香川高専

卒業研究論文

論文題目

単一話者の発話区間率とブラインド音源分離性能の関係の調査

提出年月日	令和 6 年 2 月 6 日
学科	電気情報工学科
氏名	鈴木 慶 印
指導教員（主査）	北村 大地 講師 印
副査	重田 和弘 教授 印
学科長	漆原 史朗 教授 印

香川高等専門学校

Analysis of relationship between activity ratio of single speaker and blind source separation performance

Kei Suzuki

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

Abstract

Blind source separation (BSS) is a technique aimed at estimating individual audio signals from mixed observed signals. Various methods have been proposed for BSS, and one popular and important method is independent vector analysis (IVA). For multiple speech mixtures, IVA provides better separation performance when there exist many single-utterance segments in the mixture. In particular, it was theoretically proved that IVA can achieve the optimal separation performance in the case of flame-level W-disjoint orthogonality (F-WDO). F-WDO is an assumption that one source is dominant across all frequencies in each time frame of the mixture of all the source signals. However, it has not been analyzed whether such properties hold true for BSS methods other than IVA. In this thesis, I investigate the relationship between BSS performance of IVA and F-WDO. In addition, I conduct a similar BSS experiment to investigate whether another BSS algorithm called independent low-rank matrix analysis (ILRMA) has the same property as IVA. The experimental results demonstrate that BSS performance of IVA is improved when the rate of single-utterance segment in the mixture increases, resulting in approaching F-WDO. On the other hand, ILRMA did not show the same phenomenon, and it is hypothesized that this discrepancy may be attributed to the more complex optimization problem associated with ILRMA compared to IVA.

Keywords: single-utterance segment, blind source separation, independent component analysis, frame-level W-disjoint orthogonality

（和訳）

ブラインド音源分離（blind source separation: BSS）は，複数の音声信号が混合した観測信号から混合前の個々の音声信号を推定する技術であり，様々な手法が提案されている．ブラインド音源分離手法の 1 つである独立ベクトル分析（independent vector analysis: IVA）は，複数の話者がいる状況において，混合信号の全体時間区間に対して単一話者の発話時間区間の占める割合が多いほど分離性能が良いことが理論的に解析されている．特に，混合音源がフレーム単位相互排他直交性（flame-level W-disjoint orthogonality: F-WDO）を持つ状態であるとき IVA の分離性能が高くなることが IVA の目的関数から証明されている．F-WDO とは，混合音源の時間周波数表現において，ある短時間フレームでは 1 人の話者の音声信号が全周波数において支配的（排他的）であることをいう．しかし，このような性質が IVA 以外の BSS 手法のについても言えるか否かは解析されていない．そこで本論文ではまず，F-WDO の状態に近づくほど IVA の分離性能が向上することを実験的に追試する．その後，IVA の改良手法として知られている独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）についても同様の性質の有無について実験的に調査する．結果，IVA では先行研究の示唆する通り，混合信号が F-WDO に近い条件ほど分離性能が向上したが，ILRMA ではそのような現象が確認されなかった．この理由として，ILRMA の最適化問題が IVA のそれよりも複雑であることが考えられる．

目次

第 1 章

緒言

1.1 本論文の背景

音源分離とは，複数の音声信号を混合した観測信号から，混合される前の音声信号を推定する技術である．具体的な例を Fig. 1.1 に示す．Fig. 1.1 では，観測された音声信号を目的音声と背景雑音に分離し，背景雑音のみを除去することで目的音声を強調している．強調された音声はスマートスピーカーやナビゲーションシステムなど，音声認識の入力として利用される．また，音楽における利用例では Fig. 1.2 のように楽曲の混合信号から特定の楽器の演奏音を推定することができる．特定された楽器の演奏音はユーザによる既存音楽の再編集，自動採譜技術，及び実演奏補助技術などに利用される．

音源分離技術の一例としてブラインド音源分離（blind source separation: BSS）[1] があ

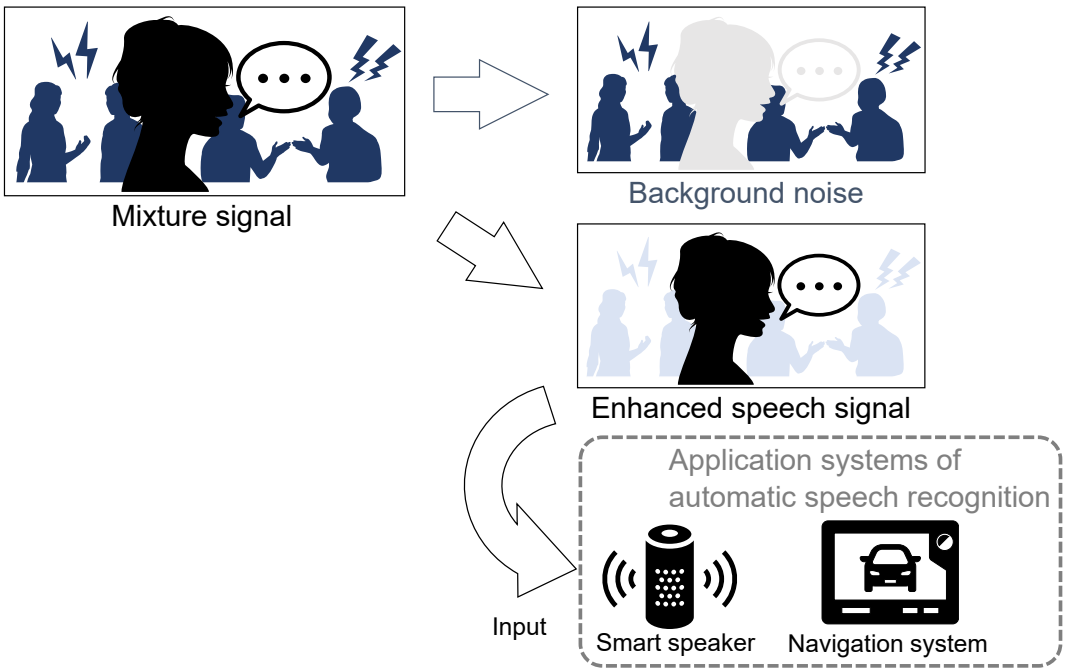

Fig. 1.1: Example of applications using speech source separation.

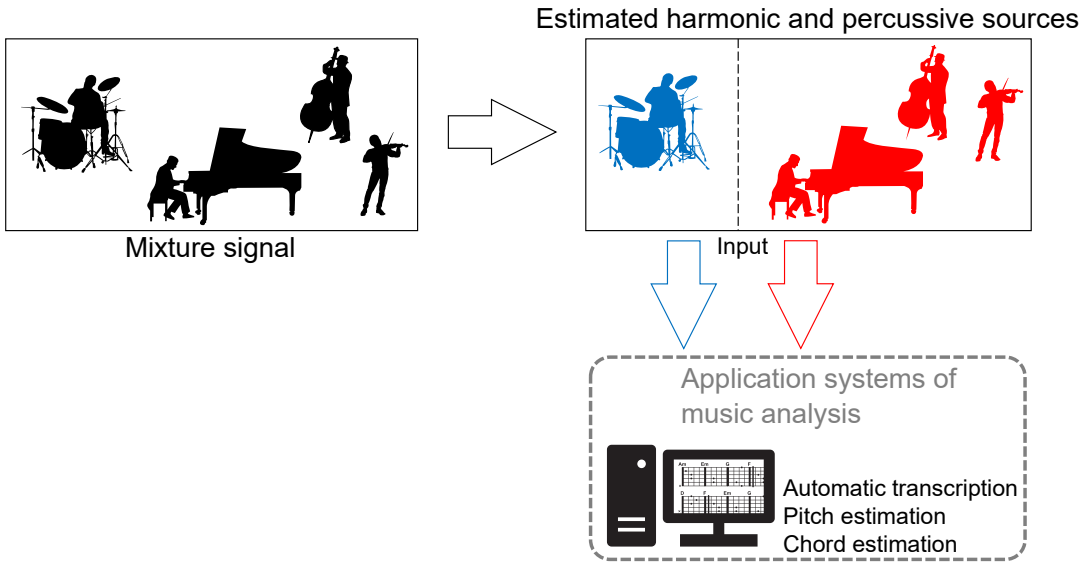

Fig. 1.2: Example of applications using musical source separation.

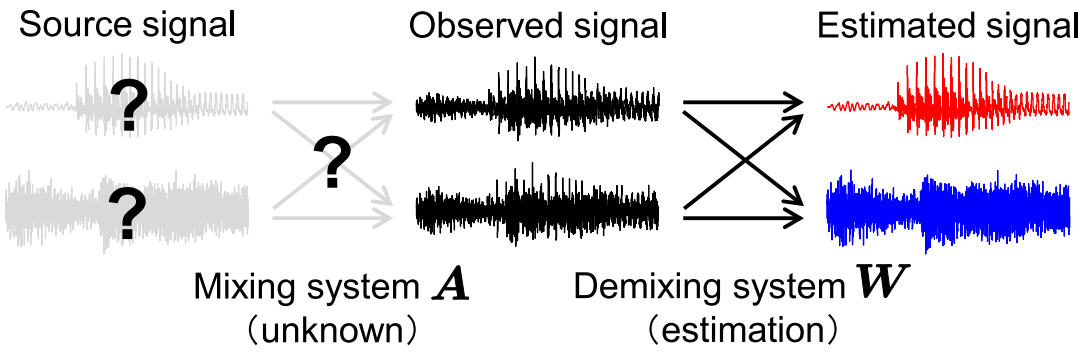

Fig. 1.3: Overview of BSS.

る．BSS とは，Fig. 1.3 に示すように，混合後の観測信号のみが与えられ，マイクロホンや音源の位置，音源の種類，周波数帯域，及びその学習データなどが与えられない（ブラインドな）条件で混合前の音声信号を推定する技術である．

現在に至るまで，音響信号処理の分野で発展してきた BSS の有名な手法として独立ベクトル分析（independent vector analysis: IVA）[2, 3] 及び，独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[4, 5] がある．これらのアルゴリズムはいずれも独立成分分析（independent component analysis: ICA）[6] を起源とする BSS である．ICA が電波や脳波等の時間領域の信号源の推定に使われたのに対し，音響信号では時間周波数領域での信号源（音源）の推定問題へと発展し，その後 IVA や ILRMA が登場したという背景がある．また近年では，IVA における高速かつ安定な更新アルゴリズム [7]，移動音源にも対応した IVA のオンライン更新アルゴリズム [8]，学習データと深層学習を ILRMA に援用するアルゴリズム [9]，時間周波数領域で満たされるべき性質を活用した ILRMA [10]，ユーザからのアノテーションを音源分離の補助情報として用いる ILRMA [11]，ILRMA が仮定する音源モ

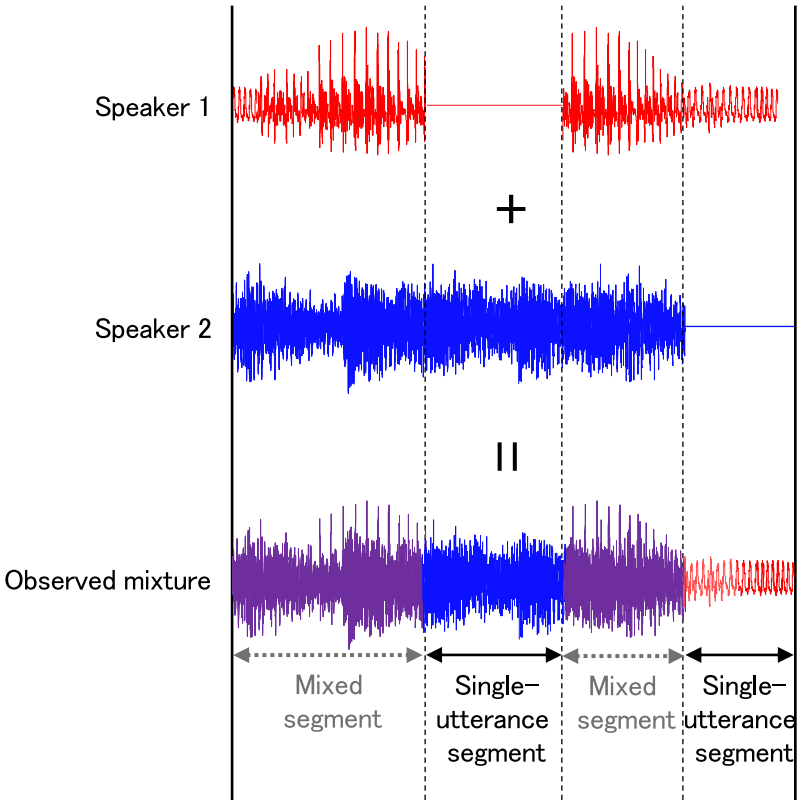

Fig. 1.4: Two speech sources and their mixture. The mixture signal contains mixed and single-utterance segments.

デルを一般化や拡張した手法 [12] 等が次々と提案されており，IVA 及び ILRMA が音響信号の BSS におけるメインストリームとなっている動向がうかがえる．本論文でも，複数の話者の音声が混合した信号に対する BSS として，以後 IVA 及び ILRMA を取り扱う．

近年の IVA に関する先行研究 [13] として，IVA は「混合信号の時間長に対して単一話者発話区間が占める割合が多いほど分離性能が高い」という性質を持つことが理論的に解析されている．単一話者発話区間とは，Fig. 1.4 に示すように混合音声の全体の時間区間のうち，2 人以上の話者が同時に発話しておらず，1 人の話者のみが発話している時間区間のことをいう．また，混合信号が単一話者発話区間のみで構成されていなくとも，単一話者発話区間が多いほど IVA の性能は向上する．よって IVA は混合信号の時間長に対して単一話者発話区間が占める割合が多いほど分離性能が高くなるといえる．

1.2 本論文の目的

本論文では 1.1 節で説明した，「混合信号の時間長に対して単一話者発話区間が占める割合が多いほど分離性能が高い」という IVA の特徴について実験的に追試する．具体的には，任意の割合で単一話者発話区間を含む観測信号を作成し，これを用いて単一話者発話区間の割

合と IVA の音源分離性能の関係性を実験的に示す．また，IVA の発展的な BSS 手法である ILRMA においても，同様に単一話者発話区間の割合と音源分離性能に相関があるかを実験的に調査する．ILRMA の最適化における目的関数は IVA の目的関数と類似する点が多いため，同様の特徴が観測されることが期待され，その知見は ILRMA による BSS の更なる性能向上につながる可能性が高いと考えられる．

1.3　本論文の構成

2 章では，音響信号の BSS における定式化，音源モデル，及び分離モデルについて説明する．また，本論文で扱う 2 種類の BSS（IVA 及び ILRMA）について，その概要を説明する．さらに，先行研究 [13] で示された IVA における性能向上の条件ついて説明し，実際の応用における活用方法について説明する．3 章では，単一話者発話区間の割合と BSS の性能の関係性について調査する実験のデータの作成方法，実験条件，及び実験結果を示す．実験方法では音源データの準備，録音環境，及び音源分離精度の客観評価尺度を説明する．実験結果では IVA 及び ILRMA について，混合音源中の単一話者発話区間の割合を変化させた場合の音源分離精度の変化を示す．また，IVA 及び ILRMA の実験結果を比較した考察を述べる．4 章では本論文のまとめ及び今後の課題を述べる．

第 2 章

基礎理論

2.1 まえがき

本章では，BSS に関する基礎理論及び単一話者発話区間と IVA の分離性能の関係性について説明する．2.2 節では，音響信号の BSS における定式化及び混合・分離モデルを説明する．2.3 節では，IVA の概要及び更新式を説明する．2.4 節では，ILRMA の概要及び更新式を説明する．2.5 節では，単一話者発話区間と IVA の分離性能の関係性について説明する．2.6 節では IVA の持つ関係性を応用した分離アルゴリズムについて説明する．

2.2 BSS における定式化と混合・分離モデル

短時間フーリエ変換（short-time Fourier transform: STFT）により時間周波数領域に変換した音声信号，観測信号，及び分離信号をそれぞれ次式で定義する．

$$\boldsymbol{s}_{ij} = [s_{ij1}, s_{ij2}, \cdots, s_{ijn}, \cdots, s_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N \tag{2.1}$$

$$\boldsymbol{x}_{ij} = [x_{ij1}, x_{ij2}, \cdots, x_{ijm}, \cdots, x_{ijM}]^{\mathrm{T}} \in \mathbb{C}^M \tag{2.2}$$

$$\boldsymbol{y}_{ij} = [y_{ij1}, y_{ij2}, \cdots, y_{ijn}, \cdots, y_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N \tag{2.3}$$

ここで，$i = 1, 2, \ldots, I$，$j = 1, 2, \ldots, J$，$n = 1, 2, \ldots, N$，及び $m = 1, 2, \ldots, M$ はそれぞれ周波数ビン，時間フレーム，音源，及び観測チャネル（マイクロホン）のインデクスを示し，$\cdot^{\mathrm{T}}$ は転置を表す．また，式 (2.1)，(2.2)，及び (2.3) の各信号においては，時間周波数行列（複素スペクトログラム）としての表記もそれぞれ $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$，$\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$ として定義する．式 (2.1)，(2.2)，及び (2.3) において，周波数毎の時不変な（時間フレーム j に依存しない）瞬時混合 Fig. 2.1 (a) のように，混合行列 $\boldsymbol{A}_i \in \mathbb{C}^{I\times J}$ を用いて次式で表せる．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{2.4}$$

この混合モデルは，時間領域において部屋の残響等の影響を受けて畳み込み混合となる観測信号を時間周波数領域で表現したものであるため，観測信号に対して式 (2.4) が成立することは，音響信号の BSS における基本的な仮定となる [1]．観測チャネル数と音源数が

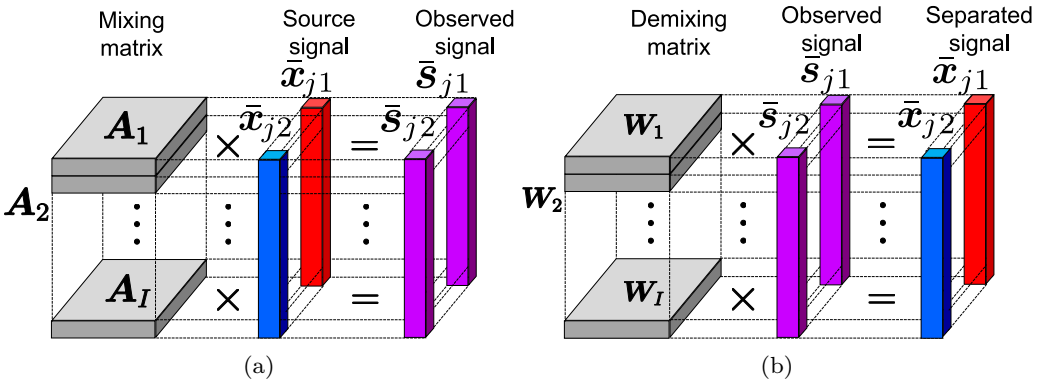

Fig. 2.1: (a) mixing and (b) demixing models in BSS. $\overline{\boldsymbol{s}}_{jn}$, $\overline{\boldsymbol{x}}_{jm}$, and $\overline{\boldsymbol{y}}_{jn}$ are vectors that contain all frequency components of nth source, mth observed, and nth separated signals at jth time frame.

等しい（$M = N$）かつ混合行列 $\boldsymbol{A}_i$ がフルランクの場合は，その逆行列である分離行列 $\boldsymbol{W}_i = [\boldsymbol{w}_{i1}\ \boldsymbol{w}_{i2}\ \cdots\ \boldsymbol{w}_{iN}]^{\mathrm{H}} \in \mathbb{C}^{N\times M}$ が存在し，これを用いて n 番目の音源の分離信号（推定信号）が Fig. 2.1 (b) や次式のように表せる．

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{2.5}$$

ここで，$\cdot^{\mathrm{H}}$ は行列及びベクトルのエルミート転置を表す．式 (2.4) 及び (2.5) を図示したものが Fig. 2.1 である．ここで，図中に登場する各ベクトルの定義は次式の通りである．

$$\overline{\boldsymbol{s}}_{jn} = [s_{1jn}, s_{2jn}, \cdots, s_{Ijn}]^{\mathrm{T}} \in \mathbb{C}^I \tag{2.6}$$
$$\overline{\boldsymbol{x}}_{jm} = [x_{1jm}, x_{2jm}, \cdots, x_{Ijm}]^{\mathrm{T}} \in \mathbb{C}^I \tag{2.7}$$
$$\overline{\boldsymbol{y}}_{jn} = [y_{1jn}, y_{2jn}, \cdots, y_{Ijn}]^{\mathrm{T}} \in \mathbb{C}^I \tag{2.8}$$

すなわち，これらはある音源又は観測チャネルの j 番目の時間フレームにおいて，全ての周波数ビンの成分をまとめたベクトルである．

BSS は，観測信号 $\boldsymbol{x}_{ij}$ から分離系である $\boldsymbol{W}_i$ を全ての周波数に関して正確に推定し，式 (2.5) で分離信号を得ることが目的となる．特に，IVA や ILRMA のように ICA に基づく BSS は，以下の 2 つの仮定を導入することでマイクロホンロホンや音源の位置，音源の種類やその学習データなどの音声信号の情報を必要とすることなく，分離系を推定することができる [14].

1. 混合前の各音源信号は互いに統計的に独立
2. 混合前の各音源信号は非ガウス分布に従う

2.3 節及び 2.4 節ではそれぞれ，BSS の一種である IVA 及び ILRMA について説明する．

2.3　IVA

本節では，音響信号処理分野で有名な周波数領域 BSS の一つである IVA について概説する．

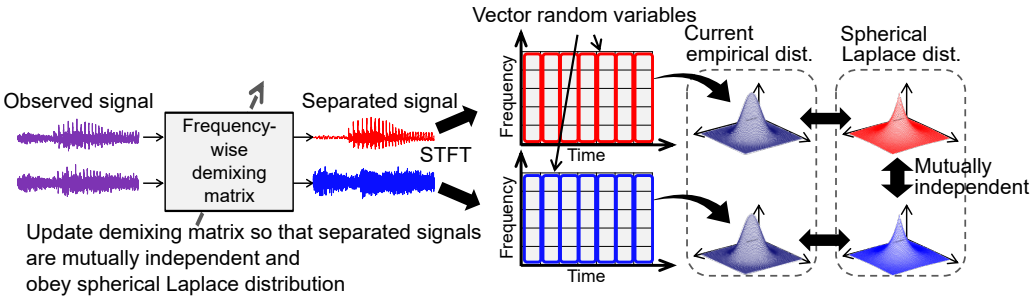

Fig. 2.2: Source and statistical models assumed in IVA.

Fig. 2.2 に $M = N = 2$ の場合における IVA の混合系及び分離系のモデル図を示す．IVA は時間周波数領域 BSS であるため，全周波数の分離行列 $(\boldsymbol{W}_i)_{i=1}^I$ を推定する．ただし，推定の過程において，全周波数を含む I 次元複素ラプラス分布を各音源 $\overline{\boldsymbol{s}}_{jn}$ 又は分離信号源 $\overline{\boldsymbol{y}}_{jn}$ の生成モデルと仮定している [2]．I 次元複素ラプラス分布は次式で定義される．

$$p(\overline{\boldsymbol{s}}_{jn}) = p(\overline{\boldsymbol{y}}_{jn}) \tag{2.9}$$

$$= \frac{1}{\pi \prod_i \sigma_{in}} \exp\left(-\sqrt{\sum_i \left| \frac{y_{ijn}}{\sigma_{in}} \right|^2} \right) \tag{2.10}$$

ここで，$\sigma_{in} > 0$ は複素ラプラス分布のスケールパラメータであり，IVA では $\sigma_{in} = 1 \ \forall i, n$ で固定される．この確率モデルに基づき導出される IVA の負対数尤度関数（目的関数）は次式で与えられる．

$$\underset{\boldsymbol{W}_1, \cdots, \boldsymbol{W}_I}{\text{Minimize}} \ -2J \sum_i \log |\det \boldsymbol{W}_i| + \sum_{j,n} G(\overline{\boldsymbol{y}}_{jn}) \tag{2.11}$$

ここで，det は行列式を表す．また，分離信号 $\overline{\boldsymbol{y}}_{in}$ 中に最適化変数の $\boldsymbol{W}_i$ が含まれる点に注意する．最適化問題 (2.10) 中の $G(\overline{\boldsymbol{y}}_{jn})$ はコントラスト関数と呼ばれ，次式で定義される．

$$G(\overline{\boldsymbol{y}}_{in}) = -\log p(\overline{\boldsymbol{y}}_{in}) \tag{2.12}$$

$$= -\log \frac{1}{\pi \prod_i \sigma_{in}} \exp\left(-\sqrt{\sum_i \left| \frac{y_{ijn}}{\sigma_{in}} \right|^2} \right) \tag{2.13}$$

$$= \log \pi + \sum_i \log \sigma_{in} + \sqrt{\sum_i \left| \frac{y_{ijn}}{\sigma_{in}} \right|^2} \tag{2.14}$$

結局，スケールパラメータを $\sigma_{in} = 1$ とおいて最適化変数に関する項だけで表すと，IVA の最適化問題 (2.10) は次に示す問題と等価になる．

$$\underset{\boldsymbol{W}_1, \cdots, \boldsymbol{W}_I}{\text{Minimize}} \ -2J \sum_i \log |\det \boldsymbol{W}_i| + \sum_{j,n} \|\overline{\boldsymbol{y}}_{jn}\|_2 \tag{2.15}$$

ここで，$\|\cdot\|_2$ は L_2 ノルムである．

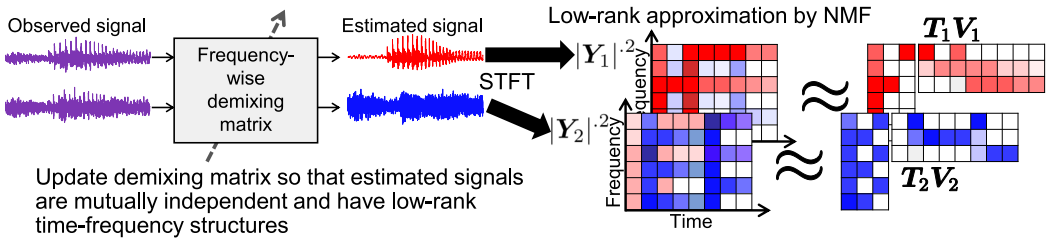

Fig. 2.3: Source and statistical models assumed in ILRMA.

IVA における分離行列 $(\boldsymbol{W}_i)_{i=1}^I$ の推定は，補助関数法 [15] 及び反復射影法（iterative projection: IP）[16] を用いた最適化アルゴリズムによって，高速かつ安定に解くことができる．この反復最適化更新則は下記の通りである．

$$\boldsymbol{G}_{in} = \frac{1}{J}\sum_j \frac{1}{\sqrt{\sum_i |\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} \tag{2.16}$$

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i\boldsymbol{G}_{in})^{-1}\boldsymbol{e}_n \tag{2.17}$$

$$\boldsymbol{w}_{in} \leftarrow \boldsymbol{w}_{in}(\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{G}_{in}\boldsymbol{w}_{in})^{-\frac{1}{2}} \tag{2.18}$$

ここで，$\boldsymbol{e}_n$ は n 番目の要素が 1，それ以外の要素が 0 の単位ベクトルである．上記の更新式に式 (2.16)–(2.18) を全ての n 及び i について計算することを 1 回の更新とし，複数回反復計算を行うことで IVA の目的関数の局所解である $(\boldsymbol{W}_i)_{i=1}^I$ を求めることができる．なお，この反復最適化アルゴリズムは 1 回の更新で負対数尤度関数 (2.11) の値が減少する又は変動しないこと（単調非増加性）が理論的に保証されている．

2.4　ILRMA

IVA よりも高精度な BSS を達成できるアルゴリズムとして ILRMA [4, 5] が提案されている．本章では ILRMA の仮定するモデルとその反復更新式について概説する．ILRMA の反復最適化の概要を Fig. 2.3 に示す．図中の $\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I\times K}$ 及び $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{K\times J}$ は，n 番目の音源のパワースペクトログラム $|\boldsymbol{Y}_n|^{\cdot 2}$ を非負値行列因子分解（nonnegative matrix factorization: NMF）[17] と呼ばれる行列分解アルゴリズムで低ランク近似した時間周波数モデルを構成する行列であり，$|\boldsymbol{Y}_n|^{\cdot 2} \approx \boldsymbol{T}_n\boldsymbol{V}_n$ としてモデル化されている．このとき，行列に対する絶対値記号とドット付き指数乗 $|\cdot|^{\cdot p}$ は要素毎の絶対値の指数乗を表す．また，$\boldsymbol{T}_n$ は基底行列，$\boldsymbol{V}_n$ はアクティベーション行列と呼ばれる．$\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の要素をそれぞれ t_{ikn} 及び v_{kjn} と定義する．ここで $k = 1, 2, \ldots, K$ は基底行列 $\boldsymbol{T}_n$ 中の列ベクトル（基底ベクトル）のインデクスを表す．ILRMA は，IVA に基づく分離行列 $\boldsymbol{W}_i$ の反復最適化と NMF の低ランク近似による分離信号のパワースペクトログラム $(|\boldsymbol{Y}_n|^{\cdot 2})_{n=1}^N$ の低ランクモデル $(\boldsymbol{T}_n\boldsymbol{V}_n)_{n=1}^N$ を反復的に交互最適化するアルゴリズムである．これにより，分離信号のパワースペクトルグラム $|\boldsymbol{Y}_n|^{\cdot 2}$ がいずれの音源も低ランクに近づくように分離行列 $\boldsymbol{W}_i$ が推定される．混合前の音源信号のパワースペクトルグラム $|\boldsymbol{S}_n|^{\cdot 2}$ が本来低ランクである場合，それらが混合した観測信号のパワースペ

クトルグラム $|\boldsymbol{X}_m|^{\cdot 2}$ はランクが大きくなるはずである．したがって，低ランクな時間周波数構造（パワースペクトログラム）を持つような分離行列 $\boldsymbol{W}_i$ を推定することは，より高精度な BSS を促す作用があり，多くの実験において ILRMA の音源分離性能が IVA の性能を上回ることが確認されている [4].

詳細な導出は省略するが，ILRMA の負対数尤度関数は以下のようになる [1, 5].

$$\underset{\boldsymbol{W}_1,\cdots,\boldsymbol{W}_I,\boldsymbol{T}_1,\cdots,\boldsymbol{T}_N,\boldsymbol{V}_1,\cdots,\boldsymbol{V}_N}{\text{Minimize}} \; -2J\sum_i \log|\det \boldsymbol{W}_i| + \sum_{i,j,n}\left(\frac{|\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2}{\sum_k t_{ikn}v_{kjn}} + \log\sum_k t_{ikn}v_{kjn}\right) \tag{2.19}$$

最適化問題 (2.19) を最小化する基底行列及びアクティベーション行列の要素は，次に示す反復最適化アルゴリズムで推定される．

$$t_{kjn} \leftarrow t_{ikn}\sqrt{\frac{\sum_j |\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2 v_{kjn}\left(\sum_{k'} t_{ik'n}v_{k'jn}\right)^{-2}}{\sum_j v_{kjn}\left(\sum_{k'} t_{ik'n}v_{k'jn}\right)^{-1}}} \tag{2.20}$$

$$v_{kjn} \leftarrow v_{kjn}\sqrt{\frac{\sum_i |\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2 t_{ikn}\left(\sum_{k'} t_{ik'n}v_{k'jn}\right)^{-2}}{\sum_i t_{ikn}\left(\sum_{k'} t_{ik'n}v_{k'jn}\right)^{-1}}} \tag{2.21}$$

一方，分離行列 $\boldsymbol{W}_i$ に関する最適化は，IVA と同様に IP を用いて分離ベクトル $\boldsymbol{w}_{in}$ を更新することで達成される．その更新式は次式となる．

$$\boldsymbol{U}_{in} = \frac{1}{J}\sum_j \frac{1}{\sum_k t_{ikn}v_{kjn}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} \tag{2.22}$$

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i\boldsymbol{U}_{in})^{-1}\boldsymbol{e}_n \tag{2.23}$$

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{U}_{in}\boldsymbol{w})^{-\frac{1}{2}} \tag{2.24}$$

これらの更新式も，IVA と同様に，1 回の更新の前後で目的関数 (2.19) の値が単調非増加となることが保証されている．

2.5 IVA の性能向上の条件

一般に音源分離問題において，各音源が時間周波数領域でどの程度お互いに重なっているかは分離信号の推定難易度に影響を与えるため重要である．特殊な音源の混合状況として，相互排他直交性（W-disjoint orthogonality: WDO）[18] という概念がある．WDO とは，Fig. 2.4 に示すように，観測された混合信号が，全ての周波数ビン及び時間フレームについていずれか一つの音源のみが常に支配的であり，すなわち排他的で直交している状況のことである．Fig. 2.4 (a) は WDO が成立していない状況であり，いくつかの時間周波数スロットにおいて赤色の音源と青色の音源が重なっている（紫色のセルで表している）．一方，Fig. 2.4 (b) は WDO が成立している状況であり，どの時間周波数スロットを見ても複数音源の重なりが無く，互いに排他的で直交している状況にある．そのため，次式に示す様に，WDO の関係性が

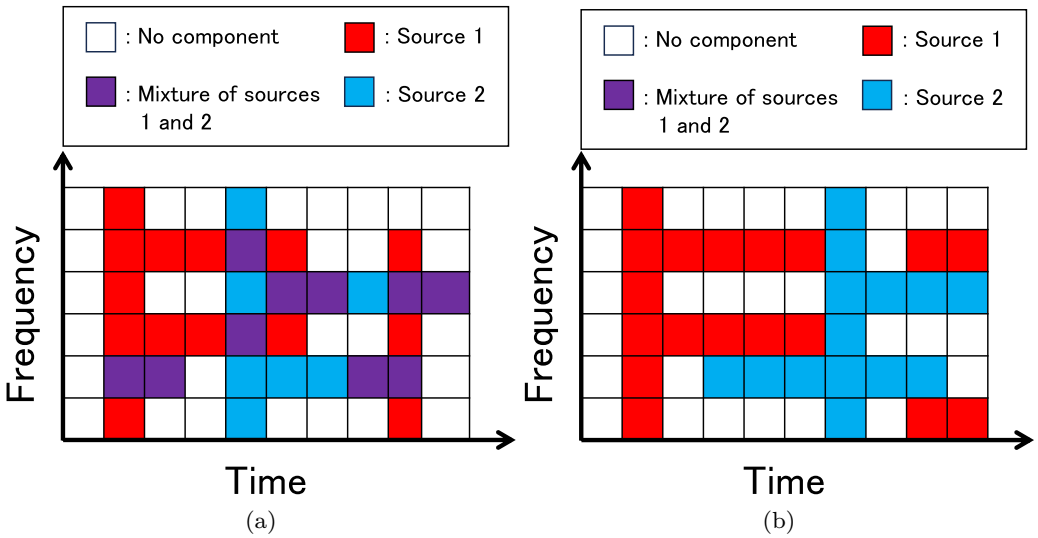

Fig. 2.4: Examples of (a) non-WDO and (b) WDO mixture signals, where red and blue source components are mixed ($N = 2$). In (a), some time-frequency slots has two source components as depicted in purple, whereas they are completely not overlapped in (b).

ある音源信号同士の積は 0 となる．

$$|s_{ijn}| \cdot |s_{ijn'}| = 0 \quad \forall i, j, n \neq n' \tag{2.25}$$

式 (2.25) は，同じ時間周波数スロット (i, j) において非零の値を持つ s_{ijn} が全ての音源 $n = 1, \cdots, N$ に対して 1 個しかないことを表している．このような WDO 仮定は，少ない音源数でかつ音声信号に対して比較的成立しやすいことが知られている [18]．これは，音声信号が時間周波数領域で比較的スパースな構造を持っていることに起因している．但し，たとえ音声信号であっても混合している音源の数 N が増加すると，当然このような WDO 仮定は成立しづらくなる．

WDO よりさらに強い直交性仮定として，フレーム単位相互排他直交性（flame-level W-disjoint orthogonality: F-WDO）が定義されている [13]．これは Fig. 2.5 に示すように，時間フレーム単位での WDO 仮定である．具体的には，ある時間フレームではすべての周波数ビンにおいていずれか一つの音源のみが常に支配的であり，すなわち排他的で直交している状況のことである．時間フレームは STFT を適用する際の「短時間区間」に相当するため，この短時間区間内では同時にアクティブとなっている音源や同時に発話している音声が無い，という強い仮定に相当する．Fig. 2.5 を見てわかるように，WDO は F-WDO よりも強い仮定であるため，Fig. 2.5 (b) のように WDO は成立するが F-WDO ではない状況もあり得る．以上より，F-WDO が成立する条件は次式のように表現できる．

$$\|\overline{\boldsymbol{s}}_{jn}\|_2 \cdot \|\overline{\boldsymbol{s}}_{jn'}\|_2 = 0 \ \forall i, j, n \neq n' \tag{2.26}$$

式 (2.26) は，同じ時間フレーム j において非零の L_2 ノルム値を持つ $\overline{\boldsymbol{s}}_{jn}$ が全ての音源

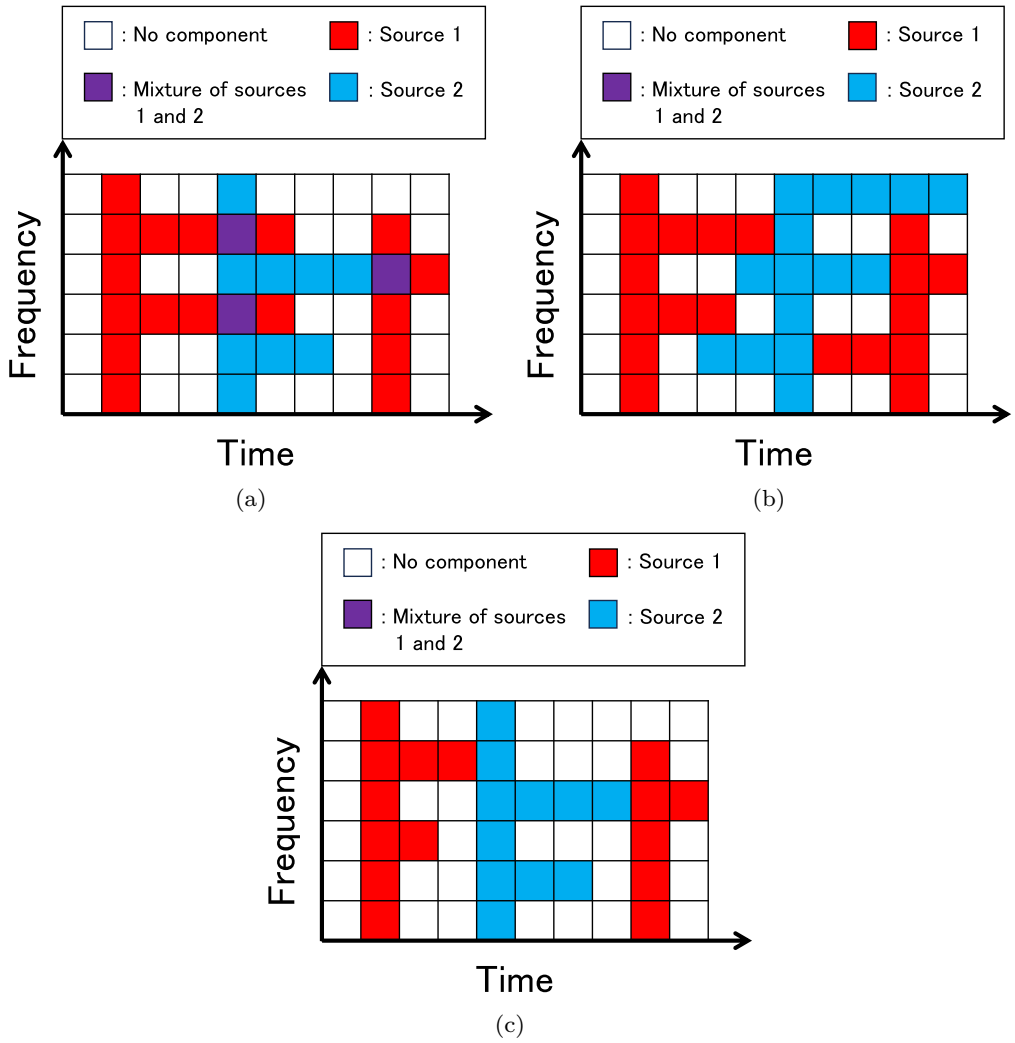

Fig. 2.5: Examples of (a) non-WDO, (b) WDO but non-F-WDO, and (b) F-WDO mixture signals, where red and blue source components are mixed ($N = 2$). In (a), some time-frequency slots has two source components as depicted in purple. In (b), there is no mixed (purple) slot, but some time frames include two source components in different frequency bins, whereas such time frames do not exist in (c).

$n = 1, \cdots, N$ に対して 1 個しかないことを表している．従って，アクティブでない音源は全周波数ビンの成分が 0 でなければならないため，時間フレーム j においてその音源は完全に無音であることを意味している．音声の混合においては，このような強い仮定は一見成立しないように思えるが，複数の話者が会話をしている状況の混合信号を考えると必ずしも起こり得ないことではない．通常の会話はターンテイキングを基本としており，誰かが話している時間は，他の話者は沈黙をしていることが自然である．実際には相槌等による多少の発話のオーバーラップは存在するとしても，通常の会話音声から成る混合信号は比較的 F-WDO の状態

に近い信号ということができる．

文献 [13] では，観測信号に含まれる全ての音源信号が F-WDO の条件を満たす場合に，最適化問題 (2.19) の大域最小解が最も音源分離性能を高めるような分離行列 $(\boldsymbol{W}_i)_{i=1}^I$ に対応する解となることを理論的に導出している．より厳密には，IVA や ILRMA でしばしば直面するパーミュテーション問題 [19, 20] が全く生じていない解 $(\boldsymbol{W}_i)_{i=1}^I$ が，F-WDO 成立時に最適化問題 (2.19) の第 2 項の大域最小解となることを示している．この事実は，観測信号が F-WDO の状態に近づくほど IVA の音源分離性能が向上することを意味している．前述の会話音声による混合信号のことを考慮すると，F-WDO に近づく状況とは，観測信号の全体時間区間に対して単一話者発話区間の割合が高くなることを意味する．従って IVA は，原理的には観測信号の単一発話発話区間の割合が高くなるほど，高精度な BSS を達成できるといえる．

一方，ILRMA については，観測信号の F-WDO 性や単一話者発話区間の割合と BSS の精度の関係は不明である．しかしながら，IVA と ILRMA の最適化問題の目的関数が類似していることや，分離行列の更新が同じ IP という最適化アルゴリズムに基づいていること考慮すると，ILRMA も IVA と同様の性質を持つことが予想される．

2.6　実際の信号への応用

前節で述べた通り，IVA は F-WDO の状態に近い観測信号程良い分離結果をもたらす性質がある．しかしながら，複数の音声信号が混合している観測信号がどの程度 F-WDO の状態に近いかは観測時の音源信号によって現象として決まるため，事前に制御できるようなものではない．それでも，IVA を適用する前に，得られた観測信号に対して何らかの操作を加えることで，F-WDO の状態に近づけることは現実的に可能である．具体的には，Fig. 2.6 に示すように，観測信号中の単一話者発話区間以外の波形を間引いてしまえば，完全に F-WDO 仮定が成立する状態に変形することも可能である．この処理の実現には，観測信号から単一話者発話区間を推定する必要が生じるが，これは音源分離よりもはるかに簡単な問題であるため発話区間検出 [21] や話者ダイアライゼーション [22] 等の既存技術を活用すれば実現できる可能性が高い．本論文では，観測信号中の単一話者発話区間が完全に推定でき，Fig. 2.6 のように単一話者発話区間以外の波形を間引くことができるという仮定をおいて以後の議論を進める．実際に単一話者発話区間を検出する技術については今後の課題とする．

IVA や ILRMA 等の BSS は式 (2.4) の時不変な（j に非依存な）混合系を仮定しているため，推定される分離行列 $(\boldsymbol{W}_i)_{i=1}^I$ も時不変である．そのため，Fig. 2.7 のように本来の観測信号を F-WDO になるように変形したとしても，その前後で分離行列 $(\boldsymbol{W}_i)_{i=1}^I$ の最適解が変化するわけではない．これはすなわち，Fig. 2.7 のような間引き処理で観測信号を F-WDO の状態に近づけ，これを新しい観測信号として IVA に入力して求まる分離行列 $(\boldsymbol{W}_i)_{i=1}^I$ が，間引き処理を行う前の（元々の）観測信号に対しても適用可能であることを意味する．理想的には，時間フレームを間引いた混合信号は F-WDO の状態に近づいているため，これを用いて推定される分離行列は元々の観測信号を用いて推定される分離行列よりも高精度であり，分離

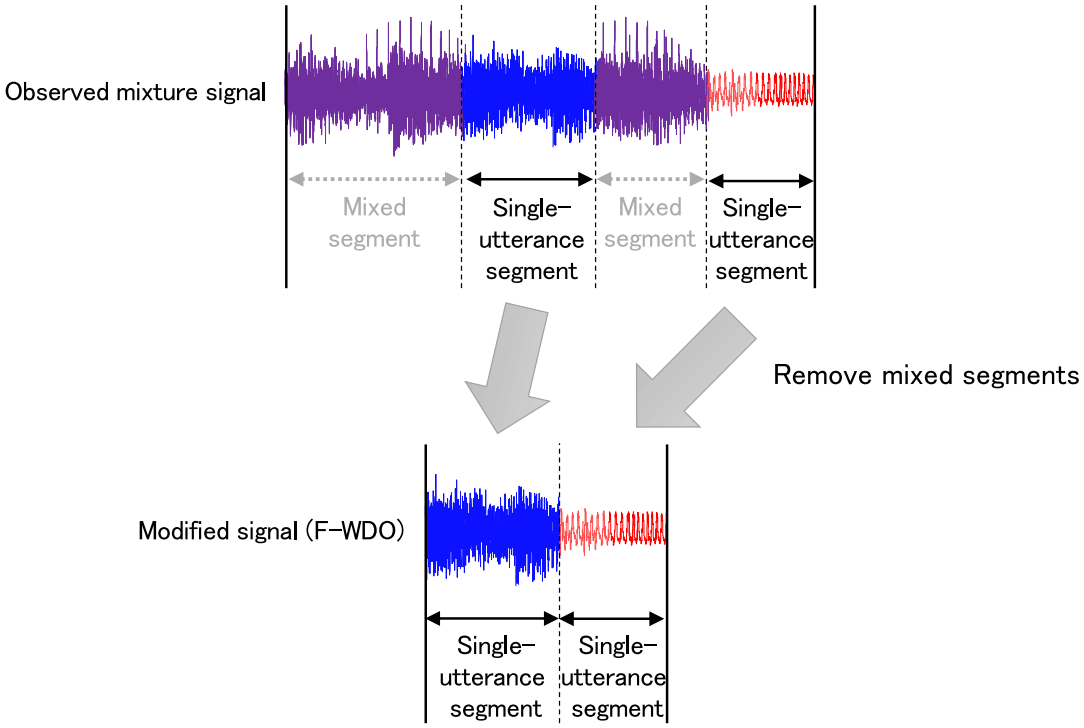

Fig. 2.6: Modification of observed signal to validate F-WDO assumption.

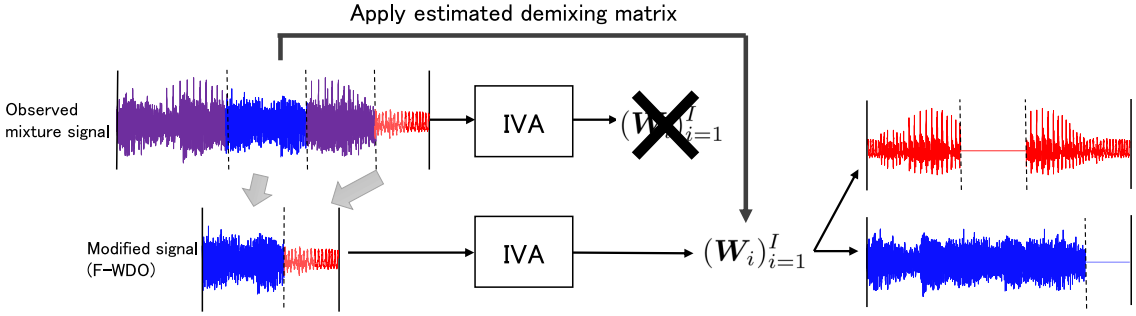

Fig. 2.7: Application of demixing matrix estimated by using modified signal that satisfy F-WDO. Since demixing matrix is time-invariant parameter, estimated demixing matrix is applicable to original observed mixture signal.

精度を大きく向上させられることが期待できる．このような方法で IVA や ILRMA の BSS 性能向上のアプローチを考えた場合，まず調査すべき事項は下記の 3 点となる．

- F-WDO に近い観測信号に対して IVA の BSS 性能はどの程度の向上が期待できるのか
- F-WDO の状態にどの程度近いかは単一話者発話区間の割合として同一視できるか
- IVA よりも高度な BSS である ILRMA においても同様の傾向は確認されるのか

従って本論文では，次章にて BSS の実験を構成し，その結果から上記の 3 点を詳しく考察する．

2.7　本章のまとめ

本章では本論文に関する基礎理論及びその応用について説明した．時間周波数領域 BSS の定式化を行い，その代表的な手法である IVA 及び ILRMA において，音源分離仮定及び反復更新式について説明した．また，先行研究 [13] で示された IVA の分離性能が向上する条件について説明し，実際の信号への応用について説明した．次章では IVA 及び ILRMA の BSS 性能と F-WDO や単一話者発話区間の割合との関係性を実験的に調査する．具体的には，音源データや録音環境などの実験方法と，単一話者発話区間を変化させたときの IVA 及び ILRMA の分離性能の変化についての実験結果を示し考察する．

第 3 章

単一話者発話区間の割合と BSS の性能の関係性の実験的調査

3.1 まえがき

本章では，観測信号が F-WDO の状態にどの程度近いか，すなわち単一話者発話区間の割合と IVA 及び ILRMA の性能の関係性を実験的に調査する．具体的には，本実験における観測データの生成方法，録音環境，IVA 及び ILRMA の実験条件，分離精度の客観評価尺度，及び実験結果について示す．実験結果では，IVA と ILRMA の両 BSS 手法に関して観測信号の単一話者発話区間の割合と分離精度の関係性を示し，その比較に基づく考察を述べる．

3.2 単一話者発話区間率

本研究では，観測信号がどの程度 F-WDO の状態に近いかという条件に着目し，BSS の精度を調査する．この F-WDO の状態に近いか否かを表す尺度としては，観測信号全体の時間長に対して単一話者発話区間がどれだけの時間を占めているかが対応する．したがって，F-WDO にどの程度近いかを表す指標として，単一話者発話区間率（active ratio of single speaker: ARSS）を定義する．

まず，時間領域の n 番目の音源信号，m 番目のマイクロホンの観測信号，及び n 番目の音源の分離信号をそれぞれ $\tilde{s}_n[l]$，$\tilde{x}_m[l]$，及び $\tilde{y}_n[l]$ と定義する．ここで，$l = 1, 2, \cdots, L$ は離散時間インデクスである．これらの信号をベクトル形式でも定義しておく．

$$\tilde{\boldsymbol{s}}_n = [\tilde{s}_n[1], \tilde{s}_n[2], \cdots, \tilde{s}_n[l], \cdots, \tilde{s}_n[L]]^{\mathrm{T}} \in \mathbb{R}^L \tag{3.1}$$

$$\tilde{\boldsymbol{x}}_m = [\tilde{x}_m[1], \tilde{x}_m[2], \cdots, \tilde{x}_m[l], \cdots, \tilde{x}_m[L]]^{\mathrm{T}} \in \mathbb{R}^L \tag{3.2}$$

$$\tilde{\boldsymbol{y}}_n = [\tilde{y}_n[1], \tilde{y}_n[2], \cdots, \tilde{y}_n[l], \cdots, \tilde{y}_n[L]]^{\mathrm{T}} \in \mathbb{R}^L \tag{3.3}$$

今，2.6 節で述べたように，観測信号 $(\tilde{\boldsymbol{x}}_m)_{m=1}^M$ から，各話者の発話/非発話状態が完全に推定できると仮定する．この各話者の発話/非発話の状態を 2 値で表すバイナリベクトルを $\tilde{\boldsymbol{d}}_n = [\tilde{d}_n[1], \tilde{d}_n[2], \cdots, \tilde{d}_n[l], \cdots, \tilde{d}_n[L]]^{\mathrm{T}} \in \{0, 1\}^L$ と定義する．発話状態が 1，非発話状態

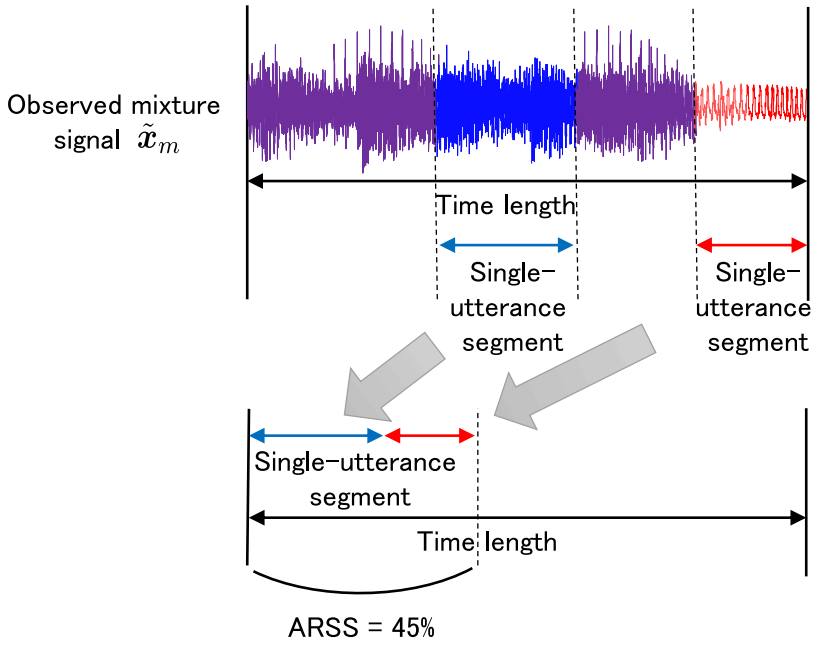

Fig. 3.1: Calculation of ARSS.

が 0 に対応する．このとき，ARSS は次式で定義できる．

$$\mathrm{ARSS} = \|\tilde{\boldsymbol{r}}\|_1 \tag{3.4}$$

$$\tilde{\boldsymbol{r}} = [\tilde{r}[1], \tilde{r}[2], \cdots, \tilde{r}[l], \cdots, \tilde{r}[L]]^{\mathrm{T}} \tag{3.5}$$

$$\tilde{r}[l] = \begin{cases} 1 & (\text{if } \sum_n \tilde{d}_n[l] = 1) \\ 0 & (\text{otherwise}) \end{cases} \quad \forall l \tag{3.6}$$

ここで，$\|\cdot\|_1$ は L_1 ノルムである．ARSS の定義の概要は Fig. 3.1 に示すとおりである．ARSS は観測信号の時間長 L に対して，いずれかの話者 1 人が発話している時間インデクスの割合に対応する．ARSS は 0 から 1（0 から 100%）の範囲の値をとり，ARSS が 0% であれば観測信号中に単一話者発話区間が存在しないことを意味する．逆に ARSS が 100% であれば，その観測信号は F-WDO の状態であることになる．

3.3　音源データの作成

本節では 3.2 節で説明した ARSS を変化させた観測信号をそれぞれ IVA や ILRMA に入力し，そのときの分離性能を比較する．本実験を実現するには，様々な ARSS 値に対応する観測信号を網羅的に用意した音源セットのデータ群が必要である．本論文では，2 人の話者を混合させて生成する観測信号について，一方の話者の非発話区間（無音区間）を制御することによって ARSS を 10% 刻みに 10% から 90% まで用意する．Fig. 3.2 に示すように，まず性別に関して「男性と男性」及び「女性と女性」の 2 種類の組み合わせを考える．「男性と女性」の組み合わせについては，基本周波数の違いから BSS の難度が低下するため，本論文ではより

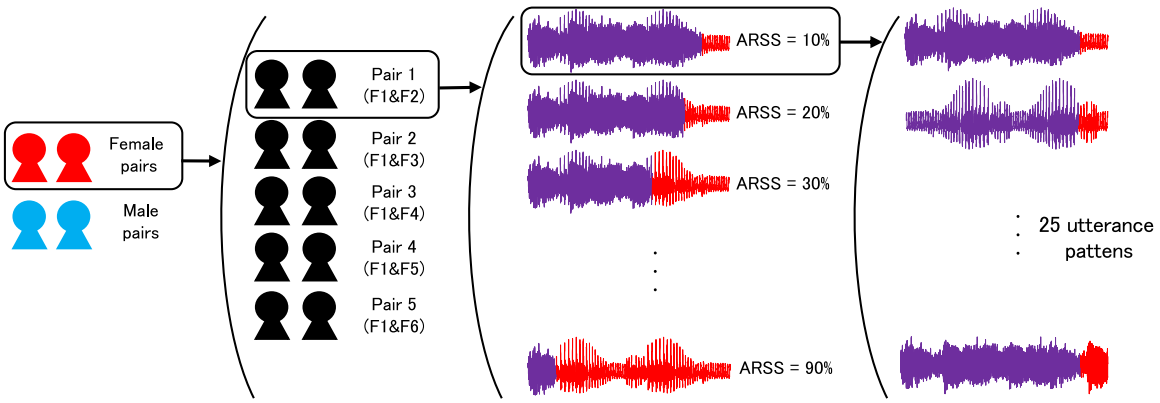

Fig. 3.2: Preparation of observed mixture signals with various ARSS, where purple waveform depicts mixture of two speech sources. In total, 2,250 observed mixture signals are prepared.

難しい条件の同性話者の音声の混合のみを対象とする．次に各性別の組み合わせにおいて，話者の異なる 5 種類の話者ペアを用意する．この各話者ペアにおいて，10% から 90% までの 9 種類の ARSS の条件値を考え，さらに 1 つの ARSS の条件値に対して異なる発話を用いた 25 パターンの音源信号のペアを用意する．結局合計としては，2,250 個の音源信号のペアを用意している．

ARSS が高い場合，一方の音源信号の発話区間が短くなってしまうため，両話者の観測信号全体のエネルギー値に差が生じてしまう．このようにエネルギーの不均衡な音源信号が混合された観測信号の BSS では，当然エネルギーの小さい（発話区間の短い）方の音源の分離が難しくなる．本実験においては，各音源信号のエネルギーの差に由来する音源分離性能の変化は一切排除し，純粋に ARSS の違いのみによって生じる音源分離性能の変化のみを観測することが目的である．そのため，Fig. 3.3 に示すように，エネルギーの不均衡な音源信号ペアでも混合した後の観測信号全体のエネルギー値が音源間で等しくなるように，発話区間の短い音源信号の振幅を調整して正規化する．この処理を音源信号 $\tilde{\boldsymbol{s}}_1$ と $\tilde{\boldsymbol{s}}_2$ に対して適用することを考えると，$\tilde{\boldsymbol{s}}_1$ と $\alpha\tilde{\boldsymbol{s}}_2$（$\alpha > 0$ は振幅の倍率）のエネルギー比が等しくなるという条件より，振幅の倍率が次式として求まる．

$$\alpha = \sqrt{\frac{{\|\tilde{\boldsymbol{s}}_1\|_2}^2}{{\|\tilde{\boldsymbol{s}}_2\|_2}^2}} \tag{3.7}$$

$$= \sqrt{\frac{\sum_l |\tilde{s}_1[l]|^2}{\sum_l |\tilde{s}_2[l]|^2}} \tag{3.8}$$

従って，$\tilde{\boldsymbol{s}}_1$ と $\alpha\tilde{\boldsymbol{s}}_2$ を混合前の音源信号ペアとして用い，次節で述べる方法で観測信号を生成する．

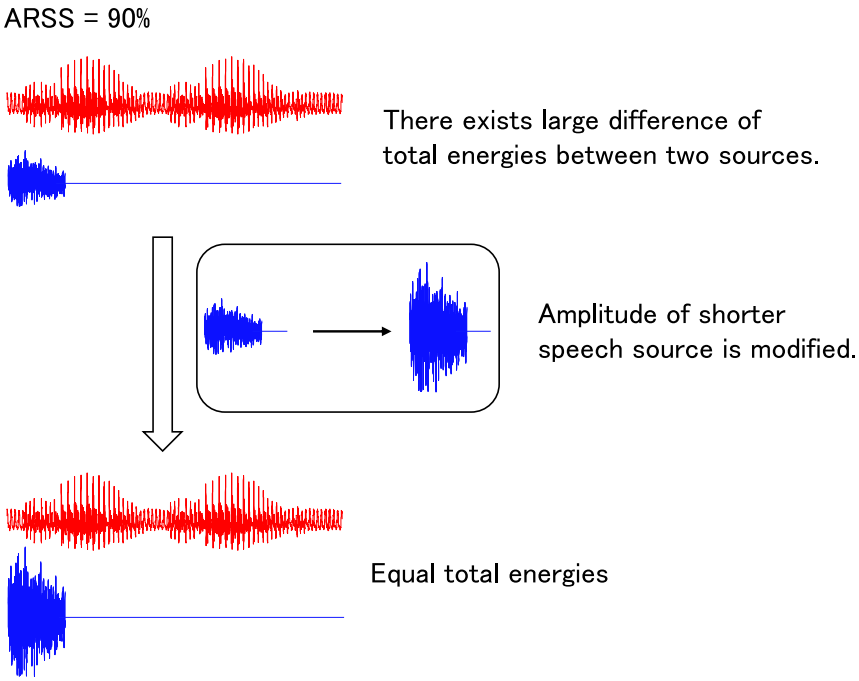

Fig. 3.3: Equalization of speech energies between two sound sources. Amplitude of shorter speech source is modified so that total energies of both sources are equal.

3.4　インパルス応答の畳み込み

3.3 節で用意した音源信号は混合される前の状態である．音源セットを混合するにあたって，新情報処理開発機構（real world computing partnership: RWCP）データベース [23] 収録のインパルス応答 E2A（残響時間 $T_{60} = 300$ ms）による 2 音源の畳み込み混合シミュレーションを行う．E2A は，様々な空間配置のマイクロホンや音源間のインパルス応答を収録しており，このインパルス応答を音源信号に畳み込むことは，実際にその環境で音源信号を再生した際に録音される信号を模擬することに相当する．本実験で用いた E2A 収録のインパルス応答のマイクロホン及び音源の配置図を Fig. 3.4 に示す．音の到来方向はマイクロホンアレイに対して同角度であり，マイクロホンから音源までの距離も等しい．これは，2 人の話者が会話している状況を 1 台のマイクロホンアレイで録音するという典型的な収録環境を模擬したものである．

3.5　客観評価尺度

本研究では IVA 及び ILRMA の客観精度評価尺度として信号対歪み比（source-to-distortion ratio: SDR）[24] を用いる．SDR は各音源の分離度合いと歪みの少なさの両方を捉えた総合的な音源分離指標であり，次のように計算される．

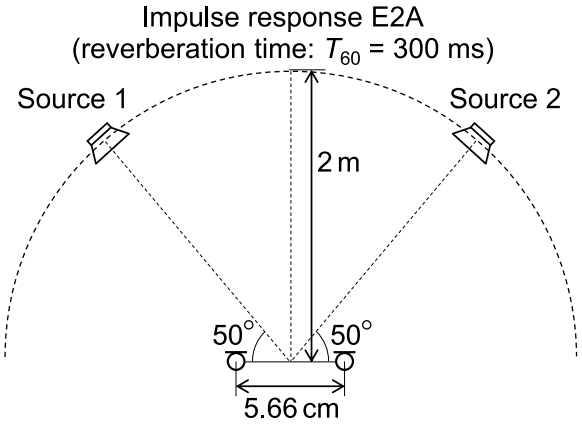

Fig. 3.4: Recording environment of impulse responses that are used for simulating observed mixture signals.

今，n 番目の目的音源 $\tilde{\boldsymbol{s}}_n$ に対応する推定信号 $\tilde{\boldsymbol{y}}_n$ が，次のような成分から構成される仮定する．

$$\tilde{\boldsymbol{y}}_n = \tilde{\boldsymbol{s}}_{\text{target}} + \tilde{\boldsymbol{e}}_{\text{interf}} + \tilde{\boldsymbol{e}}_{\text{artif}} \tag{3.9}$$

ここで，$\tilde{\boldsymbol{s}}_{\text{target}} \in \mathbb{R}^L$，$\tilde{\boldsymbol{e}}_{\text{interf}} \in \mathbb{R}^L$，及び $\tilde{\boldsymbol{e}}_{\text{artif}} \in \mathbb{R}^L$ はそれぞれ推定信号 $\tilde{\boldsymbol{y}}_n$ 中の $\tilde{\boldsymbol{s}}_n$ に対応する成分，残留した $\tilde{\boldsymbol{s}}_n$ 以外の非目的音源に対応する成分，及び音源分離によって生じた人工的な歪み成分を表す．このとき，SDR は，次式で定義される．

$$\text{SDR} = 10 \log_{10} \frac{\|\tilde{\boldsymbol{s}}_{\text{target}}\|_2}{\|\tilde{\boldsymbol{e}}_{\text{interf}} + \tilde{\boldsymbol{e}}_{\text{artif}}\|_2} \text{ [dB]} \tag{3.10}$$

したがって，高い SDR 値を達成するには，推定信号 $\tilde{\boldsymbol{y}}_n$ に残留する非目的音源成分 $\tilde{\boldsymbol{e}}_{\text{interf}}$ と人工的な歪み成分 $\tilde{\boldsymbol{e}}_{\text{artif}}$ のエネルギーがどちらも小さく，本来推定されるべき目的音源成分 $\tilde{\boldsymbol{s}}_{\text{target}}$ のエネルギーが大きい必要がある．SDR は音源分離における標準的な客観評価尺度として用いられており，10 dB を上回ると比較的良い分離性能を達成しているといえる．

3.6 その他の実験条件

3.3 節で述べた話者の発話音源信号には，様々な話者の日本語発話データが収録されている Japanese versatile speech（JVS）corpus [25] の一部分を使用した．このとき，JVS corpus では音源データのサンプリング周波数が 48 kHz で提供されているが，本論文では計算量削減のために全ての音声信号を 48 kHz から 16 kHz にダウンサンプルした．なお，音声信号は 8 kHz 以下にエネルギーが集中しているため，サンプリング周波数を 16 kHz にダウンサンプルしたうえで信号処理を適用することは一般的である．本実験では JVS corpus 内の parallel 100 というデータを用いて，女性及び男性の話者をそれぞれ 6 名ずつ用いた．ここで，女性話者 6 名には F1 から F6，男性話者 6 名には M1 から M6 というラベルをそれぞれ付与し，Fig. 3.2 に示すように 5 種類の話者ペアを女性と男性で 5 種類ずつ構成した．具体的には，女性の

話者ペアが F1&F2，F1&F3，F1&F4，F1&F5，及び F1&F6 の 5 種類，男性の話者ペアが M1&M2，M1&M3，M1&M4，M1&M5，及び M1&M6 の 5 種類である．

その他，STFT のパラメータである窓長及びシフト長はそれぞれ 4096 点（256 ms）及び 2048 点（128 ms）とし，窓関数は blackman 窓を用いた．また，IVA 及び ILRMA の更新式は 2 章で示したものを使用し，反復更新回数をいずれの手法も 200 回とした．IVA と ILRMA の両方において分離行列 $(\boldsymbol{W}_i)_{i=1}^{I}$ の初期値は全て単位行列とし，また ILRMA の基底行列 $(\boldsymbol{T}_n)_{n=1}^{2}$ 及びアクティベーション行列 $(\boldsymbol{V}_n)_{n=1}^{2}$ の初期値はいずれも区間 $(0, 1)$ の一様分布乱数とした．このように ILRMA の初期値は $(\boldsymbol{T}_n)_{n=1}^{2}$ と $(\boldsymbol{V}_n)_{n=1}^{2}$ に関して乱数要素を含むが，本実験では 1 つの話者ペア・ARSS 値条件に対して 25 パターンもの観測信号を用意しているため，乱数の種類（乱数シードの種類）は 1 種類として結果を集計した．

最後に，本実験では SDR の改善量（分離信号の SDR 値から観測信号の SDR を減算した値）を結果として示す．この原因として，3.3 節で述べた通り音源信号のエネルギーは等しくなるように正規化されているが，3.4 節で述べたようにインパルス応答を畳み込む段階で再びわずかなエネルギー差が生じ，観測信号の SDR が 0 dB にならない為である．

3.7　実験結果

3.7.1　IVA

分離アルゴリズムが IVA の場合の SDR 改善量を Figs. 3.5 及び 3.6 に示す．Figs. 3.5 が女性話者のペア 5 種類，Figs. 3.6 が男性話者のペア 5 種類の結果に対応している．各グラフはいずれも箱ひげ図を用いている．青色の箱の内側の水平線は中央値，赤い丸印は平均値を表し，箱の下端及び上端はそれぞれ第 1 四分位数及び第 3 四分位数を示す．さらに，箱の下と上に伸びるひげはそれぞれ，箱の下端及び上端から箱の長さの 1.5 倍離れた範囲内にある最小値と最大値を示し，前述の範囲を超えるサンプルは外れ値として青い丸印で描かれている．

結果より，Figs. 3.5 (b) 及び 3.6 (b) のように ARSS の増加に対して SDR 改善量が継続的に増加するパターンと Figs. 3.5 (a)，3.5 (c)，3.5 (d)，3.5 (e)，3.6 (a)，3.6 (c)，3.6 (d)，及び 3.6 (e) のように改善が ARSS 50% 程度から飽和するパターンの 2 種類が見られた．前者のパターンの結果に関しては ARSS が高くなるにつれて SDR の中央値が増加した．特に Fig. 3.5 (b) の話者ペア M1&M3 の結果については，ARSS の増加に対して明確に音源分離性能が改善していく様子が確認できる．この結果は 2.5 節で説明した IVA の特性からもうかがえるため，予想された結果である．一方，後者の改善量が飽和するケースについては，性能の改善が止まってしまったというよりは，比較的低い ARSS の時点で十分な SDR の改善が得られて飽和しているように見える．すなわち，ARSS が 10% から 30% 程度は音源分離性能の変化が非常に大きいが，ARSS が 50% 程度まで大きくなると，IVA が達成することのできる性能の限界まで改善されている，と解釈するのが妥当な結果と思われる．なお，一部の結果においては，ARSS が 70% から 80% を超えたあたりから，SDR がやや減少する傾向を確認した．これは 2.5 節で説明した IVA の特性からはうかがえない予想に反した結果である．この原因

については未解明であるが，ARSS が 100% に近づくほど分離行列の更新式 (2.16)–(2.18) の計算における数値不安定性が顕著になってくるのではないかと推測している．具体的には，式 (2.16) で分離信号の j 番目の時間フレームの全周波数パワー $\sum_i |\boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{x}_{ij}|^2$ を計算し分母に用いているが，これが 0 に近づくほど行列 $\boldsymbol{G}_{in}$ は特異行列に近づく（条件数が増加しランク落ちの状態に近づく）．その結果，式 (2.17) に含まれる逆行列演算の数値不安定性が大きくなり，これが分離行列の正確な推定に悪影響を及ぼしているのではないか，と予想される．但し，もしこの予想が真実であったとしても，これは最適化アルゴリズムに起因する現象であるため，文献 [13] で提唱された「F-WDO の場合に IVA の性能が向上する」という結論が否定されるわけではない．

3.7.2 ILRMA

分離アルゴリズムが ILRMA の場合の SDR 改善量を Figs. 3.7 及び 3.8 に示す．Fig. 3.7 が女性話者のペア 5 種類，Fig. 3.8 が男性話者のペア 5 種類の結果に対応している．

これらの結果を見ると，まず IVA の結果よりも顕著な SDR の向上が確認できなくなっていることが分かる．例えば，Fig. 3.7 (a) では ARSS が 20% の時が最良となり，それより大きい ARSS の場合は性能が低下していく様子が確認できる．一方，IVA のときに SDR が大きく向上していた Figs. 3.5 (b) 及び 3.6 (b) と同じ観測信号である Figs. 3.7 (b) 及び 3.8 (b) では，ILRMA の場合もやや SDR が上昇する傾向がみられる．それでも，ARSS が 50% または 60% 付近が最高性能となり，それ以上の ARSS では性能が低下している．これらの結果は，ILRMA においても IVA と同様に ARSS が高いほど性能が向上すると思われた当初の予想と反している．推測となるが，考えられる要因は IVA の場合と同様に，最適化アルゴリズム中の数値計算の不安定性と予想される．

ILRMA の最適化アルゴリズムは式 (2.20)–(2.24) を反復的に計算するが，分離行列 $(\boldsymbol{W}_i)_{i=1}^{I}$ の更新式 (2.22) で低ランク近似された音源毎のパワースペクトログラム（モデルスペクトログラム）$\boldsymbol{T}_n \boldsymbol{V}_n$ の各要素が分母に現れる．例えば $\boldsymbol{T}_n \boldsymbol{V}_n$ がある時間周波数要素 (i, j) でほとんど 0 となってしまえば，やはり行列 $\boldsymbol{U}_{in}$ の条件数が増加しランク落ちの状態に近づく．その結果，式 (2.23) の逆行列演算に大きな数値誤差が含まれることになる．ILRMA の最適化アルゴリズムは元来このような数値不安定性をはらんでいるが，通常は前述の現象を避けるために，モデルスペクトログラム $\boldsymbol{T}_n \boldsymbol{V}_n$ の要素を計算機イプシロンでフロアリングする処理を式 (2.20) 及び (2.21) の計算後に毎反復適用する．しかしながら，これは本質的な解決ではなく，観測信号や初期値の条件によってはやはり式 (2.23) の逆行列演算が不安定となることがある．本質的な解決として，逆行列演算を伴わない最適化アルゴリズムで IVA 及び ILRMA の最適化問題を解くことが挙げられる．そのようなアルゴリズムは，IP の代替として iterative source steering（ISS）法が提案されており [6]，前述のような行列の条件数の悪化が起こる観測信号において，数値誤差が結果に影響しにくいことが予想される．もし ISS に基づく IVA や ILRMA で本章の実験結果の傾向が変わるならば，前節及び本節で述べた推測の信憑性が上

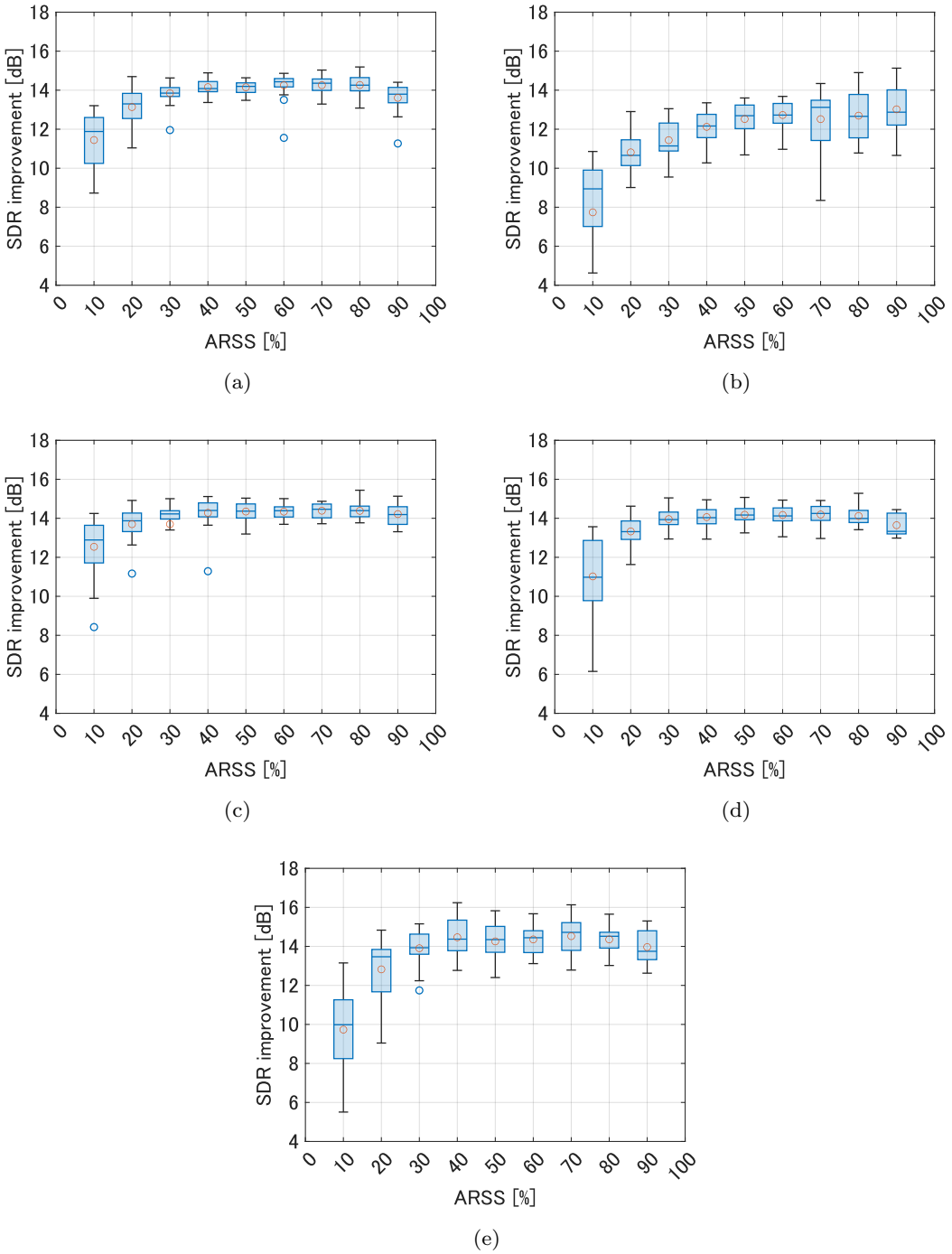

Fig. 3.5: Results of SDR improvement obtained by IVA for speaker pair (a) F1&F2, (b) F1&F3, (c) F1&F4, (d) F1&F5, and (e) F1&F6.

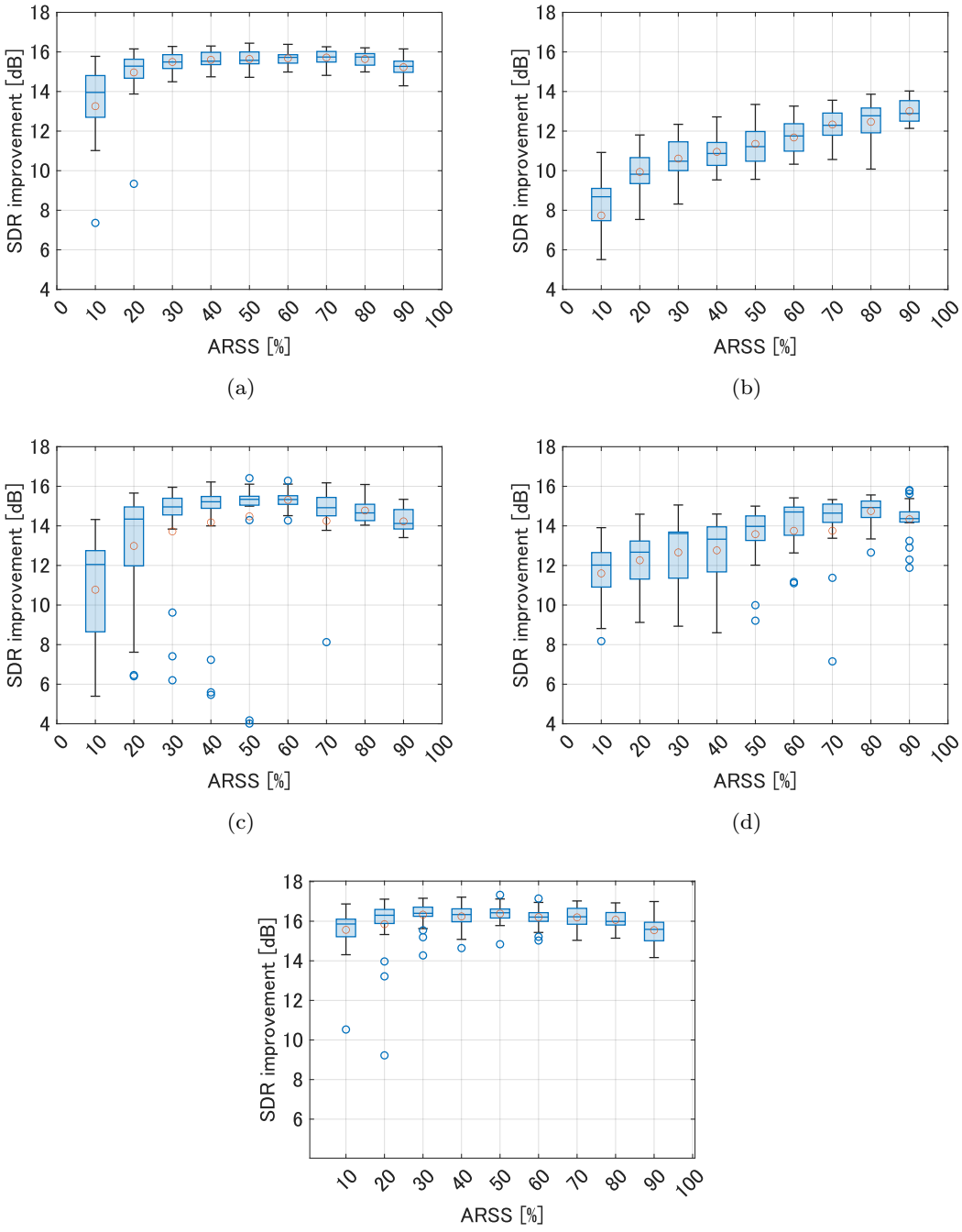

Fig. 3.6: Results of SDR improvement obtained by IVA for speaker pair (a) M1&M2, (b) M1&M3, (c) M1&M4, (d) M1&M5, and (e) M1&M6.

がると思われるため，これを今後の課題とする．

最後に，ILRMA の分離結果は IVA よりもやや分布が広がっている傾向や外れ値が多い様子が確認できる．この現象は原理的に起こる問題であり，音源モデル $\boldsymbol{T}_n\boldsymbol{V}_n$ だけ最適化変数が増えている ILRMA がより難しい問題を解いていることに起因する．さらに，$\boldsymbol{T}_n\boldsymbol{V}_n$ の初期値を乱数にしたことも，本質的に IVA よりも ILRMA の方が性能に不安定さを生じさせる原因となっている．NMF や ILRMA の最適化を安定させる初期値等も提案されており（例えば [26]），これらの技術を活用することで，不安定性を解消できる可能性がある．あるいは，IVA の分離信号を NMF に適用して $\boldsymbol{T}_n\boldsymbol{V}_n$ の初期値を得るという方法等も考えられる．

以上をまとめると，ILRMA で IVA と同様の性能向上が得られなかったことには複数の要因があると推測される．行列の条件数に起因する数値的不安定性と，最適化初期値に起因する性能の不安定性の 2 点を何らかの方法で改善した場合に，ILRMA の ARSS に対する性能の変化が IVA と同様にさらに向上するならば有用な知見となる．

3.7.3　IVA と ILRMA の比較と考察

3.7.1 項及び 3.7.2 項で示した実験結果を比較すると，次の 2 点が確認できる．

- IVA は ARSS が低い観測信号に対して十分な性能を達成できない
- ILRMA は ARSS が高い観測信号に対して性能の劣化が生じる

特に，これまで [6] 等多くの文献で実証されてきた ILRMA の IVA に対する優位性は，常に全て話者が発話している音声信号や常に多くの楽器が演奏されている音楽信号の BSS の実験で結論付けられたものであり，これはすなわち ARSS が 0% から 10% 程度と低い観測信号に対して確認できる優位性であった可能性が高い．しかしながら，文献 [13] で証明された高い ARSS の観測信号に対する IVA の性能向上という原理を考慮すると，本来高い ARSS の観測信号の BSS 性能は向上するはずであるため，ILRMA の更なる性能向上が前項で述べた数値不安定性等に要因により阻害されているのではないか，という推測ができる．この推測の検証については今後の課題とする．

3.8　本章のまとめ

本章では，観測信号が F-WDO の状態に近いか否かを表す指標として ARSS を定義し，観測信号の ARSS と IVA 及び ILRMA の BSS の性能の関係性について実験的に調査した．実験結果では，先行研究 [13] で提唱されていた現象が IVA の BSS 結果で確認されたが，同じ最適化アルゴリズムに基づいている ILRMA ではその現象が確認されなかった．その理由について，現時点で考えられる推測を述べた．次章では，本論文全体をまとめる．

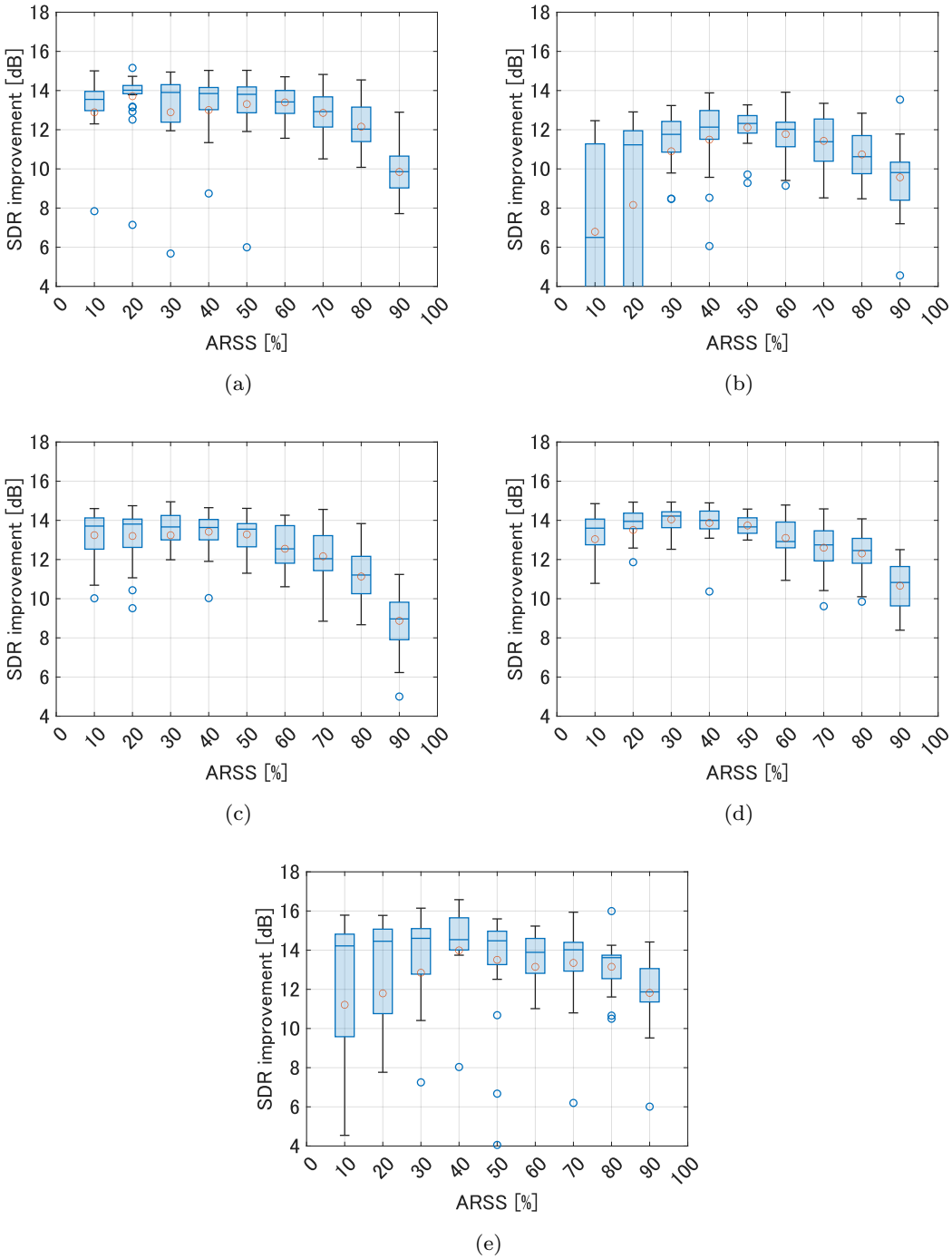

Fig. 3.7: Results of SDR improvement obtained by ILRMA for speaker pair (a) F1&F2, (b) F1&F3, (c) F1&F4, (d) F1&F5, and (e) F1&F6.

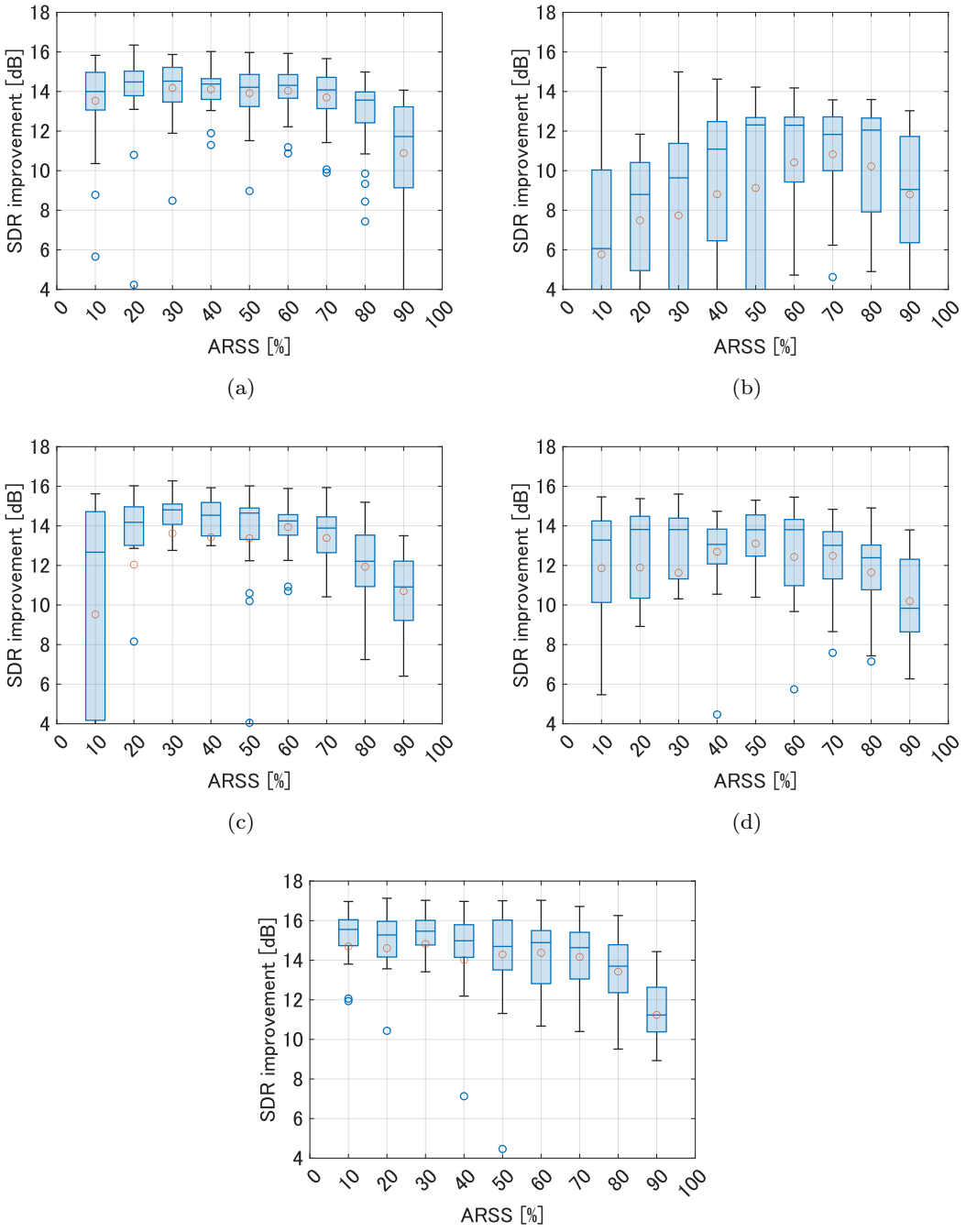

Fig. 3.8: Results of SDR improvement obtained by ILRMA for speaker pair (a) M1&M2, (b) M1&M3, (c) M1&M4, (d) M1&M5, and (e) M1&M6.

第 4 章

結言

本論文では，「混合音源が F-WDO の状態に近づくにつれて，分離精度が向上する」という性質を IVA については追試し，ILRMA については同様の結果が得られるか検証することを目的として，ARSS を変化させて IVA 及び ILRMA の分離精度がどのように変化するか実験を行った．IVA は上記の性質が見られたが，実験結果の多くがある一定の ARSS で分離精度が飽和していた．また，ILRMA には IVA と同様の結果が得られなかった．この原因として，ILRMA の数値計算の不安定性（行列の条件数に起因する数値的不安定性及び最適化初期値に起因する性能の不安定性）が挙げられるが，詳細は現時点では不明である．これまでの文献で実証されてきた ILRMA の IVA に対する優位性は，ARSS が 0% から 10% 程度と低い観測信号に対して確認できる優位性であった可能性が高いことが言える．さらに，反復更新式の不安定性が本論文の結果に大きく作用していると予想できる．IP の代案として提案された ISS では ILRMA の不安定性が本論文の分離性能の結果に影響しにくいことが予想される．IP に基づく ILRMA の結果が IP の数値不安定性によるものであれば，ISS に基づく ILRMA では IVA と同様の結果が得られるはずであるため，ISS においても IP と同様の分離性能の結果が示されれば，3.7 節で述べた推測の信憑性が上がると思われる．以上を踏まえて，今後の課題として以下の 3 つが挙げられる．

- 単一話者発話区間と分離性能の関係における精度飽和の原因調査
- IVA 及び ILRMA の単一話者発話区間と分離性能の関係性の違いの原因調査
- ISS に基づく IVA 及び ILRMA の単一話者発話区間と分離精度の関係の調査

また，F-WDO と IVA の分離性能の関係を応用する例として，観測信号中の単一話者発話区間以外の波形を間引いて，F-WDO 仮定が成立する状態に変形した混合音源を用いて分離系を求める方法を述べた．F-WDO の状態に編集した混合信号から推定する分離系と元の混合信号から推定される分離系は等しく，さらに F-WDO の状態に編集した混合信号の方が分離性能が高くなる傾向がある．そのため．この応用例は従来の方法よりも分離精度が高くなると予想される．

謝辞

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

まず，本研究を進めるにあたり，ご多忙のところ熱心にご指導くださいました指導教員の北村大地講師に心より感謝申し上げます．北村大地講師には，論文執筆や研究に関する議論など，細部にわたるまで丁寧にご指導いただきました．音響信号処理や数学などの知識や研究に用いる計算機設備の設置および利用など，貴重な経験となりました．

本論文の副査である重田和弘教授には，論文の構成や記述に関して有益な助言を頂き，大変お世話になりました．ここに厚く御礼申し上げます．

北村研究室の先輩である専攻科 2 年の川口翔也氏，蓮池郁也氏，溝渕悠朔氏，村田佳斗氏及び専攻科 1 年の綾野翔馬氏には研究や論文執筆の基礎や MATLAB に関するアドバイスなど数々のご支援をいただきました．また，北村研究室同期の加藤大輝氏，松本愛花氏，和気祐弥氏には研究のみならず一年間の学校生活を様々な面で支えていただきました．研究室で大変充実した時間を過ごすことができました．心より感謝申し上げます．

最後になりますが，現在に至るまで私の学生生活を金銭的に支え，暖かく見守って下さった両親には感謝の念に堪えません．これまで本当にありがとうございました．

参考文献

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

[2] 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.

[3] 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*, pp. 189–192, 2011.

[4] 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.

[5] 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.

[6] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 236–240. 2020.

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

[8] T. Nakashima, R. Ikeshita, N. Ono, S. Araki, and T. Nakatani, "Fast online source steering algorithm for tracking single moving source using online independent vector analysis," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. 2023.

[9] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for de-

termined audio source separation," *IEEE/ACM Transaction on Audio, Speech, and Language, Processing*, vol. 27, no. 10, pp. 1601–1615, 2019.

[10] D. Kitamura and K. Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, no. 46, 2020.

[11] 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.

[12] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 28, 2018.

[13] J. Gu, L. Cheng, D. Yao, J. Li, and Y. Yan, "The effect of source sparsity on independent vector analysis for blind source separation," *Signal Processing*, vol. 213, pp. 3–8, 2023.

[14] 北村大地, 小野順貴, 澤田宏, 亀岡弘和, 猿渡洋, "独立低ランク行列分析に基づくブラインド音源分離," 電子情報通信学会 技術研究報告, EA2017-56, vol. 117, no. 255, pp. 73–80, 2017.

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

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

[17] 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.

[18] S. Rickard and Ö. Yılmaz, "On the approximate W-disjoint orthogonality of speech," *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I-529–I-532, 2002.

[19] 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 Transaction on Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[20] 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*, pp. 3247–3250, 2007.

[21] M. Sharma, S. Joshi, T. Chatterjee, and R. Hamid, "A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows," *Neurocomputing*, vol. 494, pp. 116–131, 2022.

[22] S. E. Tranter and D. A. Reynolds, “An overview of automatic speaker diarization systems,” *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[23] 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,” *in Proceedings of International Conference on Language Resources and Evaluation*, pp. 965–968, 2000.

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

[25] 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, pp. 1–4, 2019.

[26] D. Kitamura and N. Ono, “Efficient initialization for nonnegative matrix factorization based on nonnegative independent component analysis,” *in Proceedings of International Workshop on Acoustic Signal Enhancement*, 2016.

[27] J. Wang, S. Guan, J. Chen, and J. Benesty, “Independent low-rank matrix analysis based on the Sinkhorn divergence source model for blind source separation,” *arXiv Preprint*, 2401.01762, 2024.