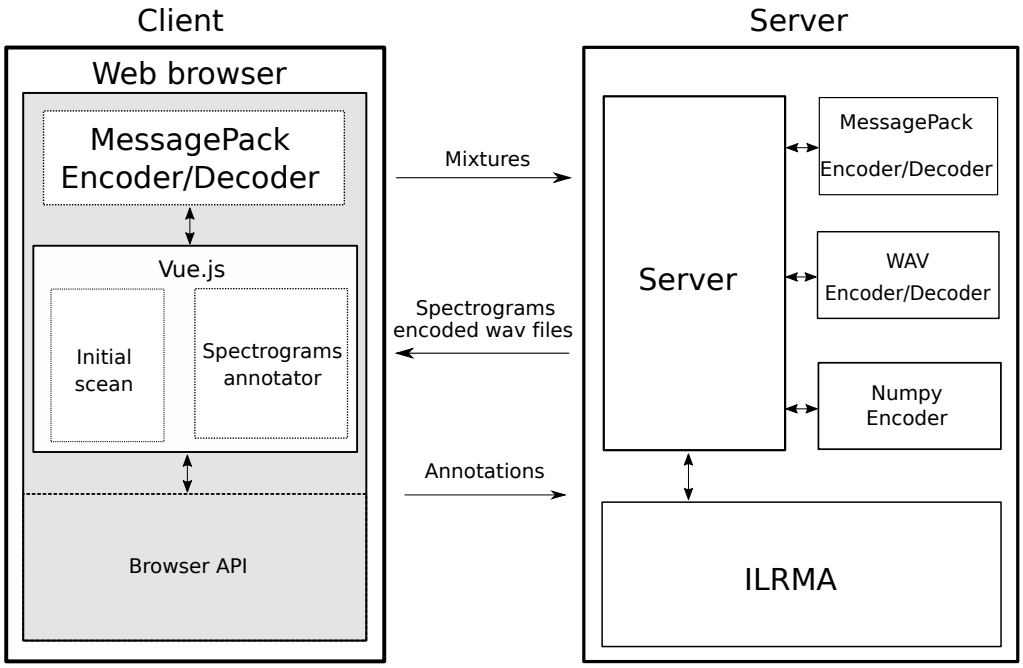

Fig. 2.10. Outline of interactive audio source separation system.

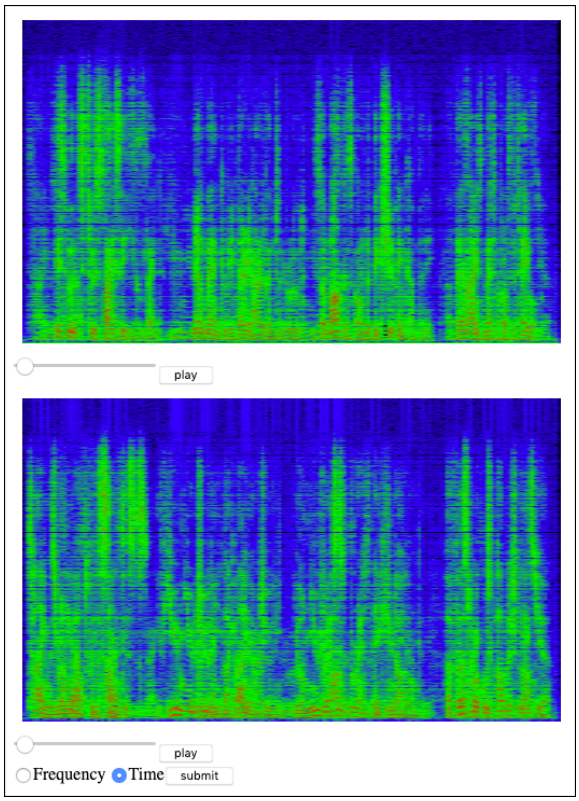

Fig. 2.11. UI of interactive audio source separation system.

2.10　アノテーションの活用方法

インタラクティブ音源分離システムで，ユーザがアノテーションを与えたときの処理方法について説明する．処理方法はブロックパーミュテーションを起こしている周波数帯を直接指定する方法と，多音源が沈黙し特定の音源のみが発音している時間を指定する方法が文献 [25] で提案されている．なお，後者の方法は二種類提案されている．

2.10.1　ブロックパーミュテーション発生周波数帯域の指定情報の活用

周波数アノテーションの活用方法の概要を Fig. 2.12 に示す．今，周波数ビン $i = i_s$ から $i = i_e$（$1 \leq i_s < i_e \leq I$）の範囲でブロックパーミュテーション問題が発生し，それに該当する音源インデクスが $n = n_s$，ブロックの移動先となる正しいパーミュテーションの音源が $n = n_t$ であるというアノテーション情報が，ユーザーから与えられた状況を考える．この場合，以後の ILRMA の更新では該当周波数ビンの分離フィルタ $\boldsymbol{w}_{i,n}$ と基底ベクトル $\boldsymbol{T}_n$ の成分 $t_{ik,n}$ について入れ替え処理を行えば良いので，次式のような処理を行う．

$$\boldsymbol{w}_{i_s,n_s},\ \boldsymbol{w}_{i_s+1,n_s},\cdots,\boldsymbol{w}_{i_e,n_s} \Leftrightarrow \boldsymbol{w}_{i_s,n_t},\boldsymbol{w}_{i_s+1,n_t},\cdots,\boldsymbol{w}_{i_e,n_t} \tag{2.42}$$

$$t_{i_sk,n_s},\ t_{(i_s+1)k,n_s},\cdots,t_{i_ek,n_s} \Leftrightarrow t_{i_sk,n_t},t_{(i_s+1)k,n_t},\cdots,t_{i_ek,n_t} \quad \forall k \tag{2.43}$$

$$v_{kj,n} \leftarrow \rho \quad \forall k,j,n \tag{2.44}$$

ここで $\Leftrightarrow$ は左辺と右辺の変数を入れ替えすることを意味する．また，ρ は区間 $(0,1)$ の一様乱数である．式 (2.44) では，アクティベーション行列 $\boldsymbol{V}_n$ をリセットすることで，現在捕らわれている悪い局所解から一度抜け出すことを目的としている．分離行列 $\boldsymbol{W}_i$ 及び基底行列 $\boldsymbol{T}_n$ は，式 (2.42) 及び (2.43) で入れ替えたうえでそのまま引き継いで ILRMA 最適化の反復更新

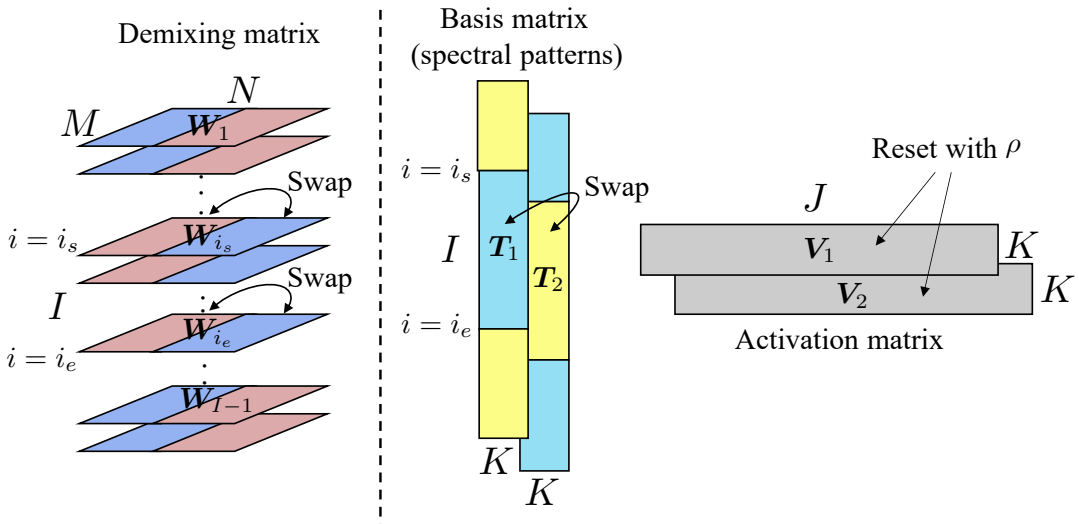

Fig. 2.12. Overview of frequency annotation.

を再開するため，ブロックパーミュテーション問題を回避しながらより高精度な音源分離ができる解へと誘導されることを期待している．

2.10.2　他音源が沈黙している時間範囲の指定情報の活用方法（a）

時間アノテーションの活用方法 (a) の概要を Fig. 2.13 に示す．今，時間 $j = j_s$ から $j = j_e$（$1 \leq j_s < j_e \leq J$）の範囲で音源 $n = n_t$ の音源が沈黙しているというアノテーション情報がユーザーから与えられた状況を考える．この場合，以後の ILRMA の更新では，次式のように該当時間フレームのアクティベーション成分 $v_{kj,n}$ に微小値で置き換え，同時に分離フィルタ $\boldsymbol{w}_{i,n}$ についても乱数でリセットする処理を行う．

$$v_{kj_s,n_t}, v_{k(j_s+1),n_t}, \cdots, v_{kj_e,n_t} \leftarrow \varepsilon, \varepsilon, \cdots, \varepsilon \quad \forall k \tag{2.45}$$

$$\boldsymbol{w}_{i,n} \leftarrow \rho \quad \forall i, n \tag{2.46}$$

ここで，$\varepsilon > 0$ は適当な微小値である．この処理では音源モデル $\boldsymbol{W}_i$ をリセットすることで，現在捕らわれている悪い局所解から一度抜け出すことを目的としている．この場合は基底行列 $\boldsymbol{T}_n$ とアクティベーション行列 $\boldsymbol{V}_n$ の一部は引き継いで ILRMA 最適化の反復更新を再開するため，より良い解へと誘導されることを期待している．

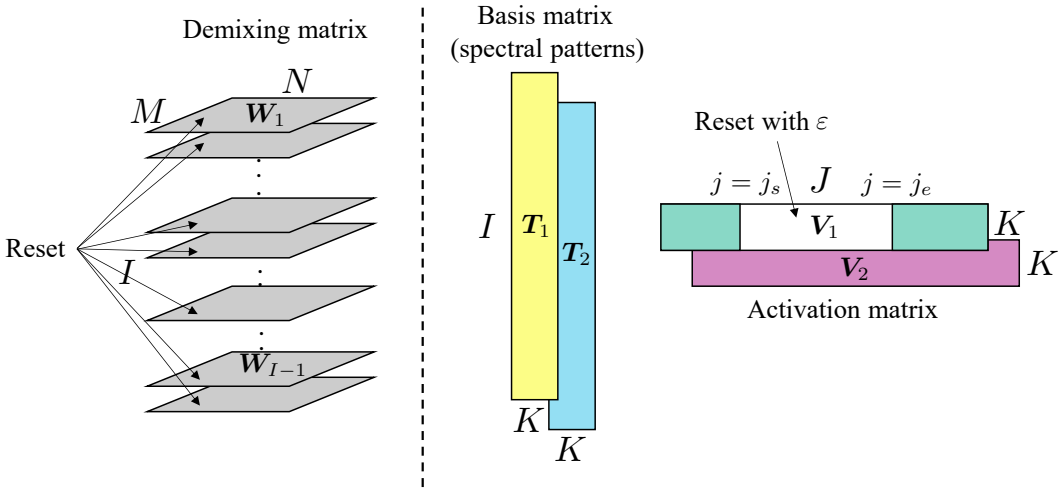

Fig. 2.13. Overview of time annotation (a).

2.10.3　他音源が沈黙している時間範囲の指定情報の活用方法（b）

時間アノテーションの活用方法（b）の概要を Fig. 2.14 に示す．前項の処理 (2.45) 及び (2.46) では，沈黙音源の当該時間フレームのアクティベーションに微小値を代入していたが，同時にアクティベーション行列のその他の要素を次式のようにリセットする手法も考えられている．

$$v_{k1,n_t}, v_{k2,n_t}, \cdots, v_{k(j_s-1),n_t} \leftarrow \alpha, \alpha, \cdots, \alpha \quad \forall k \tag{2.47}$$

$$v_{k(j_s+1),n_t}, v_{k(j_s+2),n_t}, \cdots, v_{kJ,n_t} \leftarrow \alpha, \alpha, \cdots, \alpha \quad \forall k \tag{2.48}$$

$$v_{kj,n} \leftarrow \alpha \quad \forall k, j, n \neq n_t \tag{2.49}$$

ここで，α は区間 $[1.0 \times 10^5, 1.1 \times 10^5]$ の一様乱数（ε と比較して十分大きな一様乱数）である．前項の処理 (2.45) 及び (2.46) のみの場合と比べ，本項の処理 (2.47)–(2.49) はより多くのパラメタをリセットしている．

2.11　本章のまとめ

本章では，ICA と FDICA の問題点について説明した．また，周波数領域 BSS の定式化を行い，その代表的な手法，そして ILRMA で導入された ISNMF について説明した．それに加えて 4 章で拡張するインタラクティブ音源分離システムとアノテーションの利用方法についても説明した．次章では，2.8 節で説明した一般化反復更新式の更新速度と分離性能の相関について調査を行う．

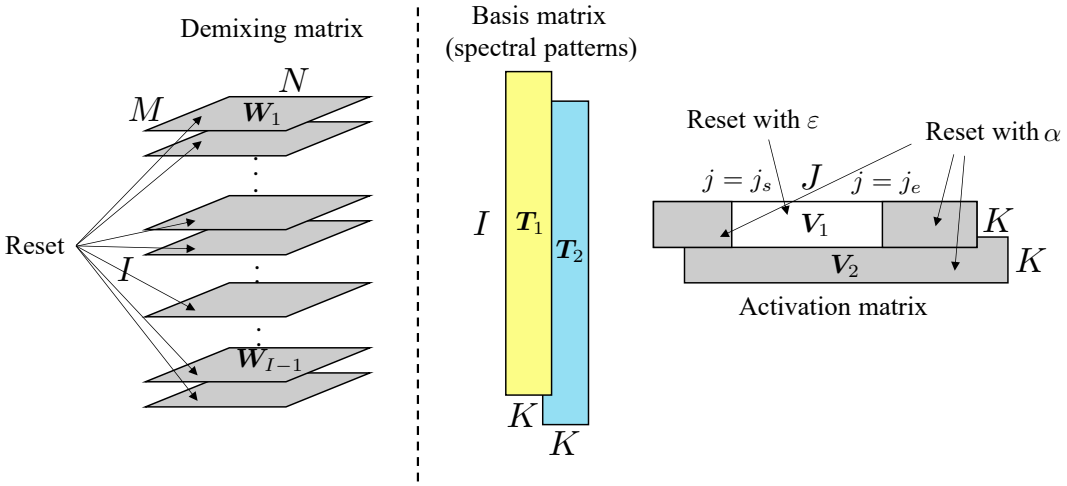

Fig. 2.14. Overview of time annotation (b).

第 3 章

目的関数の挙動と音源分離性能の相関の調査

3.1 まえがき

本章では，目的関数の挙動と音源分離性能の相関を調査する．まず，3.2 節では ILRMA の初期値依存性と分離失敗に対するアプローチについて述べる．次に，3.3 節 では最適化速度と収束値を別々に議論するために ILRMA の目的関数から IVA に由来する目的関数及び ISNMF に由来する目的関数を定義する．そして，3.4 節では最適化速度と分離性能の関係を明らかにし，それを元に分離性能が高いときと低いときの二つのモデルの目的関数を観測する．

3.2 ILRMA の初期値依存性

ILRMA のパラメタである空間モデル $\boldsymbol{W}_i$ 及び音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ の初期値は，それぞれ単位行列及び乱数で初期化することが一般的である．しかし，分離性能のばらつきはこの初期化方法でも解消されず，場合によっては極端に分離性能が低くなる局所最適解に陥ってしまう．この原因の多くは先述したブロックパーミュテーション問題である．これを解決するには，常に高い分離性能をもたらすようなパラメタの初期化方法を明らかにする必要があるが，現在その方法は不明である．また，ILRMA の目的関数が非凸関数なので，あらゆる観測信号に対して常に最良の分離性能をもたらすような初期化方法を確立することは困難であると考えられる．

そこで，別の観点から解決を検討する．空間モデル $\boldsymbol{W}_i$ 及び音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ の最適化速度のバランスを操作すれば，初期化方法が同じでも解は変わる．2.8 節で述べた音源モデルの一般化反復更新式を用いれば，空間モデル $\boldsymbol{W}_i$ 及び音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ の最適化速度を変化させることができる．そこで，本章では解決が困難である初期値依存性の問題を「空間モデル $\boldsymbol{W}_i$ 及び音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ の最適化速度の調整」という観点から解決の可能性があるか調査する．

3.3　モデルごとの目的関数

空間モデル $\boldsymbol{W}_i$ と音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ についての目的関数を式 (2.34) より以下のように定義する．

$$\mathcal{L}_{\mathsf{W}} = -2J\sum_i \log|\det \boldsymbol{W}_i| + \sum_{i,j,n}\left(\frac{|\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2}{\sum_k t_{ik,n}v_{kj,n}}\right) \tag{3.1}$$

$$\mathcal{L}_{\mathsf{TV}} = \sum_{i,j,n}\left(\frac{|\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2}{\sum_k t_{ik,n}v_{kj,n}} + \log\sum_k t_{ik,n}v_{kj,n}\right) \tag{3.2}$$

式 (3.1) 及び式 (3.2) は式 (2.34) のそれぞれ IVA に対応する項と ISNMF に対応する項を抽出したものとなっている．したがって，$\mathcal{L}_{\mathsf{W}}$ と $\mathcal{L}_{\mathsf{TV}}$ の反復に対する収束カーブを観測することで，空間モデル $\boldsymbol{W}_i$ と音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ の最適化速度と収束値を別々に議論でき，より良い最適化条件を調査することができる．

なお，式 (3.1) 及び式 (3.2) は式 (2.34) における第二項が重複しているため，両モデルの目的関数の和は式 (2.34) に一致しない．また，式 (2.34) と異なって式 (3.1) 及び式 (3.2) に単調非増加性は保証されていない点に注意する．

3.4　実験

まず，2.8 節で述べた一般化乗算型反復更新式の最適化速度と分離性能の関係について調査する．次に，分離性能が高いとき及び低いときの空間モデル $\boldsymbol{W}_i$ と空間モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ の目的関数の挙動を観察することで目的関数の挙動と分離性能の関係について調査する．分離性能の評価には次式で表される信号対歪比（source-to-distortion ratio: SDR）[26] を用いる．

$$\hat{S}[l] = e_t[l] + e_i[l] + e_a[l] \tag{3.3}$$

$$\mathrm{SDR} = 10\log\frac{\sum_{l=1}^{L}|e_t[l]|^2}{\sum_{l=1}^{L}|e_i[l]+e_a[l]|^2}\ [\mathrm{dB}] \tag{3.4}$$

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．

3.4.1 音源モデルの収束速度と分離性能に関する実験

分離実験は Fig. 3.1 に示すような 2 種類の音源を 2 本のマイクロホンで観測した場合を想定した．観測信号は SiSEC2011 [27] で利用されたギターとシンセサイザー混合信号（dev2__ultimate_nz_tour__snip_43_61）に対し，RWCP データベース収録の E2A インパルス応答 [28] を畳み込むことで作成した．その他の実験条件を Table 3.1 に示す．また，音源モデルの最適化速度を制御する指数パラメタ p を $(0, 1)$ の範囲の一様乱数から生成し，$\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の初期値と p を 2000 通り変化させた．その結果を Fig. 3.2 に示す．縦軸が SDR 改善量を，横軸が指数パラメタ p を示す．音源分離性能は 2 層の層状に分布しており，音源分離性能が高い結果と低い結果の 2 つの大まかな局所解に収束している様子が確認できる．また，指数パラメタ p が小さいとき，即ち音源モデルの最適化速度が速いよりもむしろ遅い方がより良い音源分離を達成できる傾向にある．

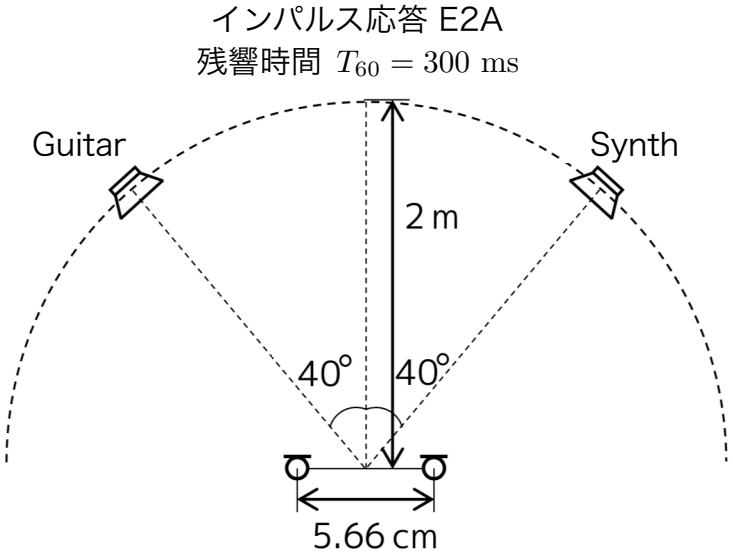

Fig. 3.1. Audio recording environment.

Table 3.1. Experimental conditions

Parameter	Value
Sampling frequency	16000 Hz
FFT length	256 ms (4096 samples)
Shift length	128 ms (2048 samples)
Initial value of spatial model	Uniform random values $(0, 1)$
Initial value of source model	$N \times N$ identity matrix
Number of iterations	200

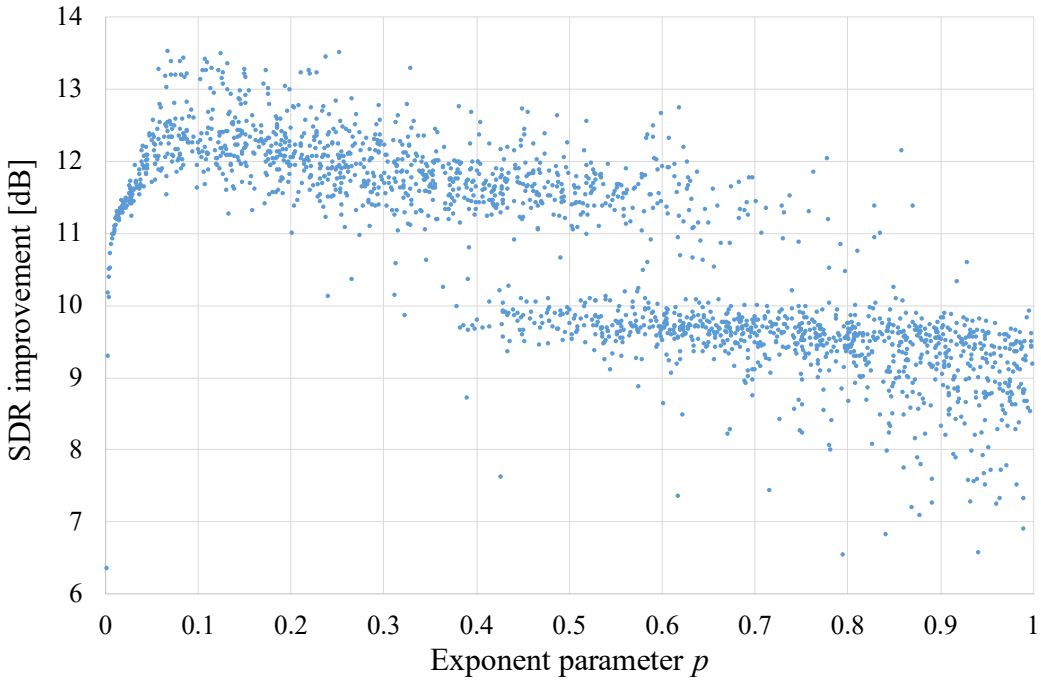

Fig. 3.2. SDR improvements of guitar and synth mixture in terms of balance of optimization speeds between spatial and source models.

3.4.2　目的関数の挙動に関する実験

前項の実験で得られた SDR 改善量が大きい場合及び小さい場合の例をそれぞれ 5 つ抽出したものを Table 3.2 に示す．これらのパラメタにおける ILRMA の反復最適化の際に式 (3.1) 及び式 (3.2) で得られる各モデルの目的関数の挙動を Fig. 3.3 及び Fig. 3.4 に示す．両図とも SDR 改善量が大きいときの目的関数の挙動を実線で，低いときを破線で表しており，横軸が反復回数を，縦軸が各モデルの目的関数値を示している．

まず，Fig. 3.3 の空間モデルの場合では分離性能が高いときの方が目的関数は小さい値に収束している．これは分離信号間の独立性をより高めるような空間モデルつまり分離行列を推定していることを示しており，この結果からも p の値を小さくする方がより良い音源分離結果をもたらす傾向にあることが示唆される．また，反復回数が 3 回程度の最適化がほとんど進んでいないとき，分離性能が低いときは高いときと比較してスパイクが大きい．

次に Fig. 3.4 では，音源モデルの目的関数の場合では分離性能が高いときの方が目的関数は大きい値に収束している．つまり，目的関数である負対数尤度関数を最小化することがより良い音源分離につながってないと言える．また，目的関数値のオーダから，音源モデルの目的関数が支配的であることから式 (2.34) の全体の目的関数でも同様の現象が見られる．一方で，目的関数の値ではなく挙動の方に注目すると，分離性能が高いときは低いときと比較してカーブが緩やかであることが見て取れる．

上記の結果がなぜより良い分離結果をもたらすかについて考察する．ILRMA の反復最適化

Table 3.2. Exponent parameters and initial values in case of good and poor separations

	SDR [dB]	指数パラメタ p	初期値（乱数シード）
高性能	13.51	0.067	1656
	13.50	0.252	789
	13.48	0.125	854
	13.43	0.237	984
	13.41	0.084	1580
低性能	6.53	0.795	769
	6.56	0.941	1052
	6.81	0.841	1040
	6.89	0.990	1513
	7.23	0.960	1823

は 2.7 節で述べたように，分離信号 $\boldsymbol{Y}_n$ の推定とそれに対する低ランクモデリングが交互に行われる．音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ の最適化速度が速いと $\boldsymbol{T}_n\boldsymbol{V}_n$ は直ちに $\boldsymbol{Y}_n$ をモデル化し収束するが，空間モデル $\boldsymbol{W}_i$ の最適化がまだ進んでいない反復初期の段階では，空間モデルによって推定される分離信号 $\boldsymbol{Y}_n$ の推定精度は低く，複数の音源が混合した（非目的音源が多く残留した）状態にある．このような状態で音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ が $\boldsymbol{Y}_n$ を正確にモデリングし収束してしまうと，混合信号を低ランク近似する不適切なモデリングが行われ，それに対する分離信号 $\boldsymbol{Y}_n$ もまた，分離性能が低い局所最適解に陥ってしまう．つまり，音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ は空間モデル $\boldsymbol{W}_i$ がある程度最適化された状態（ある程度音源分離が進んだ状態）でモデリングする必要がある．一般化反復最適化式における p 値を小さくし，音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ の最適化速度を遅くすることは，$\boldsymbol{T}_n\boldsymbol{V}_n$ が混合信号をモデル化してしまう現象を防ぐ効果があり，より音源分離性能の高い解へと誘導していると考えられる．

最後に，両モデルに共通する事実として，分離性能が高いときと低いときでは明らかに目的関数の挙動が異なっている．分離性能が低いときは空間モデルと音源モデルの両方の目的関数値の収束が速く，反復回数が 40 回を超えると目的関数値はほとんど変化しない．一方で，分離性能が高いときは反復回数が 100 回を超えても収束せず緩やかに目的関数値が推移している．この事実から，目的関数の挙動を観察することで分離が未了である反復最適化の途中であっても分離の成否を判断できることがわかる．

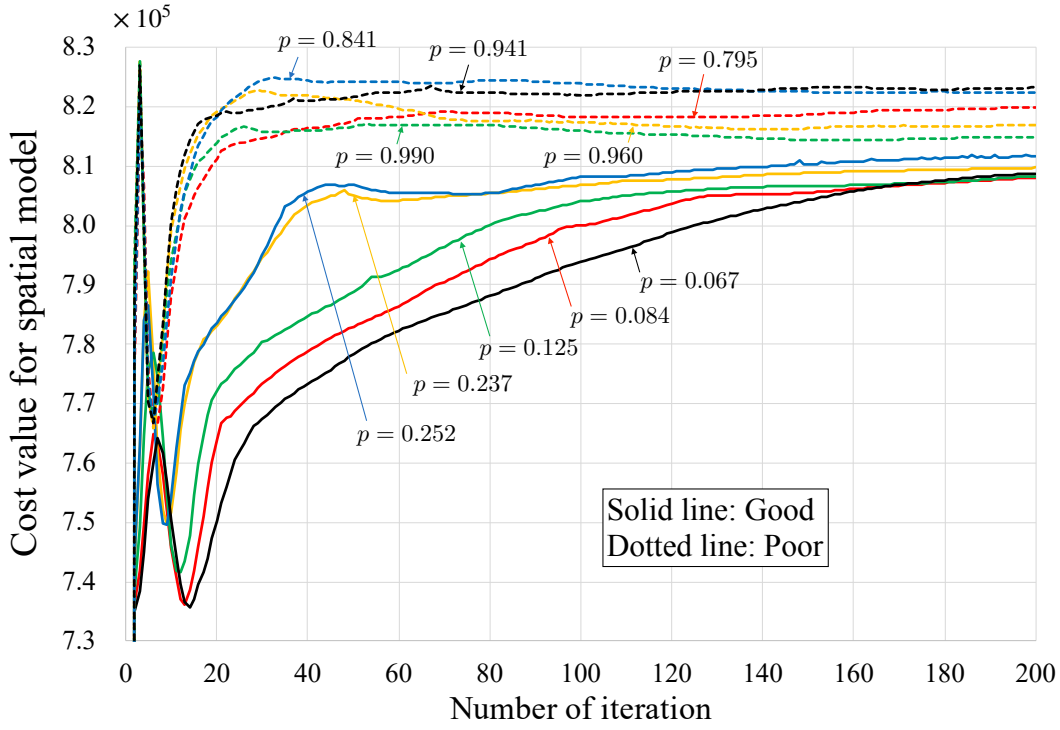

Fig. 3.3. Convergence behaviors of cost value for spatial model.

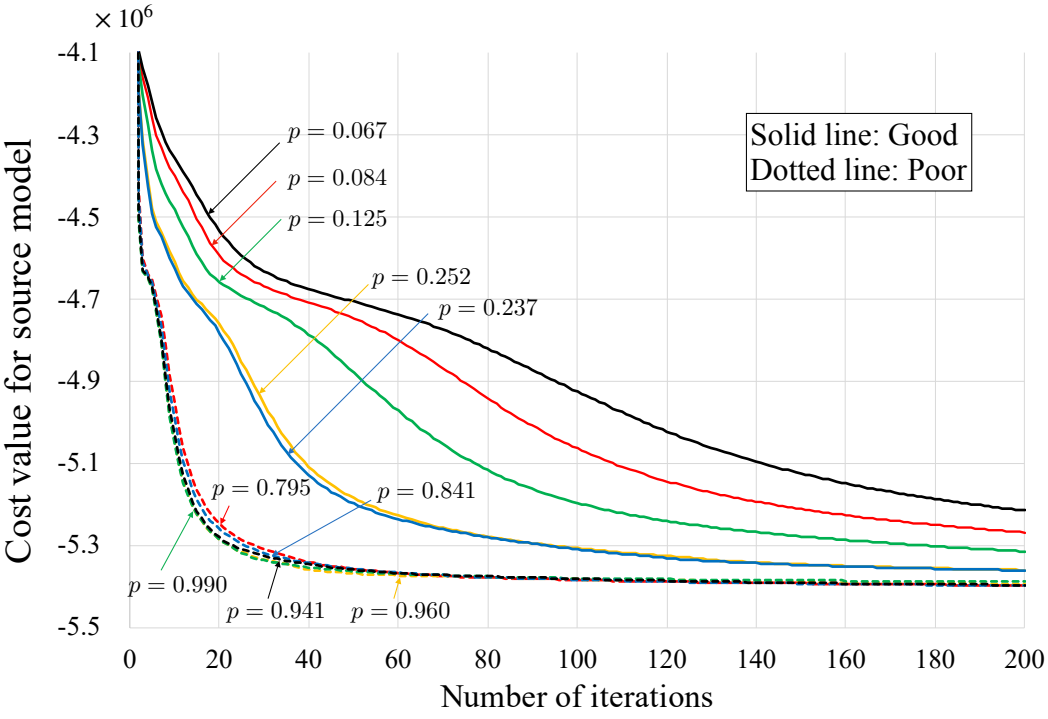

Fig. 3.4. Convergence behaviors of cost value for source model.

3.5 本章のまとめ

本章では初期値依存性により分離性能が低下する問題を回避するための実験を行った．まず，最適化速度を制御する指数パラメタ p と分離性能の関係を明らかにした．そして，分離性能が高いときと低いときの各モデルごとの目的関数の挙動を調べることで反復最適化の途中であっても分離の成否を判定できると言う知見を得た．次の 4 章ではこの知見を 2.9 節で述べたインタラクティブ音源分離システムに拡張する．

第 4 章

インタラクティブ音源分離システムへの拡張

4.1　まえがき

本章では 3 章で得られた知見を元に 2.9 節で述べたインタラクティブ音源分離システムに拡張し，初期値頑健性を持つ音源分離を達成できるシステムの開発について述べる．まず，4.2 節ではインタラクティブ音源分離システムの問題点と 3 章で得られた知見の活用方法について述べる．そして，4.3 節では開発したシステムの解説を行う．

4.2　動機

従来のインタラクティブ音源分離システムにおいて，ユーザはスペクトログラムと分離音から分離失敗を判定，アルゴリズムにアノテーションを与えることで音声信号を分離するときに発生しやすいブロックパーミュテーション問題を回避している．しかし，実用で発生するブロックパーミュテーションは Fig. 2.9 のように一見して判断することは，システムを初めて使うようなユーザにとっては幾分困難である．そしてユーザが誤ったアノテーションを与えた場合，正しく分離できている周波数帯に悪影響を与え，分離性能がさらに低くなることが予想される．このことから，システムを初めて使うユーザであってもある程度分離の失敗を指摘できるようなシステムの開発が必要であると言える．そこで，3 章で得た知見を元に，各モデルの目的関数の挙動をユーザに提示することで，音源分離の性能がパラメタの初期値に依存しない，より初期値頑健性の強いシステムの開発を目指す．

4.3　開発システムの解説

Fig. 4.1 に拡張した開発システムの一部を示す．既存システムである Fig. 2.11 と比較すると，ダウンロードする分離音を選択する部分と目的関数のグラフを追加した．なお，図には示していないが，ページ下部には空間モデルの目的関数だけでなく，音源モデルの目的関数及び

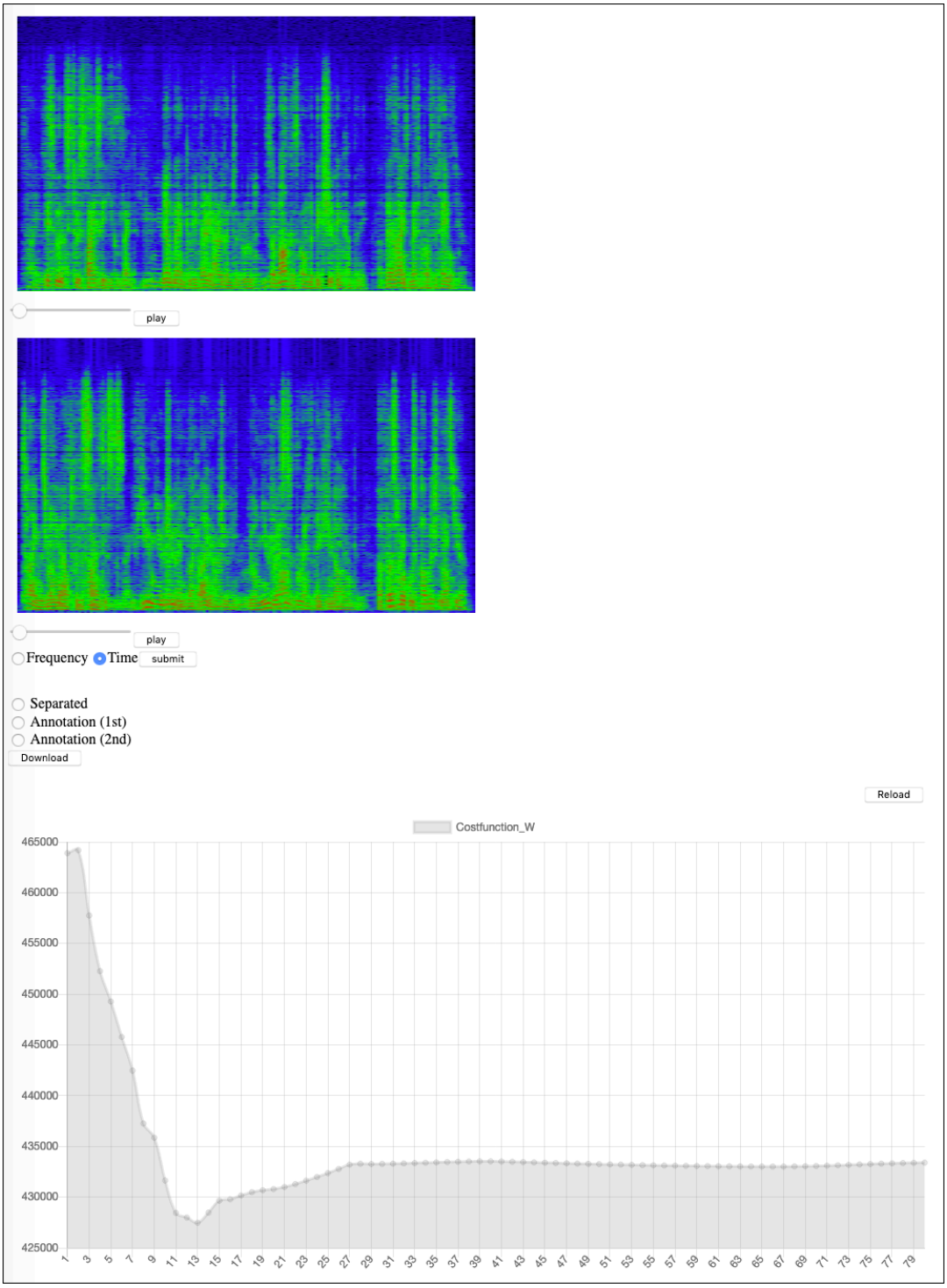

Fig. 4.1. UI of proposed interactive audio source separation system.

式 (2.34) で表される全体の目的関数を表示している．拡張したシステムでは分離がある程度進んだ状況（Fig. 4.1 では 80 回）でユーザにスペクトログラムと空間モデル，音源モデル及び全体の目的関数を示す．このとき，ユーザは分離音，スペクトログラム及び三つの目的関数

値の収束の挙動から分離の成否を判断する．以下に示す挙動が目的関数値のグラフで見られた場合，分離失敗の可能性が高い．

- ある程度分離が進んだときに，空間モデルの目的関数値が増加傾向にない
- 反復回数が 3 回程度のときの空間モデルの目的関数のスパイクが大きい
- 音源モデルの目的関数が 40 回程度の反復で収束し，以後ほとんど減少しない

もし分離失敗とユーザが判断した場合，ユーザは従来システムと同様にブロックパーミュテーションが発生している周波数帯を指定するか，特定音源のみが発音している時間を指定し再分離を行う．そして，再び分離の成否を判断する．本システムでは最初の分離音に加えて最大 2 回アノテーションを与えた 3 種類の音源を出力することができる．

拡張した部分の詳細を説明する．まず，Fig. 4.2 に UI の出力音源を選択する部分を示す．出力音源の選択は “Separated”，“Annotation(1st)” 及び “Annotation(2nd)” の三つのラジオボタンになっており，それぞれ最初の分離音，一度アノテーションを与えた再分離音及び二度アノテーションを与えた再分離音になっている．よってユーザは三つの分離音から実際に結果として出力する音源を選択することができる．

次に，Fig. 4.3 に UI の目的関数表示部を示す．表示する目的関数は上から空間モデル，音源モデル及び全体の目的関数となっている．アノテーションを与えて再分離が終了した後に空間モデルの目的関数のグラフ右上に配置してある “Reload” のボタンをクリックすることで目的関数の変更が反映される．

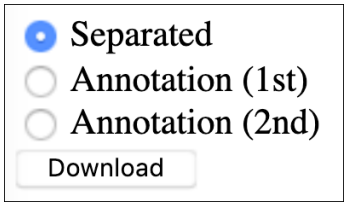

Fig. 4.2. Selection part of output source.

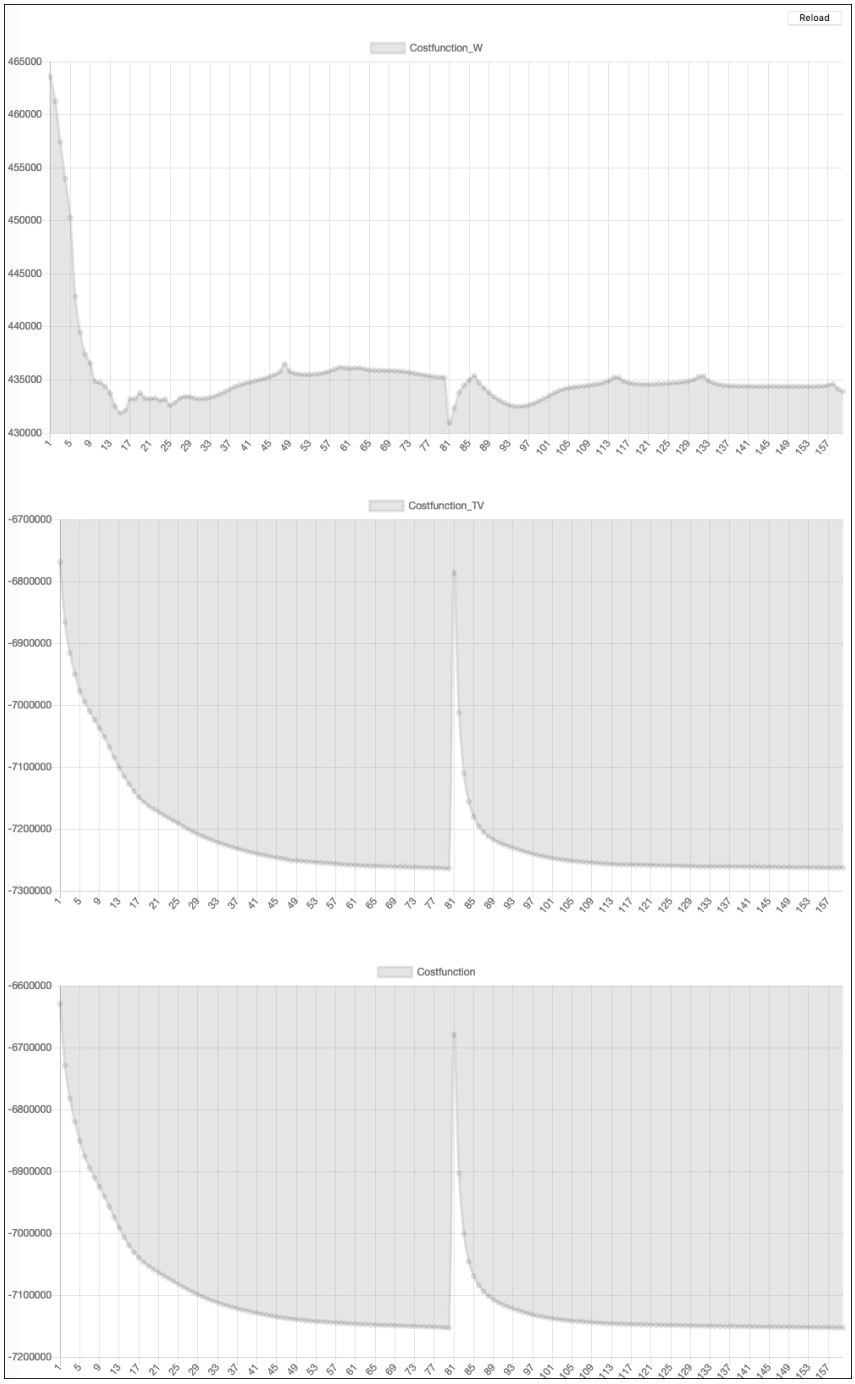

Fig. 4.3. Example of cost function behaviors.

4.4　実験

拡張したインタラクティブ音源分離システムの性能を調査するために，分離途中でユーザにスペクトログラムを提示し，アノテーションを与えた場合と与えずに分離を継続した場合の SDR 改善量を比較する．

4.4.1　実験条件

比較のための信号には SiSEC2011 の UND タスクの 6 信号を用いる．Table 4.1 に信号名を示す．また，実験条件を Table 4.2 に示す．実験では通常の ILRMA を 160 回反復した結果と，通常の ILRMA を 80 回反復したタイミングでユーザからのアノテーション情報を与えて残り 80 回反復した結果を比較する．

Table 4.1. Sources used in experiment

Mixture	Source signals
No. 1	dev1_female3_synthconv_130ms_5cm_sim_1
	dev1_female3_synthconv_130ms_5cm_sim_2
No. 2	dev1_male3_synthconv_130ms_5cm_sim_1
	dev1_male3_synthconv_130ms_5cm_sim_2
No. 3	dev1_male3_synthconv_130ms_5cm_sim_1
	dev1_female3_synthconv_130ms_5cm_sim_2

Table 4.2. Experimental conditions

Parameter	Value
Sampling frequency	16000 Hz
FFT length	128 ms (2048 samples)
Shift length	64 ms (1024 samples)
Number of bases K in source model	3
Small value ε	10^{-15}

4.4.2 実験結果

Fig. 4.4 から Fig. 4.6 はそれぞれ信号 nos. 1–3 について周波数の指定による SDR 改善量の推移を示したものである．黒の破線はアノテーションを与えずに 160 回反復したとき，赤の実線はブロックパーミュテーション問題が発生している周波数帯を直接指定したときの SDR 改善量をそれぞれ示す．SDR 改善量の推移を見ると，全ての音源においてアノテーションを与えなかった場合の ILRMA に対してアノテーションを与えた場合は，SDR 改善量が 2 dB 以上増加している．特に，Fig. 4.5 の no. 2 については SDR 改善量が約 10 dB の大きな増加が見られる．

Fig. 4.7 から Fig. 4.9 はそれぞれ信号 nos. 1–3 について沈黙時間の指定による SDR 改善量の推移を示したものである．黒の破線はアノテーションを与えずに 160 回反復したとき，赤の実線は 2.10.2 節で示した方法による時間アノテーション処理を行ったとき，青の実線は 2.10.3 節で示した方法による時間アノテーション処理をおこなったときをそれぞれ示す．(a) 及び (b) の両手法ともアノテーションを与えなかった場合と比較して SDR 改善量は大きくなっている．また，(a) と比較して (b) の方が SDR 改善量の増加量は大きい傾向にある．これは (b) の手法の方がより多くのパラメタをリセットしているためと考えられる．

これらの結果は，分離失敗している場合でも ILRMA に対してアノテーションを与えることで分離性能を向上させられることを示している．つまり，ILRMA の初期値依存性を回避し，音源分離性能が変数の乱数初期値に影響されない初期値頑健性を持っていると言える．このように，拡張したインタラクティブ音源分離システムでも従来のインタラクティブ音源分離システムと同様に，ILRMA の分離失敗に対してユーザがアノテーションを与えることで，初期値頑健性を持つ音源分離が実現できる．

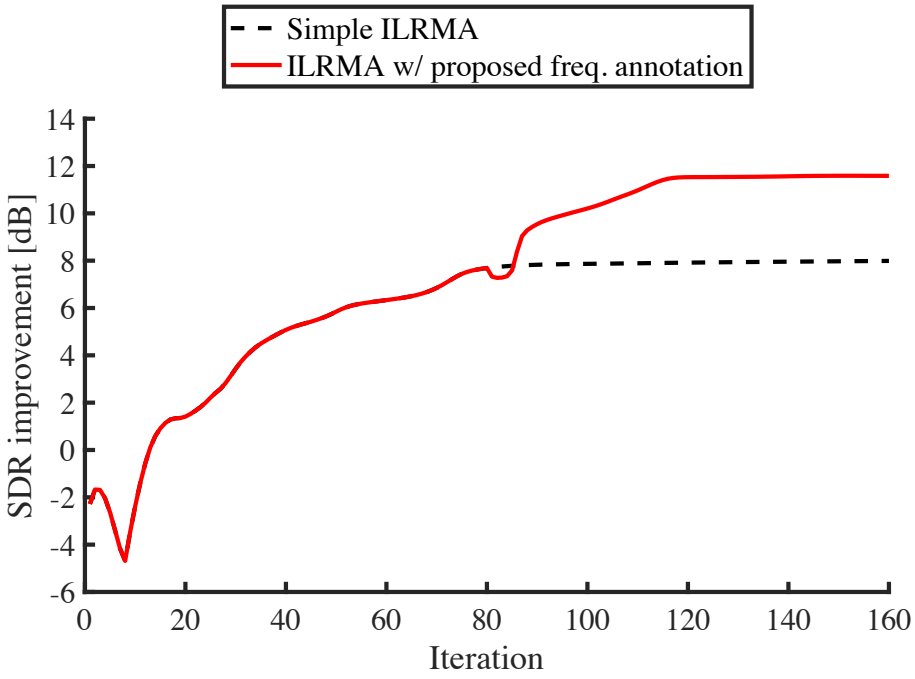

Fig. 4.4. SDR improvements by frequency annotation for no. 1 data.

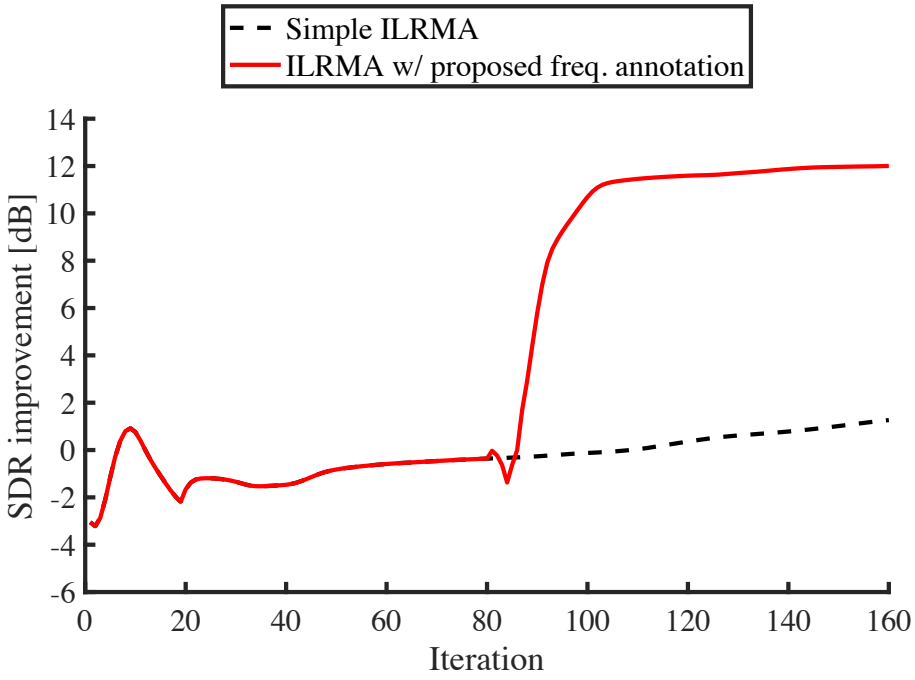

Fig. 4.5. SDR improvements by frequency annotation for no. 2 data.

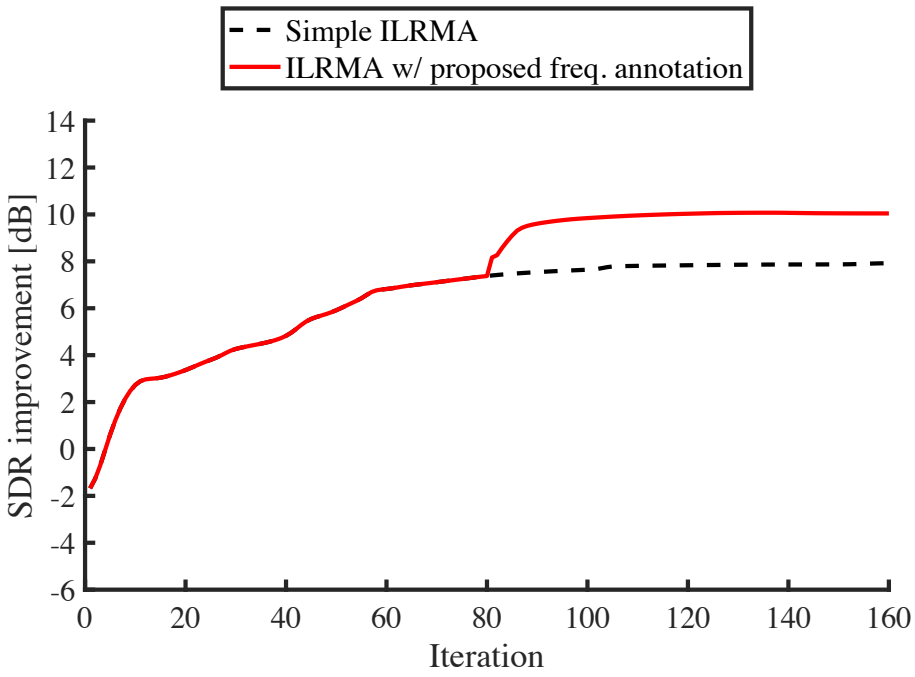

Fig. 4.6. SDR improvements by frequency annotation for no. 3 data.

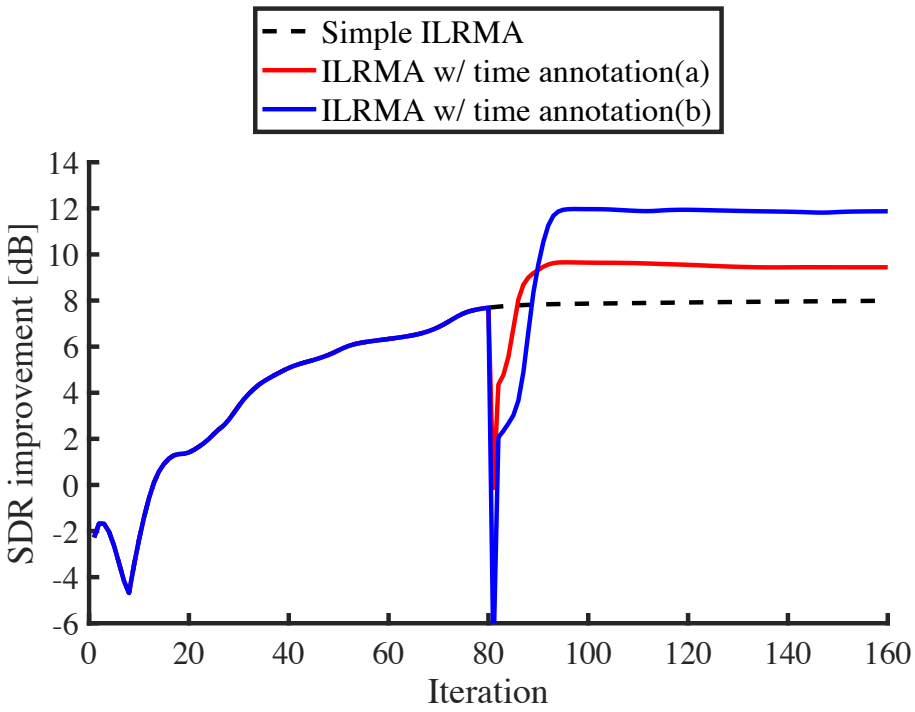

Fig. 4.7. SDR improvements by time annotation for no. 1 data.

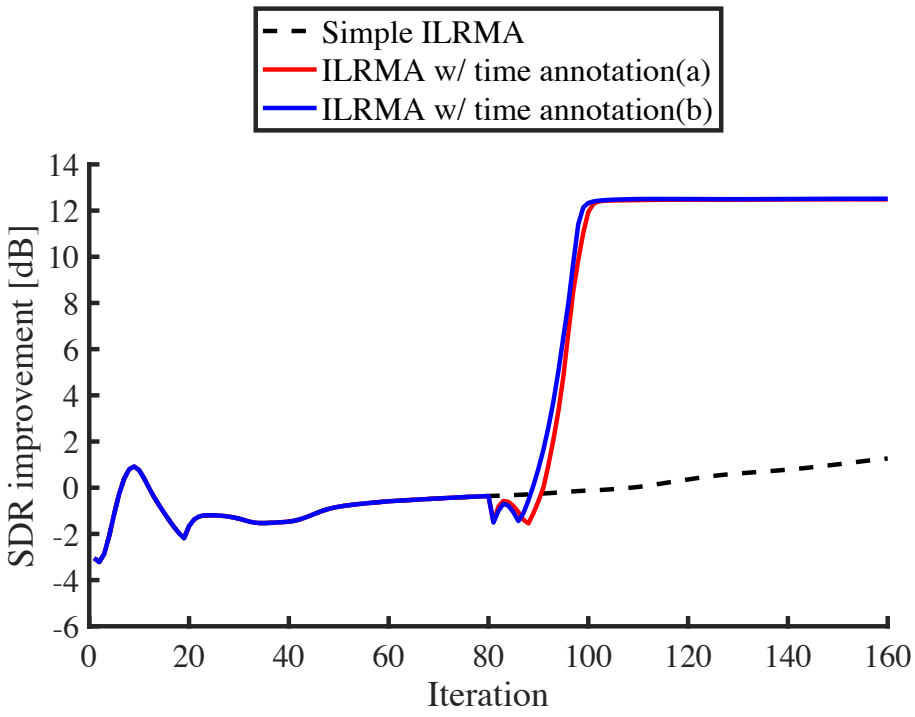

Fig. 4.8. SDR improvements by time annotation for no. 2 data.

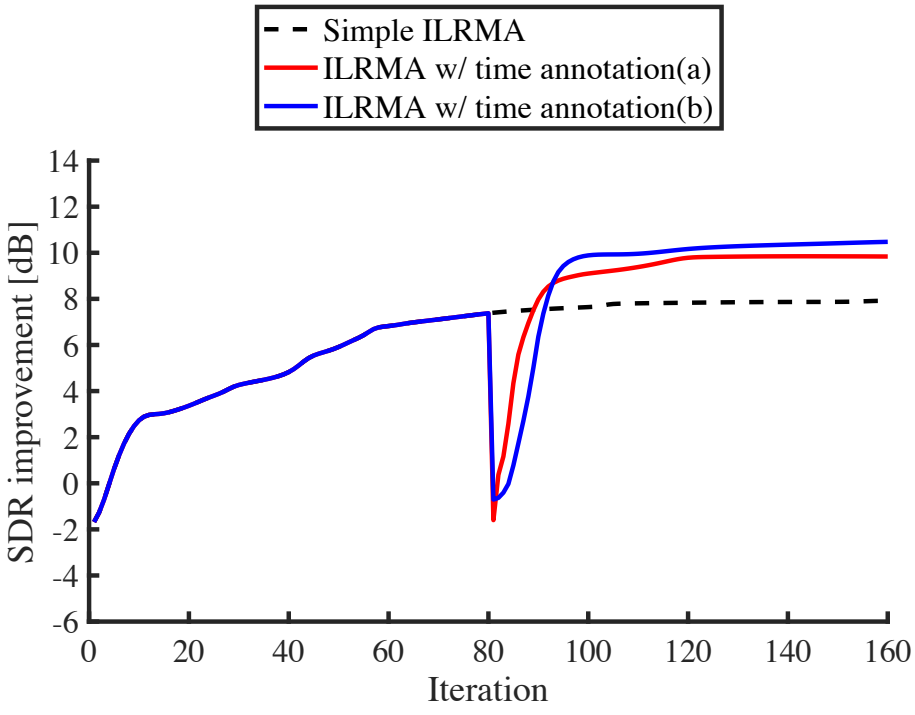

Fig. 4.9. SDR improvements by time annotation for no. 3 data.

4.5 本章のまとめ

本章では，前章で得られた知見による拡張を施した音源分離システムの開発について説明した．また，分離した音源を出力するインターフェースを実装し，実験を通して拡張したインタラクティブ音源分離システムが強い初期値頑健性を持っていることを示した．

第 5 章

結論

本論文では，頑健な音源分離が達成できる ILRMA のパラメタについて調査した．また，そこで得られた知見を元に音源分離システムを拡張した．

1 章では，音源分離技術の近年の動向と手法の概説を行なった．

2 章では，BSS の定式化，既存の ICA を発展させた BSS 手法の解説及びインタラクティブ音源分離システムについて述べた．

3 章では，分離性能及び最適化速度の相関を調査し，空間モデルと音源モデルの目的関数の挙動から最適化途中でも分離性能が予想できることを明らかにした．

4 章では，3 章で得られた知見をインタラクティブ音源分離システムに活用することで，より強い初期値頑健性を持つ音源分離が達成できるシステムを構築した．

最後に今後の課題を述べる．4 章で拡張したシステムはユーザが正しくアノテーションを与える新たな指標を導入したが，与えたアノテーションがアルゴリズムにとって有益であったかどうかの定量的な評価が必要であると考える．特にアノテーションを与えて再分離した後に分離性能が低下した場合の原因を明らかにするべきである．

謝辞

本論文は，香川高等専門学校電気情報工学科北村研究室にて行われた研究に基づくものです．

まず，本研究を進めるにあたり，ご多忙のところ熱心にご指導くださいました指導教員の北村大地助教に心より感謝申し上げます．また，専攻科入学時にに本研究室へ迎えていただいただき，音響信号処理分野の研究ができたことは自分の人生において大きな糧となりました．心よりありがたくお礼申し上げます．

本論の副査である重田和弘教授と柿元健准教授には，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，北村研究室同期の山地氏にはシステムの実装やプログラミングに際して多くの助言を頂き，後輩の岩瀬氏，大藪氏，渡辺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家族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] https://www.izotope.jp/products/rx-8/

[2] https://research.deezer.com/projects/spleeter.html

[3] https://github.com/deezer/spleeter

[4] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[5] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. International Conference on Independent Component Analysis and Signal Separation*, no. 1, 2007, pp. 414–421.

[6] D. Kitamura, H. Saruwatari, Y. Kosuke, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundamentals*, vol. 97-A, no. 5, pp. 1113–1118, 2014.

[7] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 18, no. 3, pp. 550–563, 2010.

[8] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal and Information Processing,* vol. 8, no. e12, pp. 1–14, 2019.

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

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEE/ACM Trans. Audio, Speech, and Language Processing*, 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," *Audio Source Separation*, S. Makino, Ed., pp. 125–155. Springer, Cham, 2018.

[12] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain,"

Neurocomputing, vol. 22, pp. 21–34, 1998.

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

[14] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," *Proc. International Conference on Latent Variable Analysis and Signal Separation,* pp.165–172, 2010.

[15] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proc. SICE Annual Conference*, pp. 722–727, 2001.

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

[17] S. Eguchi and Y. Kano, "Robustifing maximum likelihood estimation by psi- divergence," *ISM Research Memorandum*, 2001.

[18] A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He, "Extended SMART algorithms for non-negative matrix factorization," *Proc. International Conference on Artificial Intelligence and Soft Computing*, pp. 548–562, 2006.

[19] D. Kitamura, N. Ono, H. Saruwatari, Y. Takahashi, and K. Kondo, "Effective basis learning for sound source separation by semi-supervised nonnegative matrix factorization," *IEICE Technical Report*, EA2015–130, vol. 115, no. 521, pp. 355–360, 2016.

[20] C. Févotte, N. Bertin, J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis." *Neural Computation*, vol. 21, 793–830, 2009

[21] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence," *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, 2010, pp. 283–288.

[22] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician,* vol. 58, no. 1, pp. 30-–37, 2004.

[23] Y. Mitsui, D. Kitamura, N. Takamune, H. Saruwatari, Y. Saruwatari and K. Kondo, "Independent low-rank matrix analysis based on parametric majorization-equalization algorithm," *Proc. IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pp. 98–102, 2017.

[24] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[25] 大島風雅, 中野将生, 北村大地, "ユーザーからの補助情報を用いる独立低ランク行列分析" **日本音響学会** *2020* **年秋季研究発表会講演論文集**, pp. 269–272, 2020.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio

source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Q. K. Duong, "The 2010 signal separation evaluation campaign (SiSEC2010): Audio source separation," *Proc. International Conference on Latent Variable Analysis and Signal Separation*, 2010, pp. 114–122.

[28] 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. International Conference on Language Resources and Evaluation*, 2000, pp. 965–968.

修了予定者の研究実績一覧

創造工学専攻

著者名	論　文　名	発 表 機 関	発表年月
大島 風雅 （Fuga Oshima）	ユーザーからの補助情報を用いる独立低ランク行列分析	日本音響学会2020年秋季研究発表会講演論文集, 2-R1-6, pp. 269-272, 2020	2020.9