# 連想記憶を用いた
# 線形ブラインド音源分離の研究

# A Study on Linear Blind Source Separation using Associative Memory Model

2017年7月

大町 基
Motoi OMACHI

# 連想記憶を用いた
# 線形ブラインド音源分離の研究

# A Study on Linear Blind Source Separation using Associative Memory Model

2017年7月

早稲田大学大学院 基幹理工学研究科
情報理工学専攻　知覚情報システム研究

大 町　基
Motoi OMACHI

# 目 次

# 表 目 次

# 図 目 次

# 第1章　序論

## 1.1　背景

音源分離は，複数の音源から発せられた音が混じりあった信号から元の音を再現する技術であり，音声・音響信号を用いたパターン認識・合成を行う上で重要な役割を担っている．例えば，マイクロホンで観測された信号から，実環境で起きている音響イベントを認識するためには，音源分離が必要不可欠な一要素を担っている [1]．また，実環境において音声認識を行う場合には，認識対象話者の音声信号とそれ以外の音源から発せられた音とを分離することで，性能を改善することができる [2]．他には，音源分離を用いることで，音源ごとに音像の定位を制御することが可能となる [3]．音源分離をパターン認識の前処理として用いるときには，目的音源から発せられた音のみを高精度に分離する技術が求められる．さらに遠隔会議システムや収音システムなど，聴取を目的とした場面に応用する場合では，分離音を聴取したときに違和感を与えないことも重要な課題となる．

我々を取り巻く環境ではあらゆる音源から音が発せられており，人間はそれらが同時に発音された状態でも，任意の音を聞き分けることができる [4]．機械にも同様の機能を持たせるために，ビームフォーミングや適応ビームフォーミングなどの空間フィルタリングが広く用いられている．ビームフォーミングは，ビームや死角（Null）などの指向特性を形成することにより，任意の方向から到来する信号成分を強調・抑圧する．代表的なものとして，遅延和（Delay and sum: DS）ビームフォーマ [5] や死角制御型ビームフォーマ（Null Beamformer: NBF） [6] などが提案されている．また，適用ビームフォーミングは，周囲の雑音環境に応じて指向特性を制御することで，目的の音源方向の音を強調する．最尤（Maximum likelihood: ML）ビームフォーマ [7]，最小分散（Minimum variance distortionless responce: MVDR）ビームフォーマ [8]，一般化サイドローブキャンセラ（Generalized sidelobe canceller: GSC） [9] などがこの手法にあたる．しかし，これらの技術を用いるためには，音源の方向や雑音空間相関行列などの収音環境に関する事前情報が必要となる．実際の利用場面においては，それらの情報が得られないという場合が起こる．そこで，収音環境に関する事前情報が不要なブラインド音源分離（Blind Source Separation: BSS） [10] の研

究が盛んに行われるようになった．近年でも SiSEC(Signal Separation Evaluation Campaign) [11] といったコンペティションが開催されるなど，活発に研究が行われている分野である．

現在提案されている BSS 手法は，1 本のマイクロホンを用いるシングルチャンネル BSS と，2 本以上のマイクロホンを用いるマルチチャンネル BSS とに大別することができる．シングルチャネル BSS は，小規模なシステムへの応用が可能であるが，1 本のマイクロホンでの観測信号から得られる情報のみから音源を分離することは難しく，分離性能に関してはまだ改善の余地がある．一方マルチチャンネル BSS では，各マイクロホンにおける観測信号から得られる情報に加えて，空間的な情報も扱うことが可能なため，高精度な音源分離の実現が期待できる．マルチチャンネル BSS 手法はさらに，非線形 BSS と線形 BSS とに大別できる．非線形 BSS は，目的音源以外の成分を抑圧する分離性能は高いが，ミュージカルノイズといった非線形歪が発生する．一方線形 BSS は，非線形歪が原理的に発生しないという優れた特徴があるものの，条件によっては十分な分離性能が得られない．すなわち，現存の BSS では分離性能と分離音の音質との間にトレードオフの関係が存在する．応用先に縛られることなく、汎用的な場面で BSS を応用するためには，分離性能と分離音の音質が共に高い技術の開発が望まれる．この問題を解決するための方法として，複数の音源分離手法を組み合わせたタンデム接続型音源分離が提案されているが，どのように複数の音源分離手法を組み合わせるのかが課題となる．

## 1.2 既存のブラインド音源分離（BSS）手法

前述の通りに，既存の BSS は非線形 BSS と線形 BSS とに大別できる．ここでは，それぞれの手法の概要を説明するとともに，それらを組み合わせたタンデム接続型音源分離についても概観する．

### 1.2.1 非線形 BSS

非線形 BSS として，時間周波数マスクに基づく手法 [12, 13, 14]，Denoising autoencoder（DAE）を用いた手法 [15, 16] が挙げられる．

観測信号に対して適切な時間長で短時間フーリエ変換（Short time Fourier transform: STFT）を適用すると，音源に主要な成分は時間周波数空間上において疎に存在することが知られている [17]. この性質に基づき，時間周波数マスクに基づく手法では，目的音源成分が支配的に存在する時間周波数ビンのみを通過させるバイナリマスク [17, 18, 19] を用いて，不要な成分を効果的に取り除く．しかし，ミュージカルノイズなどの非線形歪が発生するため，分離音を聴取した際に違和感

を与える．ミュージカルノイズの影響は，ソフトマスク [20, 12, 21] を用いることで緩和できることが期待できるが，分離性能が劣化するという問題がある．ミュージカルノイズの影響を低減する他の試みとして，時間周波数マスクや分離信号をケプストラム領域にて平滑化する方法が提案されているが，残響感といった聴感上の違和感を与えてしまうことが報告されている [22, 23]．時間周波数マスクの設計は，各時間周波数ビンにおいて音声成分が含まれるかどうかの識別問題と考えることができる．そこで近年では，識別器として Deep neural network(DNN） [24] を用いることで，時間周波数マスクを高精度に推定する試みがなされている [25, 26, 27]．これらの方法では，学習データと入力データとのミスマッチに対する頑健性をどう担保するのかが課題となる．

DAE [28] は，ノイズが含まれるパターンから元のパターンを再現するニューラルネットワークであり，雑音や残響環境下の音声認識において効果的であることが報告されている [29, 30, 31, 32, 33]．Xu らは，雑音を重畳したスペクトルから元のスペクトルを再現する音声強調の観点でも DAE が有効であること示した [34]．Tu らは，DAE により観測信号から特定の目的話者音声と任意の妨害話者音声への写像関数を実現し，目的話者音声と任意話者音声を分離する方法を提案した [15]．この方法の問題点として，学習データと入力とのミスマッチが生じた場合における性能劣化があげられる．この問題を解消するために SNR 毎に DAE を設計し，入力時の SNR を推定した上で DAE を選択するという枠組みが提案された [16]．しかし，DAE により出力されるスペクトルは統計処理に起因する平滑化の影響を受ける [35]．近年では，DAE を時間周波数マスク [36]，Wiener filter [37]，非負値行列分解 (Non-negative matrix factorization: NMF) に基づく方法 [38] などの音声強調または音源分離手法の一部に組み込む方法が提案されている．

### 1.2.2 線形 BSS

無響環境において，複数の音源から発せられた信号 $\mathbf{s}(t)$ が同時にマイクロホンに到達するならば，観測信号 $\mathbf{z}(t)$ は混合行列 $\mathbf{A}$ を用いた線形変換 $\mathbf{z}(t) = \mathbf{A}\mathbf{s}(t)$ で表現することができる．線形 BSS は，$\mathbf{A}$ の影響を打ち消す線形分離行列 $\mathbf{W}$ を用いた線形変換 $\mathbf{W}\mathbf{z}(t)$ によって，音源信号 $\mathbf{s}(t)$ を再現する．代表的な手法として，独立成分分析（Independent component analysis: ICA） [39]，独立ベクトル分析（Independent vector analysis: IVA） [40] が広く用いられている．ICA は，音源信号 $\mathbf{s}(t)$ が互いに独立であるという音源の独立性の仮定を用いる．この仮定に基づき，分離信号 $\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t)$ が互いに独立となるような線形分離行列 $\mathbf{W}$ を推定し，$\mathbf{s}(t)$ を再現する [39]．線形変換されたものを線形変換で戻すという枠組みであるため，非線形歪が原理的に発生しないという利点がある．しかし実際の環境では，伝達系の影響は畳み込みにより表現されるため，瞬時混合の

仮定をそのまま用いることはできない．そこで，ICA を畳み込み問題に拡張した上で $\mathbf{W}$ を推定する方法 [41, 42]，周波数領域に変換することで瞬時混合に近似した上で $\mathbf{W}$ を推定する方法 [39, 43] が提案されている．前者の方法は収束点に近い初期値を与えることができれば高い性能が保証される．しかし，実環境における分離問題では求めるパラメータ数が多くなり，それらを同時に最適化するためには計算コストが高くなる．後者は，少ないパラメータの最適化を行うことが可能となるため計算コストが低いという点で前者よりも優れる．一方で，パーミュテーション問題やスケーリング問題とよばれる問題が発生する．これらの問題は，分離性能や音質劣化を引き起こす要因となるため，これらの問題を解決するための方法が，いくつか提案されている [44, 45, 46, 47]．近年では，パーミュテーション問題が発生しない枠組みとして，ICA を拡張させた IVA が提案されている [40]．IVA は分離信号のスペクトルが，互いに独立するベクトルとなるように分離行列を推定する．ベクトルの要素間の相関を考慮する多次元分布を用いて独立性を評価することにより，パーミュテーション問題の発生を防ぐことができる．しかし，依然としてスケーリング問題の発生を防ぐことは難しい．また，ICA や IVA において独立性を評価するためには，音源信号が生成される確率分布を仮定する必要があり，その分布と実際のデータの分布との間にミスマッチが生じると分離性能が劣化する．この問題に対応するため，音源の性質を表した分布を用いる手法 [48, 49, 50] が提案されているが，分布に基づく枠組みでは，音声のスペクトルにおける調波構造などの音源らしさを直接的に推定することは難しい．さらに，独立性の仮定では条件によっては十分な性能が得られないことがある．例えば，強い反射音が観測信号に混入する場合には反射音もまた独立な音とみなすことができるため，独立性のみで音源を分離することが難しくなることが予想される．

### 1.2.3 タンデム接続型音源分離

タンデム接続型音源分離は，複数の音源分離手法を組み合わせることにより個々の音源分離の欠点を補うために用いられる．

線形 BSS の分離性能を改善するための枠組みとして，ICA の後段に時間周波数マスクに基づく手法を組み合わせることで，ICA によって取り除くことができなかった信号を抑圧する方法 [51, 52] や，ICA により形成される指向特性を用いて音源のスペクトルを推定し，スペクトル減算法により音源分離を行う方法 [53] などが提案されている．しかしこれらの手法では，非線形処理が含まれているため非線形歪の発生を防ぐことは難しい．

非線形 BSS によって生じる非線形歪の影響を低減する枠組みとして，時間周波数マスクの後段

に NMF に基づく方法を適用することにより音質を改善する試み [54] が提案されている．この方法では NMF で用いる音声辞書の構築に用いる音声をどのように選択するかが重要な課題となる．他には，時間周波数マスクの出力に対して ICA を適用することにより，時間周波数マスクによって生じる非線形歪の影響が低減されるとの報告もある [55]．近年では，ビームフォーミングのパラメータを推定するために非線形 BSS である時間周波数マスクを組み合わせることで，分離性能を改善する枠組みがいくつか提案されている [56, 57]．これらの枠組みでは，ビームフォーミングに必要となるステアリングベクトルや PSD 行列の推定にのみ時間周波数マスクを用いており，非線形歪は原理的に発生しない枠組みとなっている．しかし，ビームフォーミングは空間的な情報に基づき分離を行うものであるため，音源に関する情報は陽には扱われていない．

## 1.3 本論文の目的

本論文では分離性能，分離音の自然性が共に高い BSS を提案する．ここでは，分離性能が高いとは妨害音に対する抑圧性能が高いことを，自然性が高いとは歪の発生が少ないことと等価の意味として定義する．

図 1.1 に既存の BSS と提案法の関係を示す．BSS の性能を分離性能と分離音の自然性で評価するとき，理想的な BSS は両軸において高いことが望ましい．1.2.1 節で説明したように，非線形 BSS の一手法である時間周波数マスクに基づく手法は，分離性能は高いが自然性は低い．また，1.2.2 節で説明したように，線形 BSS の一手法である IVA は，非線形歪が原理的に発生しないため時間周波数マスクよりも自然性は高いが，収音環境によっては十分な分離性能が得られないことがある．

本研究では，聴感上違和感を与える歪が少なく，高精度に音源を再現する BSS の実現方法として，2 つのアプローチを検討した．1 つ目のアプローチでは，連想記憶を用いて，音源である音声らしさを考慮しながら線形分離行列を推定する．具体的には，事前に音声のスペクトルを学習させた連想記憶モデルを用いて分離信号から音声信号（以降，参照信号とよぶ）を推定する処理と，参照信号と分離信号との誤差が最小となるように線形分離行列を補正する処理を繰り返す方法を提案する．ここでは，分離信号のスペクトルには妨害音の消し残しや分離により生じる歪などのノイズが含まれると仮定し，連想記憶を用いてそれらのノイズを取り除くことにより参照信号を推定する．すなわち，連想記憶モデルは，DAE の考え方を参考に構築する．1.2.1 で述べたように，DAE を音源分離手法に組み込む試みは既に行われているが，提案法は線形分離行列の推定に用いるという点でそれらとは異なる．提案法は，独立性の仮定が不要であるだけでなく，音源が

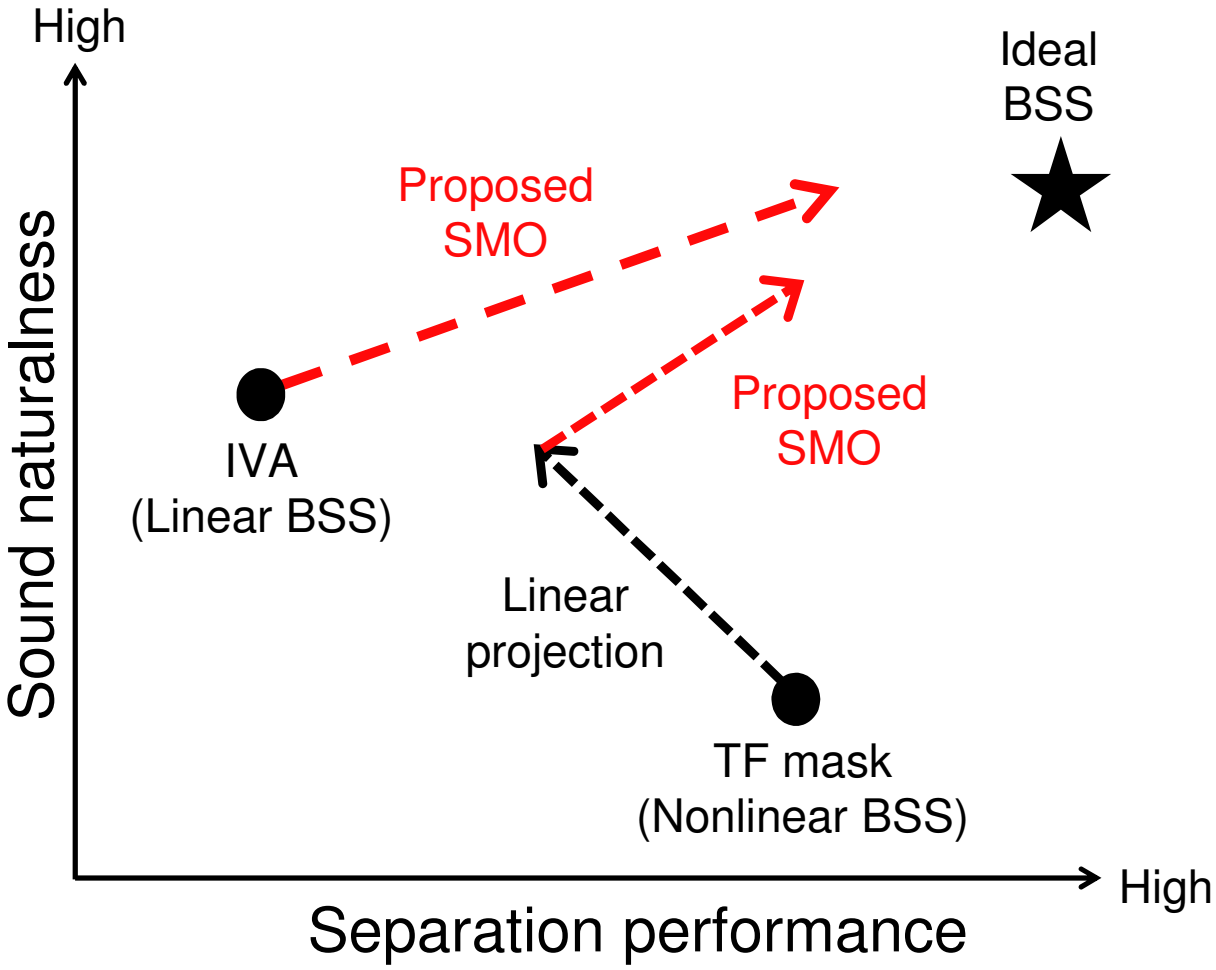



図 1.1: Relation between existing and proposed blind source separation (BSS). TF and SMO denote time-frequency and separation matrix optimization, respectively.

生成される分布を仮定した枠組みでは扱うことが難しかったスペクトルの微細構造などを考慮するため，IVA よりも高精度な分離が可能であると期待できる．

2 つ目のアプローチは，連想記憶を用いた線形分離行列推定法と非線形 BSS を組み合わせたタンデム接続型音源分離を提案する．まず，観測信号に対して非線形 BSS である時間周波数マスクを適用することにより，分離行列の推定に不要な成分を取り除く．これにより，分離行列の推定が容易になることを期待できる．しかし，時間周波数マスクを適用することにより非線形歪が発生してしまう．そこで，時間周波数マスクをそのまま用いるのではなく，時間周波数マスクの入出力から線形射影行列を求め，分離行列の初期値として用いる．さらに，連想記憶を用いて音源らしさを考慮しながら音源分離を行う．で述べたように，時間周波数マスクをビームフォーミングのパラメータの推定に用いる方法は既に提案されているが，提案法は音源らしさを陽に扱っているという点でそれらとは異なる．

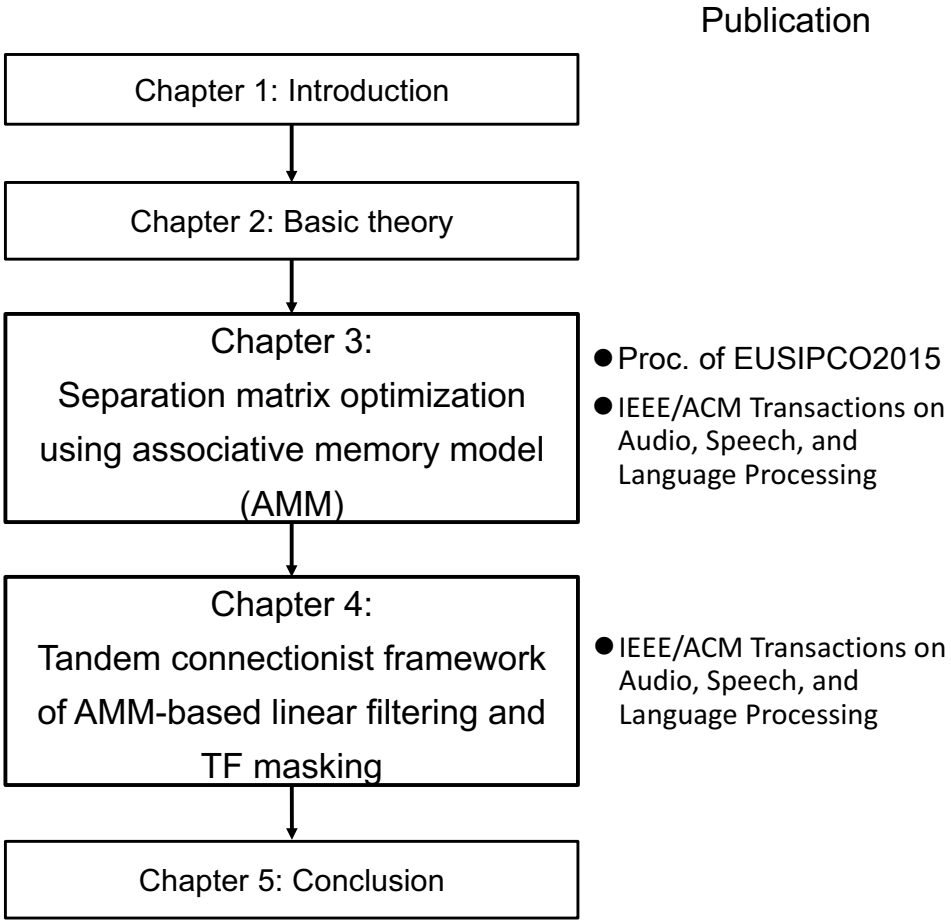



図 1.2: Structure of this study.

## 1.4 本論文の構成

本論文の構成は以下の通りである（図 1.2）.

第 1 章では，本研究の目的および従来研究との立場の違いについて述べるとともに，従来の BSS についても概観した.

第 2 章では，本研究を説明する上で重要となる基盤技術について概観する．まず，ブラインド音源分離処理の基本的な流れを説明し，続いて提案法で用いる時間周波数マスクを用いた BSS および線形分離行列を用いた BSS について概観する．特に線形分離行列を用いた BSS では，ICA と IVA の違いについても言及する．ICA では，周波数ビン毎に分離信号間の独立性を評価するために，パーミュテーション問題やスケーリング問題といった問題が発生する．一方 IVA では，周波数ビン間の関係を考慮しながら独立性を評価するためにパーミュテーション問題が原理的に発生しないという利点がある．ここではさらに，連想記憶モデルとして，Restricted Boltzmann machine

（RBM），Autoencoder（AE），Denoising autoencoder（DAE），Convolutional autoencoder（CAE），Convolutional neural network（CNN）についても概観する．

第 3 章では，提案法のアルゴリズムについて述べる [58, 59]．提案法は，連想記憶モデルを用いて参照信号を推定する処理と，参照信号と分離信号との誤差が最小となるように線形分離行列を補正する処理で構成される．ここでは，それぞれの方法について述べる．特に，参照信号の推定法では，連想記憶モデルとして，DAE，CNN，Denoising convolutional autoencoder（DCAE）について説明する．DAE は，音声スペクトルに含まれるノイズを取り除くのに効果的であることが知られており，雑音除去や残響抑圧などに応用されている．しかし，ピッチの時系列変化などの音声スペクトルの局所的な構造は陽に扱われていない．一方，CNN は，音声のスペクトルを局所的なパターンの組み合わせとして考えるため，局所的に存在するノイズに対して頑健であることが期待できる．また，Max-pooling [60] の処理を行うことにより，フィルタを効率的に学習可能であるという特長がある．しかし，Max-pooling により，フィルタの適用位置の情報が失われる．そのため，復元されたスペトルには元音声らしさが反映されない可能性がある．本研究では，この問題を解決する枠組みとして，DCAE を開発した．DCAE は，CNN と同様に音声のスペクトルを局所的なパターンの組み合わせとして考えるとともに，Max pooling の構造を有する．一方で，プーリング処理により失われる情報を陽に扱いながら音声を推定する．そのため，元音声らしさを考慮しながら音声スペクトルを推定する枠組みであるといえる．評価実験では，二話者同時発話音声の分離およびエコー除去の実験を行う．二話者同時発話音声の分離実験では，IVA と提案法とで分離性能を比較し，音声らしさを考慮することの効果を検証する．ここでは，上述の DAE，CNN，DCAE それぞれを連想記憶として用いた場合の性能を比較し，提案法に用いる連想記憶モデルとして最適な構造を検討する．エコー除去の実験では，独立性の仮定が成立しにくい状況における提案法の有効性を検証する．IVA では，音源の独立性を仮定しているため，エコー信号のように音源に対して独立とは言い難い信号が含まれるとき，分離性能が劣化する．一方，提案法は音源の独立性の仮定が不要であるため，そのような状況においても分離可能であることが期待できる．

第 4 章では，AMM を用いた線形分離行列推定法の前段に時間周波数マスクを組み合わせたタンデム接続型音源分離の枠組みを提案した．時間周波数マスクを用いて，線形分離行列の推定に不要な成分を抑圧することで，後続する線形分離行列の最適化がしやすくなることを期待している．しかし，時間周波数マスクの出力に対して分離行列を適用すると，分離信号は非線形歪の影響を受ける．そこで，時間周波数マスクをそのまま用いるのではなく，時間周波数マスクの入力

から出力への射影行列を求め，その値を提案法における分離行列の初期値として用いる方法を提案した．評価実験では，二話者同時発話音声に対する分離性能を調査し，前段に周波数マスクを組み込むことにより，提案法の分離性能が改善し，歪の発生量を低減することができることを確認した．さらに，連続音素認識による評価を行い，提案するタンデム接続型音源分離が最も高い性能を示すことも確認した．

第 4 章では，提案する線形分離行列推定法と時間周波数マスクに基づく BSS を組み合わせたタンデム接続型音源分離を提案し，有効性を検証した結果について報告する [59]．時間周波数マスクを用いて，線形分離行列の推定に不要な成分を事前に抑圧することで，後続する線形分離行列の最適化がしやすくなることが期待できる．しかし，時間周波数マスクの出力に対して分離行列を適用すると，分離信号は非線形歪の影響を受ける．そこで，時間周波数マスクをそのまま用いるのではなく，時間周波数マスクの入出力関係を表す射影行列を求め，その値を提案法における分離行列の初期値として用いる方法を検討する．評価実験では，二話者同時発話音声に対する分離性能を調査し，前段に周波数マスクを組み込むことにより，提案法の分離性能が改善し，歪の発生量を低減することができることを確認する．さらに，連続音素認識による評価を行い，提案するタンデム接続型音源分離が音声認識にも有効であることを検証する．

第 5 章では，本研究のまとめと今後の展望について述べる．

# 第2章　基礎技術

本章では本研究を説明する上で重要となる基盤技術について概観する．まず最初に，音信号処理の基本的な流れを説明する．BSS は，時間波形に対して行う方法と，時間波形を時間周波数表現に変換した上で行う方法とが存在する．本研究では，計算量が少ないこと，学習が収束しやすいという観点で後者の方法を採用している．そこでまずは，時間波形から時間周波数表現に変換し，時間波形を合成する方法について説明する．次に，BSS の基礎技術として，線形分離行列を用いた方法 [40]，および，時間周波数マスクを用いた方法 [12] について概説する．時間周波数マスクは，非線形 BSS において広く用いられており，目的音以外の成分の抑圧性能に優れているが，非線形歪が発生するという問題がある．一方線形分離行列を用いた手法は，非線形歪が原理的に発生しないという優れた特長がある．最後に，連想記憶モデルとして，Restricted boltzman machine，Autoencoder，Denoising autoencoder，Convolutional autoencoder，Convolutional neural network を紹介する．

## 2.1　音信号処理の流れ

図 2.1 に，時間周波数表現を用いた音信号処理の基本的な流れを示す．

まず，AD 変換によりデジタル化された観測信号の系列 $\mathbf{x} = \{x[i]|i = 0, \cdots, N_t - 1\}$ に対し，短時間フーリエ変換（Short time Fourier transform: STFT）を適用することで，時間周波数表現であるスペクトルの時系列 $\mathbf{X} = \{X[k, l]|k = 0, \cdots, N_\omega - 1; l = 0, \cdots, N_\tau - 1\}$ を計算する．ここで $N_t$，$N_\omega$，$N_\tau$ はそれぞれ，観測信号，離散周波数，離散フレームの総数を表す．次に，$\mathbf{X}$ に対して時間周波数マスク処理や線形変換などのスペクトルの加工を行い，目的のスペクトルの時系列 $\mathbf{X}'$ を計算する．最後に，$\mathbf{X}'$ に対して，逆短時間フーリエ変換（Inverse STFT: ISTFT）を適用することで，加工された時間波形 $\mathbf{x}'$ を得る．

時間領域から周波数領域への変換は，フーリエ変換（Fourier Transform: FT）が広く用いられる．これは周期信号 $x(t)$ を，無限個の周波数 $\omega$ の複素正弦波で表現される基底空間に射影するものであり，以式のように定義される．

$$X(\omega) \quad = \quad \int_{-\infty}^{\infty} x(t) \exp\left(-j2\pi\omega t\right) dt \ , \tag{2.1}$$

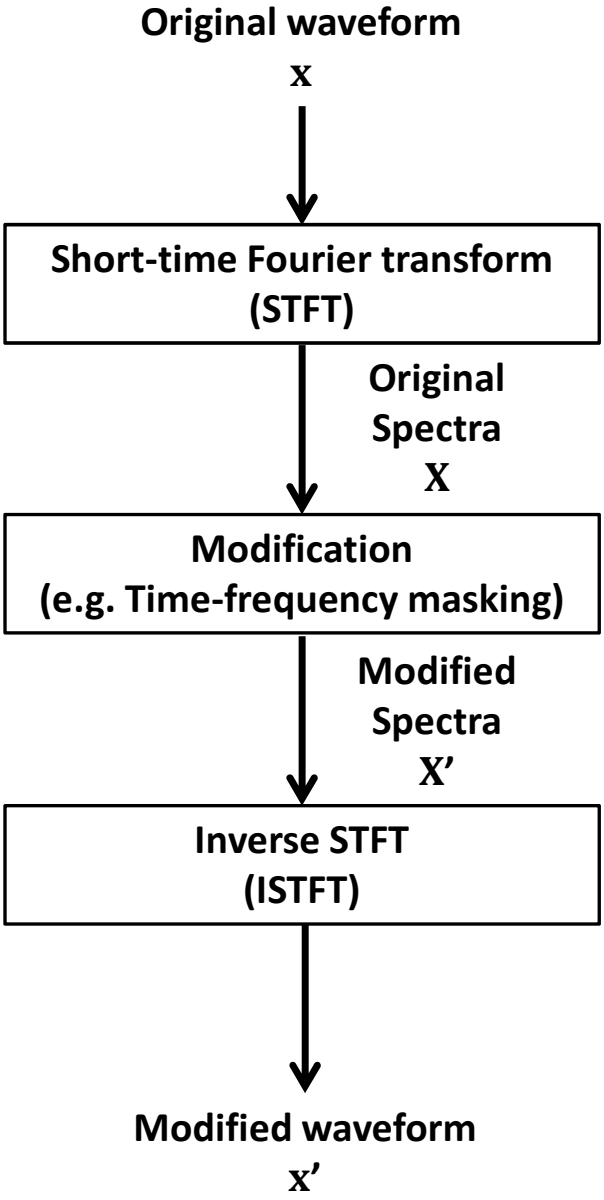



図 2.1: Procedure of sound signal processing via frequency domain.

$$|X(\omega)| = \sqrt{|\mathrm{Im}[X(\omega)]|^2 + |\mathrm{Re}[X(\omega)]|^2}\ . \tag{2.2}$$

$$\angle X(\omega) = \arctan\left(\mathrm{Im}[X(\omega)]/\mathrm{Re}[X(\omega)]\right)\ , \tag{2.3}$$

$$|X(\omega)|^2 = X(\omega)X^*(\omega). \tag{2.4}$$

ここで，$\mathrm{Re}[X(\omega)]$ および $\mathrm{Im}[X(\omega)]$ はそれぞれ，$X(\omega)$ の実部および虚部を表す．また，$^*$ は複素共役を表す．$X(\omega)$ はスペクトルと呼ばれ，特に，$|X(\omega)|$，$\angle X(\omega)$，および，$|X(\omega)|^2$ はそれぞれ振幅スペクトル，位相スペクトル，パワースペクトルと呼ばれる．周波数領域から時間領域に戻す場合は，以下に示すような逆フーリエ変換（Inverse Fourier transform: IFT）を $X(\omega)$ に適用

する．

$$x(t) \quad = \quad \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)\exp\left(j2\pi\omega t\right)d\omega \ . \tag{2.5}$$

FT および IFT は，連続時間信号において時間領域-周波数領域間の変換を行う．しかしコンピュータで音を処理する際には，$x(t)$ は，サンプリング周期（サンプリング周波数の逆数）$1/f_s$ 秒間隔でサンプルされたデジタル信号系列 $\mathbf{x} = \{x[i]|i = 0, \cdots, N_t - 1\}$ として扱われる．離散化されたデータ系列を周波数領域に変換する際には，以下に示すような離散フーリエ変換（Discrete Fourier transform: DFT）が用いられる．

$$X[k] \quad = \quad \sum_{i=0}^{N_t-1} x[i]\exp\left[-\frac{j2\pi ki}{N_t}\right] \tag{2.6}$$

また，周波数領域から時間領域に戻す際には，以下のような逆離散フーリエ変換（Inverse DFT: IDFT）を用いる．

$$x[i] \quad = \quad \frac{1}{N_\omega}\sum_{k=0}^{N_\omega-1} X[k]\exp\left[\frac{j2\pi ki}{N_t}\right] \ . \tag{2.7}$$

実際に演算を行う場合には，DFT や IDFT を高速化した高速フーリエ変換（Fast Fourier Transform: FFT）や逆高速フーリエ変換（Inverse FFT: IFFT）が広く用いられている．

FFT や IFFT では，解析対象のデータ系列を周期的な定常信号であることを前提としている．しかし音声などの音響信号は，時刻によって周波数成分が変化する非定常信号であることが多い．そこで図 2.2 に示すように，$N_d$ サンプル分のデータを $d$ サンプル間隔で切り出しながら，FFT を適用する STFT が用いられる．このとき，切り出した波形の単位をフレーム，$N_d$，$d$ をそれぞれフレーム長，フレームシフトとよぶ．$l$ 番目のフレームにおけるスペクトルは，

$$X[k,l] \quad = \quad \sum_{n=ld}^{ld+N_d-1} w^{(a)}[n-ld]x[n]\exp\left[-\frac{j2\pi k(n-ld)}{N_d}\right] , \tag{2.8}$$

$$w_a[i] \quad = \quad 0.5 - 0.5\cos\left[\frac{2\pi i}{N_d}\right] \ . \tag{2.9}$$

と計算することができる．ここで $w^{(a)}[n]$ は分析窓と呼ばれるもので，切り出した波形の中心が 1，両端が 0 となるように重みづけるものである．フーリエ変換は，切り出したフレームを，同一フレームが無限に繰り返された $[-\infty,\infty]$ の周期信号とみなす．分析窓をかけずにフレームを切り出した場合には，両端で不連続点が生じる可能性があり，そのような場合においては，周波数領域において不要な成分が発生する．時間窓はフレーム両端の不連続点を解消することで，そのような影響を緩和するという効果がある．

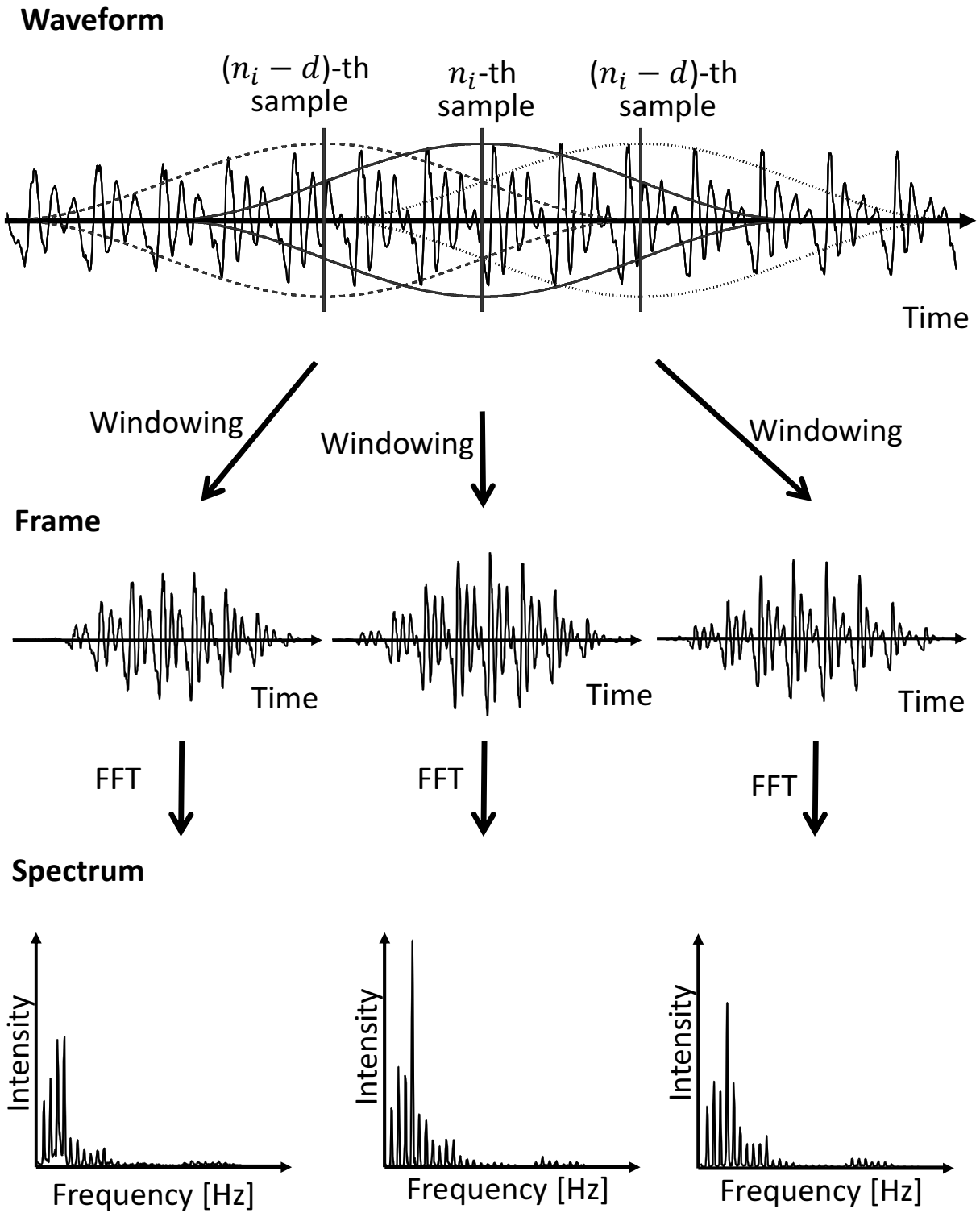



図 2.2: Short time Fourier transform.

音信号処理では，こうして得られたスペクトルの時系列に対して，線形変換や時間周波数マスキングなどを適用することで，スペクトルを加工する．線形変換は複素スペクトル $X[k]$ を加工す

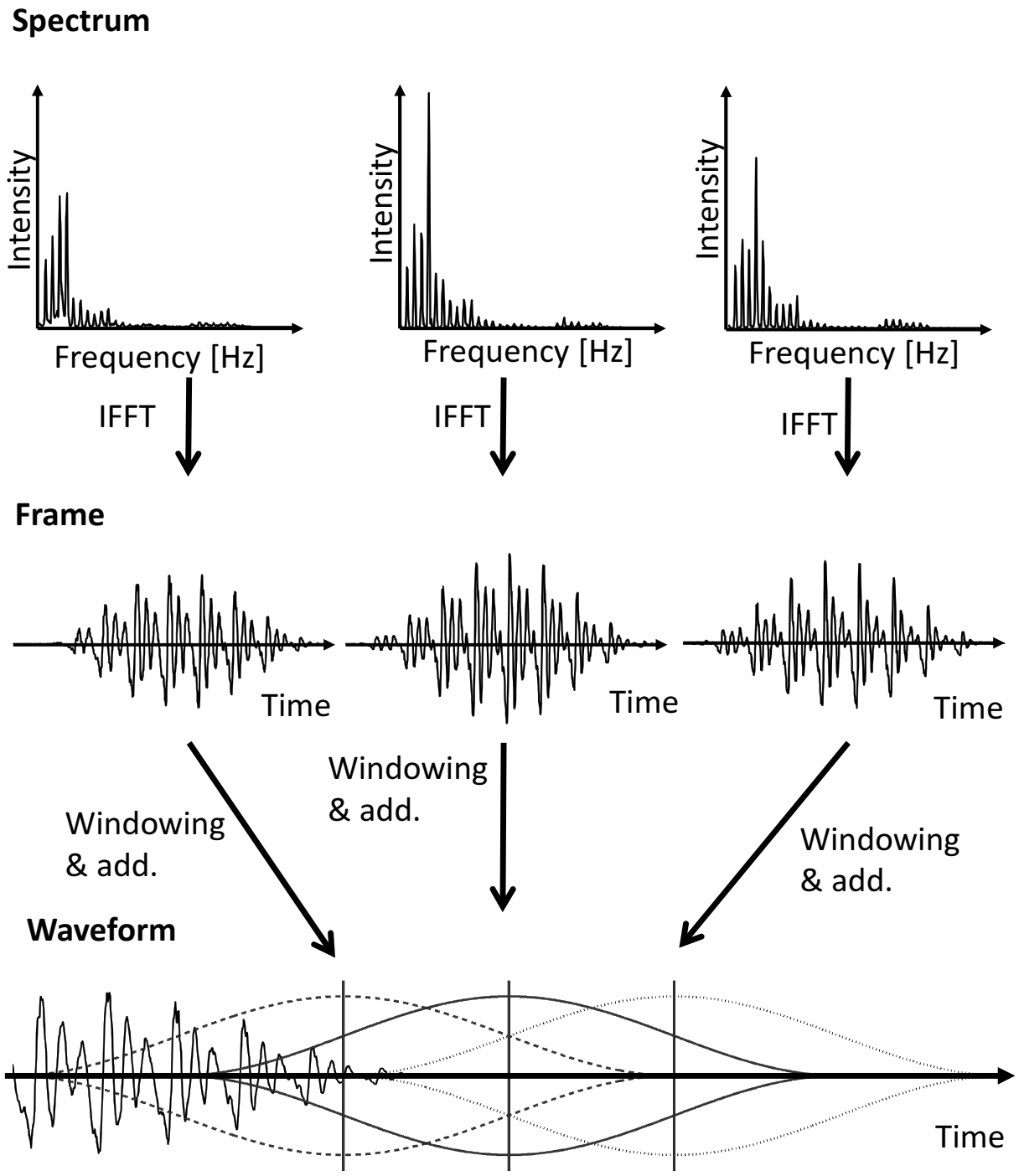



図 2.3: Inverse STFT.

るのに対し，時間周波数マスキングでは，振幅スペクトル $|X[k]|$ のみを加工し，位相スペクトル $\angle X[k]$ は元信号のものをそのまま用いる．これは，人間は位相スペクトルよりも振幅スペクトルの違いに敏感であるということに起因する．加工した振幅スペクトルに対して整合性のある位相スペトルを得たいならば，信号再構築 [61] を用いてもよい．

加工されたスペクトルから，時間波形 $\mathbf{x}'$ を得るには，図 2.3 に示すような ISTFT を各フレームのスペクトルに適用する．ISTFT ではまず，各フレームにおけるスペクトル $\mathbf{X}'[l]$ に IFFT を適用することで，波形 $\mathbf{x}'[i]$ を計算する．このとき，フレーム両端の連続性を保証するために再び時間窓 $w_s[i]$ を乗じる．そして時間窓を乗じた波形 $w_a[i]\mathbf{x}'[i]$ を，分析時のサンプル位置に足し合わせていく，波形重畳法（Overlap and add: OAA）と呼ばれる処理によって加工した時間波形 $\mathbf{x}'$ を合成する．$N_d$ および $d$ は，アプリケーションによって異なる値が用いられるが，加工したスペクトルを波形に戻す際には，$d$ を $N_d/4$ 以下にするとよい [62]．

## 2.2 ブラインド音源分離

ここでは，$N_m$ 個の観測信号から $N_s$ 個の音源（$N_s \leq N_m$）を再現する問題を例に，線形分離行列に基づく手法および，時間周波数マスクに基づく手法について概観する．

### 2.2.1 線形分離行列に基づく手法

線形分離行列に基づく音源分離の目的は，音源方向などの事前情報が未知という制約の下で，混合過程の影響を取り除く線形フィルタを推定することである．

$n$ 番目の音源信号を $\mathbf{s}_n = \{s_n[i]|i = 0, \cdots, N_t\}$，$m$ 番目のマイクロホンにおける観測信号を $\mathbf{z}_m = \{z_m[i]|i = 0, \cdots, N_t + N_h\}$，$n$ 番目の音源から $m$ 番目のマイクロホンまでのインパルス応答を $\mathbf{h}_{mn} = \{h_{mn}[i]|i = 0, \cdots, N_h\}$ とすると，時間領域における混合過程は以下のように書くことができる．

$$z_m[i] = \sum_{n=0}^{N_s-1} \sum_{d=0}^{N_h-1} h_{mn}[i] s_n[i - d + 1] \tag{2.10}$$

ここで，$N_t$，$N_h$ は観測信号のデータ数およびインパルス応答長を表す．時間領域において逆フィルタを求めるとき，初期値が最適解の近傍であるならば収束性が保証されるが，残響などが含まれる場合には，最適化が難しくなる．また，畳み込み演算が必要なため計算量が多くなる．

そこで，$\mathbf{z}_m$ を $N_h$ よりも十分に長い分析長で STFT し，混合過程を以下のような瞬時混合で近似する．

$$Z_m[k, l] = \sum_{n=1}^{N_s} H_{mn}[k] S_n[k, l] \tag{2.11}$$

$k$，$l$ は離散周波数およびフレームを表す．また，$Z_m[k, l]$，$S_n[k, l]$ はそれぞれ観測信号のスペクトルと音源のスペクトルを，$H_{mn}[k]$ は伝達関数を表す．すなわち，周波数領域に変換すると畳み込み演算を行う必要がなくなり，計算量が削減できる．また，インパルス応答長の影響を考慮する必要がなくなるため，フィルタの推定が容易となる．ここで，観測信号のスペクトルを束ねたベクトルを $\mathbf{Z}[k, l] = [Z_1[k, l], \cdots, Z_{N_m}[k, l]]^{\mathrm{T}}$ （$^{\mathrm{T}}$ は転置を表す），混合行列を $\mathbf{H}[k]$ とすると，音源信号のスペクトルを束ねたベクトル $\mathbf{S}[k, l] = [S_1[k, l], \cdots, S_{N_s}[k, l]]^{\mathrm{T}}$ は，以下のような線形変換で表現することができる．

$$\mathbf{Z}[k, l] = \mathbf{H}[k]\mathbf{S}[k, l] \tag{2.12}$$

$$\mathbf{H}[k] = \begin{bmatrix} H_{11}[k] & \cdots & H_{1N_s}[k] \\ \vdots & \ddots & \vdots \\ H_{N_m1}[k] & \cdots & H_{N_mN_s}[k] \end{bmatrix} \tag{2.13}$$

線形分離行列に基づく BSS では，$\mathbf{Z}[k,l]$ に対して以下のような線形分離行列 $\mathbf{W}[k]$ による線形変換を適用することにより，出力信号のスペクトルを束ねたベクトル $\mathbf{Y}[k,l]$=$[Y_1[k,l], \cdots, Y_{N_s}[k,l]]^{\mathrm{T}}$ を求める．このとき，$Y_n[k,l]$ は推定された $n$ 番目の音源のスペクトルを表す．

$$\mathbf{Y}[k,l] = \mathbf{W}[k]\mathbf{Z}[k,l] = \mathbf{W}[k]\mathbf{H}[k]\mathbf{S}[k,l] \tag{2.14}$$

$$\mathbf{W}[k] = \begin{bmatrix} W_{11}[k] & \cdots & W_{1N_m}[k] \\ \vdots & \ddots & \vdots \\ W_{N_s1}[k] & \cdots & W_{N_sN_m}[k] \end{bmatrix} \tag{2.15}$$

式 (2.14) において，$\mathbf{W}[k] = \mathbf{H}^{-1}[k]$ であるならば，音源信号を完全に再現することができる．すなわち $\mathbf{H}[k]$ が既知であるならば，逆フィルタを容易に求めることができる．しかし，$\mathbf{H}[k]$ はマイクロホンと音源の位置関係や収音環境に依存する．そのため，$\mathbf{H}[k]$ を事前情報として持つことは実用上難しい．

そこで，$\mathbf{W}[k]$ の推定には，音源がそれぞれ統計的に独立であるという仮定が広く用いられる．$\mathbf{W}[k] = \mathbf{H}^{-1}[k]$ であるならば，分離された信号もまた音源と同じ分布から生成されたものと考えることができる．すなわち，分離された信号が互いに独立となるような $\mathbf{W}[k]$ を求めればよい．この考えに基づき，ICA や IVA が広く用いられている．

ICA では，出力信号のスペクトル成分 $Y_n[k,l]$ が互いに独立となるような $\mathbf{W}[k]$ を周波数ごとに推定する．具体的には，以下の式を満たす $\hat{\mathbf{W}}[k]$ を推定する．

$$\begin{aligned} \hat{\mathbf{W}}[k] &= \underset{\mathbf{W}[k]}{\operatorname{argmin}} KL\left(p(Y_1[k,l], \cdots, Y_{N_s}[k,l]) || \Pi_{n=1}^{N_s} p(Y_n[k,l])\right) \\ &= \underset{\mathbf{W}[k]}{\operatorname{argmin}} \; -\log|det(\mathbf{W}[k])| - \sum_{n=1}^{N_s} E[\log p(Y_n[k,l])] \\ &\qquad + Const. \end{aligned} \tag{2.16}$$

ここで $p(\cdot)$，$det(\cdot)$ および $KL(p||q)$ はそれぞれ音源の事前分布，行列式を求める関数および Kullback-Libler 情報量を表す．$KL(p||q)$ は，分布 q から分布 p への距離を図る尺度であり，$p$ と $q$ が一致するときに 0 となる．すなわち式 (2.16) が 0 になるのは，同時確率 $p(Y_1[k,l], \cdots, Y_{N_s}[k,l])$ と，各音源の出現確率の積 $\Pi_{n=1}^{N_s} p(Y_n[k,l])$ が一致する場合であり，すなわち各音源が統計的に独立である場合である．独立性の評価を行うためには音源の分布を仮定する必要があり，双曲線余弦やラプラス分布などがよく用いられる．

$\hat{\mathbf{W}}[k]$ を解くための方法として，自然勾配法 [63]，補助関数法 [64] などが提案されている．自然勾配法は，式 (2.16) から計算される勾配を $\Delta\mathbf{W}[k]$ を，分離行列の各基底が張る空間における勾配に拡張した上で，勾配法により $\hat{\mathbf{W}}[k]$ を推定する．安定した収束性能が保証される一方で，学習

係数や学習回数などのチューニングが必要となる．補助関数法は，目的関数 $J(\theta)$ を直接最小化する代わりに，$J(\theta) = \min_{\eta} Q(\theta, \eta)$ を満たす関数（補助関数）を最小化することにより，パラメータ $\theta$ を推定する．この方法により，チューニングパラメータの設定を行うことなく，安定した性能が保証される．

ICA により復元される信号は互いに独立であることが期待できる．一方で，出力信号の順番の不定性の問題（パーミュテーション問題）や大きさの不定性の問題（スケーリング問題）がある．これらの問題により周波数間の不整合が発生し，分離性能が劣化する．そのため，分離行列を求めた後にそれらを解消するための後処理が必要となる [44, 46]．

パーミュテーション問題を解決するための方法として，ICA を拡張した IVA が提案されている．IVA は，出力信号のスペクトル成分のベクトル $\mathbf{Y}_n[l]$=$[Y_n(1,l), \cdots, Y_n(N_\omega, l)]^{\mathrm{T}}$ が互いに独立となるような $\mathbf{W}[k]$ を求める．具体的には，以下の式を満たす $\hat{\mathbf{W}}[k]$ を求める．

$$\begin{aligned}\hat{\mathbf{W}}[k] &= \underset{\mathbf{W}[k]}{\operatorname{argmin}}\ KL\left(p(\mathbf{Y}_1[l], \cdots, \mathbf{Y}_{N_s}[l]) || \Pi_{n=1}^{N_s} p(\mathbf{Y}_n[l])\right) \\ &= \underset{\mathbf{W}[k]}{\operatorname{argmin}}\ -\sum_{j=0}^{N_k-1} \log|det(\mathbf{W}[j])| - \sum_{n=1}^{N_s} E[\log p(\mathbf{Y}_n[l])] \\ &\qquad + Const. \end{aligned} \tag{2.17}$$

ここで，$N_K$ は FFT 長を表す．$\hat{\mathbf{W}}[k]$ の推定は ICA と同様に，自然勾配法や補助関数法により行う．IVA では，$p(\cdot)$ として周波数間の関係を考慮した多次元分布を仮定するため，周波数間の整合性は保証される．そのため，パーミュテーション問題が発生しないという利点がある．スケーリング問題に対しては，最小歪定理（minimal distortion principle: MDP）がよく用いられている．MDP では $\mathbf{W}[k]$ が正方行列のとき，下式のような操作により，スケールを正規化する．

$$\mathbf{W}[k] \leftarrow \mathrm{diag}\left(\mathbf{W}^{-1}[k]\right)\mathbf{W}[k]\ , \tag{2.18}$$

ここで，diag(·) は非対角要素を 0 に置き換える操作を表す．$\mathbf{W}[k]$ が正方行列ではないときには，$\mathbf{W}^{-1}[k]$ のかわりに，$\mathbf{W}[k]$ の擬似逆行列を用いればよい．MDP により各周波数における出力は正規化されるが，正規化されたスケールは観測信号に依存しており，より正確に音源を再現するための考慮が必要である．

### 2.2.2 時間周波数マスクに基づく手法

時間周波数マスクは，目的音源が支配的に存在する時間周波数成分のみを通過させ，それ以外の成分を取り除く非線形のフィルタである．これは，音声のスペクトルは主要な成分が疎に表れ，

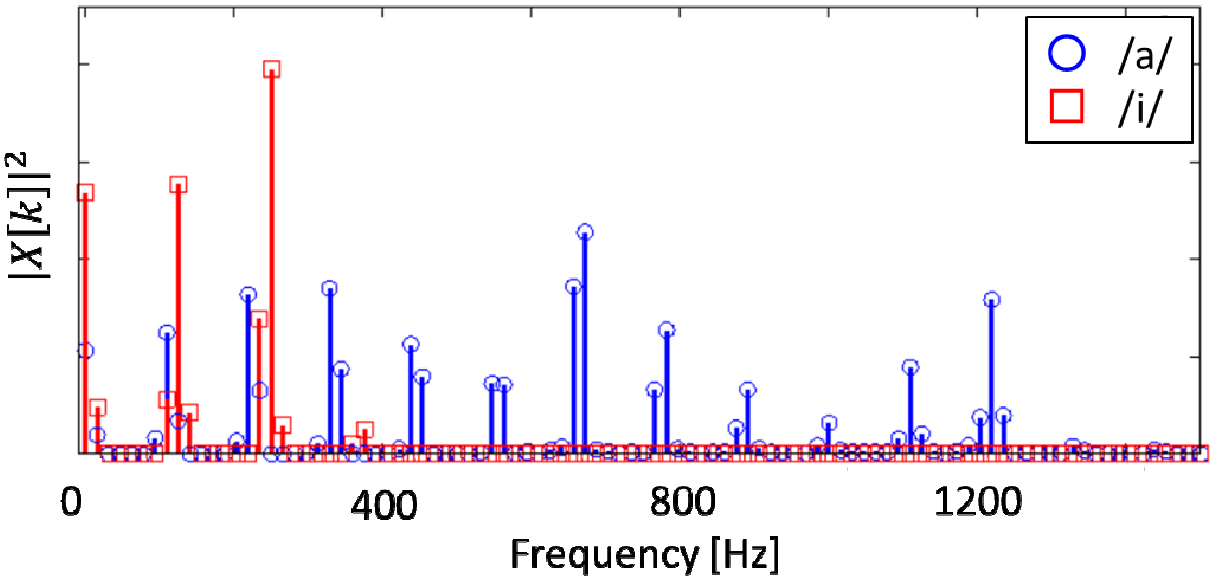



図 2.4: Example of simultaneous speech spectrum. In this case, two speakers uttered difference vowel of /a/ and /i/ simultaneously.

各周波数においては 1 つの支配的な音源の成分のみが存在するというスパース性の仮定に基づく．スパース性の仮定が成立している例を図 2.4 に示す．これは，二人の男声話者が同時に異なる母音（/a/，/i/）を発したときのパワースペクトルを重ねて表示したものであるが，/i/の主要な成分は 400 Hz 以下の範囲で疎に存在しているのに対し，/a/の主要な成分は 100 Hz から 1300 Hz の範囲で疎に存在している．また/a/と/i/それぞれの主要な成分が重なることは稀であり，重なったとしても一方の成分が支配的である様子を確認することができる．ここで課題となるのは，時間周波数マスクをどのように設計するかということである．

同時発音された 2 つの信号を 2 つのマイクロホンで観測し，位相差-振幅比の二次元のヒストグラムを作成すると，音源数分だけのピークが存在し，各音源成分はそれぞれのピーク近傍に現れる [17]．そこで，そこで観測信号から音響特徴を抽出し，音響特徴をクラスタリングすることにより，各 TF ビンが所属する音源を調べるという方法がとられる．

$\mathbf{Z}[k,n]$=$[Z_1[k,l],\cdots,Z_m[k,l]]$ を観測信号，$\psi(\mathbf{Z}[k,n])$ を振幅比や位相差などを表すスペクトル特徴，そして $P(n|\psi(\mathbf{Z}[k,l]))$ を，$n$ 番目の音源が $(k,l)$ 番目の時間周波数ビンに割り振られる事後確率とすると，$i$ 番目の音源を抽出する時間周波数マスク $M_n[k,l]$ は，以下のように設計される．

$$M_n[k,l] = \begin{cases} 1 & \text{if} \quad \arg\max_j P(j|\psi(\mathbf{Z}[k,l])) = n \\ 0 & \text{otherwise,} \end{cases} \tag{2.19}$$

クラスタリングを行う際の特徴としては，ノルムで正規化した振幅比と位相差を組み合わせた特

徴 [19] など，様々なものが提案されている．ここでは，文献 [12] で用いられているように，以下に示すようなノルムで正規化した観測信号ベクトルを用いる．

$$\psi(\mathbf{Z}[k,l]) = \frac{\mathbf{Z}[k,l]}{||\mathbf{Z}[k,l]||} \tag{2.20}$$

以降では，$\psi(\mathbf{Z}[k,l])$ を $\mathbf{x}$ と定義する．

このような特徴に対して事後確率 $P(n|\mathbf{x})$ を計算するためには，複数の複素正規分布を混合した確率密度分布が用いる．$P(n|\mathbf{x})$ は以下のように定義される．

$$P(\mathbf{x};\mathbf{a}_i,\sigma_i) = \sum_{i=1}^{N_s} \alpha_i \frac{1}{(\pi\sigma_i^2)^{N_m-1}} \exp\left(-\frac{||\mathbf{x}-\mathbf{a}_i||^2}{\sigma_i^2}\right) \tag{2.21}$$

$$\sum_{i=1}^{N_s} \alpha_i = 1 \ , \tag{2.22}$$

ここで，$\alpha_i$, $\mathbf{a}_i$, $\sigma_i^2$ はそれぞれ，混合重み，$i$ 番目の正規分布に対する平均および分散を表す．これらのパラメータは，EM アルゴリズム [65] を用いて推定される．式 2.21 に示す確率密度分布より，事後確率は以下のように計算される．

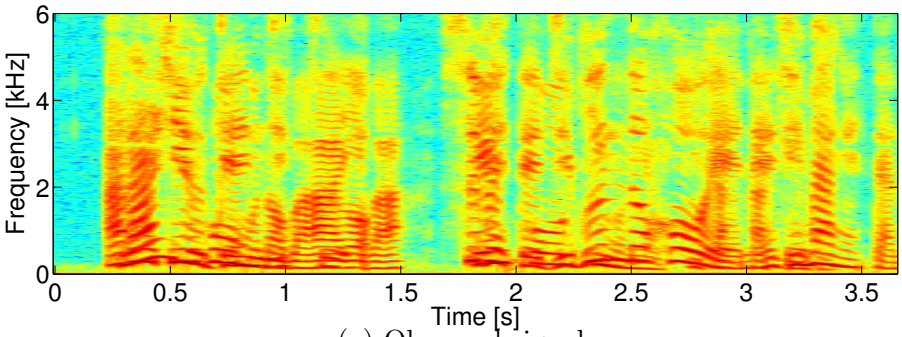
$$P(n|\mathbf{x};\mathbf{a}_i,\sigma_i) = \frac{P(\mathbf{x};\mathbf{a}_i,\sigma_i)}{\sum_{j=1}^{N_s} P(\mathbf{x};\mathbf{a}_j,\sigma_j)} \ , \tag{2.23}$$

時間や周波数関係なく一つのモデルを用いてクラスタリングするならば，パーミュテーション問題は発生しない [19]．しかし実際の環境では，伝達関数の特性は周波数毎に異なっている．そのため，周波数ごとにクラスタリングを行った上で，パーミュテーションを揃える方法が用いられている [12]．

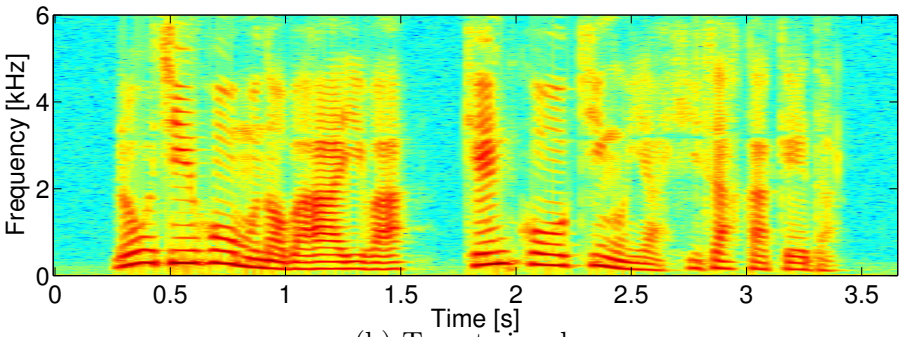
時間周波数マスク $M_n(\omega,\tau)$ を用いると，$n$ 番目の音源のスペクトルは以下のように計算される．

$$Y_n[k,l] = M_n[k,l]Z_n[k,l] \ . \tag{2.24}$$

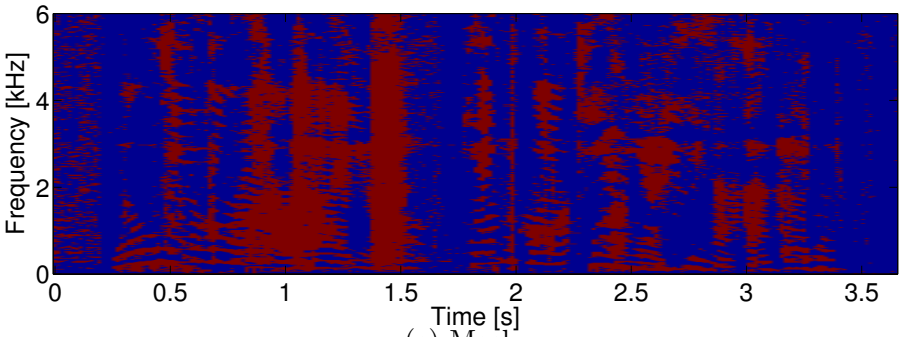
図 2.5 に，時間周波数マスクを用いた分離処理の例を示す．この例では，観測信号のスペクトル (a) より目的音のスペクトル (b) を再現する．(a) より設計された時間周波数マスク (c) を，観測信号のスペクトルを適用すると，背景雑音や妨害音などの不要な成分が取り除かれている様子が確認できる (d)．一方で，目的音が存在する部分が過剰に消されてしまっている様子も確認できる．このような歪は，音質を劣化させる要因となる．



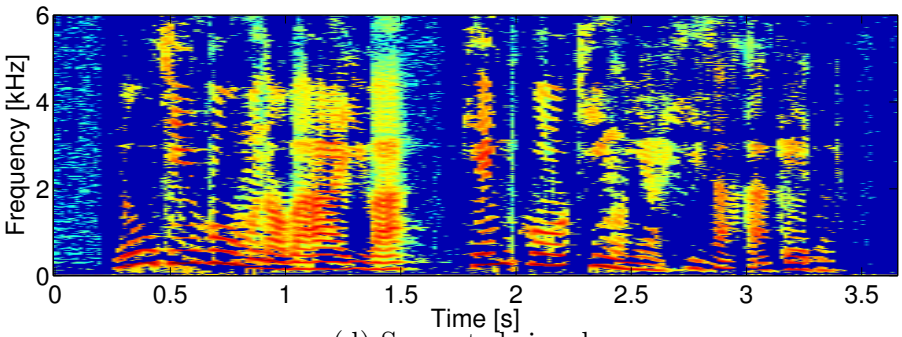
(a) Observed signal



(b) Target signal



(c) Mask



(d) Separated signal

図 2.5: Example of logarithmic power spectra when the time-frequency (TF) masking is applied.

## 2.3 連想記憶モデル

パターン A を入力した際にパターン B を想起するという関係を記憶し，A に類似する任意の入力パターン A' に対して，B に類似するパターンを出力するモデルを連想記憶モデルと呼ぶ．古典的な連想記憶モデルとして，1980 年代初頭に Hopfild らが提案したホップフィールドネットワーク (Hopfield-type associative memory) [66] が挙げられる．ホップフィールドネットワークの各ユニットの出力は，0 か 1 かの二値が決定論的に決定されるが，学習を行う際に局所解の問題が発生する．この問題を解決するために，各ユニットが確率論的に決定されるよう拡張した Boltzmann machine (BM) [67] が提案されたが，BM は全結合のネットワークであるため，学習が難しいという問題点があった．そこで，学習を容易に行うために，可視層・隠れ層の内部は結合しないという制約を加えた RBM [68] が提案されている．RBM の各ユニットの出力が確率分布に従うが，各ユニットの出力が決定論的に決定されるニューラルネットワークとして Autoencoder (AE) [69] がある．AE の主な拡張として，Denoising Autoencoder (DAE) [28]，Convolutional autoencoder (CAE) [70] が提案されている，DAE では，入力パターンにノイズを加えることにより．ノイズに対して頑健な連想記憶が実現される．CAE では，入力パターンを，局所的なパターンの組み合わせとしてみることで，入力パターンの局所的な動きを考慮することができる．CAE は CNN [71] における畳み込み層における重みの初期値として用いることができる．ここでは，連想記憶モデルとして，RBM，AE，DAE，CAE，CNN について紹介する．

### 2.3.1 Restricted Boltzmann machine

RBM は，可視層と隠れ層で構成される連想記憶型ネットワークであり，可視層・隠れ層のノードが 0 または 1 の 2 値定義される場合は Bernoulli-Bernoulli RBM (BB-RBM)，可視層のノードが実数かつ隠れ層のノードが 0 または 1 の 2 値で定義された場合は Gaussian-Bernoulli RBM (GB-RBM) とよばれる．BM の考え方を踏襲しているものの，図 2.6 に示すように,「可視層と隠れ層の間に無指向の依存関係があるものの，各層内部では依存関係が存在しない」という構造の制約を加えることにより，BM における学習困難性を解決している．また，補間能力に優れるという利点を持つ．

可視層および隠れ層のノードの集合をそれぞれ $\mathbf{v} = \{v_n | 1 \leq n \leq N\}$，$\mathbf{h} = \{h_m | 1 \leq m \leq M\}$ と定義するとき，RBM は，同時確率 $p(\mathbf{v}, \mathbf{h})$ を以下の Boltzmann 分布に従う確率モデルで表現

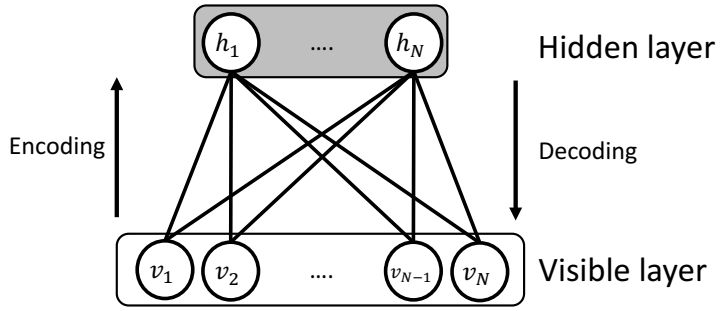



図 2.6: Restricted Boltzmann Machine.

する．

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z(\mathbf{v}, \mathbf{h})} \exp\left(-E(\mathbf{v}, \mathbf{h})\right) \tag{2.25}$$

$$Z(\mathbf{v}, \mathbf{h}) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h})\right) \tag{2.26}$$

$E(\mathbf{v}, \mathbf{h})$ はエネルギー関数とよばれ，RBM の種類によって異なる関数が定義される．具体的に BB-RBM では，

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{n=1}^{N} v_n b_n - \sum_{m=1}^{M} h_m c_m - \sum_{m=1}^{M} \sum_{n=1}^{N} w_{nm} h_m v_n \tag{2.27}$$

GB-RBM では，

$$E(\mathbf{v}, \mathbf{h}) = \sum_{n=1}^{N} \frac{(v_n - b_n)^2}{2{\sigma_n}^2} - \sum_{m=1}^{M} h_m c_m - \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{w_{nm} h_m v_n}{\sigma_n} \tag{2.28}$$

のように定義される．このとき，$b_n$，$c_m$ はバイアス，$w_{nm}$ は，$n$ 番目のノードと $m$ 番目のノード間の結合重みを表す．また，$\sigma_n^2$ は可視層の $n$ 番目のノードにおける分散を表す．BB-RBM における事後確率 $p(v_n|\mathbf{h})$，$p(h_m|\mathbf{v})$ は，それぞれ以下のように計算される．

$$p(v_n|\mathbf{h}) = sigmoid\left(\sum_{m=1}^{M} w_{nm} v_n + c_n\right) \tag{2.29}$$

$$p(h_m|\mathbf{v}) = sigmoid\left(\sum_{n=1}^{N} w_{mn} v_n + b_m\right) \tag{2.30}$$

ここで，$sigmoid(\cdot)$ は以下に示すようなシグモイド関数を表す．

$$sigmoid(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x})} \tag{2.31}$$

GB-RBM における事後確率 $p(v_n|\mathbf{h})$，$p(h_m|\mathbf{v})$ は，それぞれ以下のように計算される．

$$p(v_n|\mathbf{h}) = Norm\left(v_i; \sum_{m=1}^{M} w_{nm} h_m + c_n, \sigma^2\right) \tag{2.32}$$

$$p(h_m|\mathbf{v}) = sigmoid\left(\sum_{n=1}^{N} w_{mn}v_n + b_m\right) \tag{2.33}$$

上式のうち，$Norm(\cdot)$ は正規分布を表す．可視層から隠れ層への伝播は，入力 $\mathbf{v}$ を $M$ ビットで表現される符号系列 $\mathbf{h}$ にエンコードする過程，隠れ層から可視層への伝播は，$\mathbf{h}$ から $\mathbf{v}$ へデコードする過程であると解釈することができる．

BB-RBM および GB-RBM において決定するべきパラメータは，それぞれ，$\theta^{(BB)} = (w_{nm}, b_n, c_m)$ および $\theta^{(GB)} = (w_{nm}, b_n, c_m, \sigma^2)$ となる．GB-RBM を学習する際の入力データ $\mathbf{v}$ の平均および分散がそれぞれ 0 および 1 となるように標準化すると，$\sigma^2 = 1$ と置き換えることができる．すなわち，GB-RBM において決定するべきパラメータは，$\theta^{(RB)} = (w_{nm}, b_n, c_m)$ となり $\theta^{(BB)}$ と一致する．以降では，学習データは標準化されたものを用いることとし，$\theta^{(RB)}$ および $\theta^{(BB)}$ を $\theta$ と記述する．最終的なパラメータは，学習データが入力された際の対数尤度関数 $L(\theta)$ を最大とするものが選ばれる．対数尤度関数 $L(\theta)$ は，以下のように定義される．

$$\begin{aligned} L(\theta) &= \sum_{\mathbf{v}} \left( \sum_{\mathbf{h}} \log p(\mathbf{v}, \mathbf{h}; \theta) \right) q(\mathbf{v}) \\ &= \sum_{\mathbf{v}} \left( \log \sum_{\mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h}; \theta)\right) \right) q(\mathbf{v}) - \log Z(\mathbf{v}, \mathbf{h}; \theta) \end{aligned} \tag{2.34}$$

上式において，$q(\mathbf{v})$ は観測データの分布を示す．ここで，(2.34) 式をパラメータ $\theta$ に関して偏微分すると，以下のように変形される．

$$\begin{aligned} \frac{\partial L(\theta)}{\partial \theta} &= \sum_{\mathbf{v}} \left( \log \sum_{\mathbf{h}} \frac{\partial}{\partial \theta} E(\mathbf{v}, \mathbf{h}; \theta) \frac{\exp\left(-E(\mathbf{v}, \mathbf{h}; \theta)\right)}{\sum_{\mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h}; \theta)\right)} \right) q(\mathbf{v}) \\ &\quad - \frac{\partial}{\partial \theta} \log Z(\mathbf{v}, \mathbf{h}; \theta) \\ &= -\sum_{\mathbf{v}} \sum_{\mathbf{h}} \left( \frac{\partial}{\partial \theta} E(\mathbf{v}, \mathbf{h}; \theta) \right) p(\mathbf{v}|\mathbf{h}; \theta) q(\mathbf{v}) \\ &\quad + \sum_{\mathbf{v}} \sum_{\mathbf{h}} \left( \frac{\partial}{\partial \theta} E(\mathbf{v}, \mathbf{h}; \theta) \right) p(\mathbf{v}, \mathbf{h}; \theta) \end{aligned} \tag{2.35}$$

上式の第一項は，データ分布に対する期待値を表し，学習データを用いて容易に計算が可能である．一方，第二項はモデル分布に対する期待値を表し，マルコフ連鎖モンテカルロ法 (MCMC) などを用いて計算する必要がある．そのため，パラメータ推定するためには多大の計算が必要となる．計算量を削減するための一つの手法として，RBM で表現するモデル分布とデータ分布の間に関連があると仮定し，データ分布を初期値と用いた Contrastive Divergence (CD) 法 [72] と呼ばれる学習方法が提案されている．Algorithm 1 に，CD 法によるパラメータ推定の流れを示す．

**Algorithm 1** Contrastive Divergence

1: Initialize model parameters $\theta = (w_{mn}, b_n, c_m)$.

2: Estimate $\hat{\mathbf{h}}^{(t)}$ using a posterior probability distribution $p(\mathbf{h}|\mathbf{v})$, where $\mathbf{v}$ represent training data.

3: Estimate $\tilde{\mathbf{v}}^{(t)}$ using a posterior probability distribution $p(\mathbf{v}|\hat{\mathbf{h}}^{(t)})$.

4: Estimate $\tilde{\mathbf{h}}^{(t)}$ using a posterior probability distribution $p(\mathbf{h}|\tilde{\mathbf{v}}^{(t)})$.

5: Update $\theta$ as follows, where $\mu$ denotes learning rate:

$$w_{nm}^{(t)} = w_{nm}^{(t-1)} + \mu(v_n \hat{h}_m^{(t)} - \tilde{v}_n^{(t)} \tilde{h}_m^{(t)}) \tag{2.36}$$

$$b_n^{(t)} = b_n^{(t-1)} + \mu(v_n - \tilde{v}_n^{(t)}) \tag{2.37}$$

$$c_m^{(t)} = c_m^{(t-1)} + \mu(\hat{h}_m^{(t)} - \tilde{h}_m^{(t)}) \tag{2.38}$$

6: repeat from 2: to 5: until parameter update is converged.

対数パワースペクトルのように連続値のデータを用いる場合には，可視層に実数を想定している GB-RBM を採用することが望ましい．このとき，入力 $\mathbf{x}$ に対する推定値 $\mathbf{y}$ は以下のように計算される．

$$\mathbf{y} = \mathbf{W}^T \mathbf{h} + \mathbf{c} \tag{2.39}$$

$$\mathbf{h} = sigmoid(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{2.40}$$

$^T$ は転置を表し，$\mathbf{b}$，$\mathbf{c}$ はそれぞれ $b_m$，$c_n$ を要素にもつベクトル，$\mathbf{W}$ は $v_n$ から $h_m$ への重み $w_{mn}$ を要素にもつ行列を表す．

### 2.3.2 Autoencoder

AE は，図 2.7 に示すように，中間層のノード数が入力層のノード数よりも少ない砂時計型の構造をもつ Feed-forward 型のニューラルネットワークにより恒等写像を行う．中間層において入力されたデータの抽象表現が抽出されることを期待しており，入力層から中間層への過程を中間表現へのエンコード，中間層から出力層への過程を中間表現からのデコードと解釈することができる．

入力層，中間層，出力層のノードをそれぞれ，$\mathbf{v} = \{v_n | 1 \leq n \leq N\}$，$\mathbf{h} = \{h_m | 1 \leq n \leq N\}$，$\mathbf{o} = \{o_n | 1 \leq n \leq N\}$ とすると，次のような関係が成立する．

$$\mathbf{h} = f(\mathbf{W}^{(E)}\mathbf{v} + \mathbf{b}) \tag{2.41}$$

$$\mathbf{o} = \mathbf{W}^{(D)}\mathbf{h} + \mathbf{c} \tag{2.42}$$

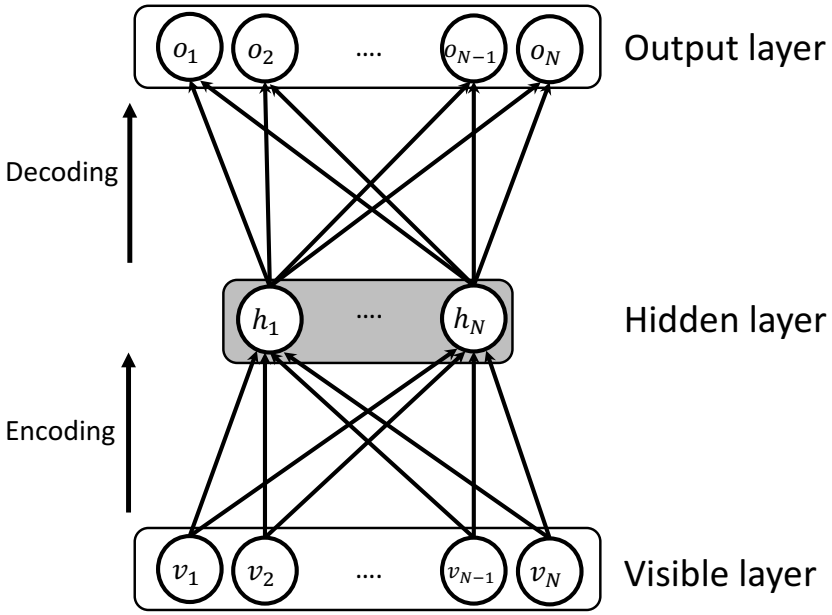



図 2.7: Autoencoder.

**b**，**c** は，**h**，**o** におけるバイアスを，$\mathbf{W}^{(E)}$ および $\mathbf{W}^{(D)}$ はそれぞれ，$v_n$ から $h_m$ への重み $w_{mn}^{(E)}$ を要素にもつ行列，および，$h_m$ から $o_n$ への重み $w_{nm}^{(D)}$ を要素にもつ行列を示す．$f(\cdot)$ は非線形関数を表し，シグモイド関数や tanh 関数などが用いられる．

AE のパラメータ $\theta = \{\mathbf{W}^{(D)}, \mathbf{W}^{(E)}, \mathbf{b}, \mathbf{c}\}$ は，誤差逆伝播法 [73] により学習する．誤差逆伝播法では，まず学習データを入力層から出力層まで伝播させ，モデルの出力と教師データとの誤差 $J(\theta)$ を計算する．このとき，教師データは入力データが与えられる．また，$J(\theta)$ は目的関数と呼ばれる．そして，誤差関数の値を出力層から入力層に逐次的に伝播させ，各層のパラメータの値を，下式に基づいた確率的降下法により補正する．

$$\theta^{(t)} \leftarrow \theta^{(t-1)} - \mu \frac{\partial}{\partial \theta} J(\theta) \tag{2.43}$$

$J(\theta)$ としては，以下に示すような，入出力データ間の二乗誤差関数が広く用いられる．

$$\begin{aligned} J(\theta) &= \frac{1}{2}\sum_{n}^{N} ||v_n - o_n||^2 \\ &= \frac{1}{2}\sum_{n=1}^{N} \left|\left| v_n - \sum_{m=1}^{M} w_{nm}^{(D)} f\left(\sum_{n=1}^{N} w_{mn}^{(E)} v_n + b_m\right) - c_n \right|\right|^2 \end{aligned} \tag{2.44}$$

AE を学習する際に，$\mathbf{W}^{(E)} = \left(\mathbf{W}^{(D)}\right)^T$ の制約（"tied weights"）を加えると，ネットワーク構造としては RBM に類似する．しかし，RBM はモデルに対する尤度の最大化を目的関数にしているのに対し，Autoencoder は入出力の二乗誤差の最小化を目的関数にしているという点で異なる．

**Algorithm 2** Greedy Layer-wise training for deep denoising autoencoder

**Require:** weight matrix and bias vector on $n$-th layer, $\mathbf{W}^{(n)}$ and $\mathbf{b}^{(n)}$, #layers N

1: Update $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ with contrastive divergence method assuming that first layer as GB-RBM.
2: Calculate output of the first layer $\mathbf{h}^{(1)}$ using $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ .
3: **for** $i$=2:$N/2$
4: Update $\mathbf{W}^{(i)}$ and $\mathbf{b}^{(i)}$ with contrastive divergence method assuming that $i$-th layer and $\mathbf{h}^{(i-1)}$ as BB-RBM and input.
5: Calculate output of the first layer $\mathbf{h}^{(i)}$ using $\mathbf{W}^{(i)}$ and $\mathbf{b}^{(i)}$ .
6: **end for**
7: **for** $j$=1:$N/2-1$
8: $\mathbf{W}^{(N/2+j)} \leftarrow \mathbf{W}^{(N/2-j)T}$ ($^T$ denotes transposition)
9: $\mathbf{b}^{(N/2+j)} \leftarrow \mathbf{b}^{(N/2-j)}$
10: **end for**
11: Update all parameters of denoising autoencoder using back propagation method.

### 2.3.3 Denoising autoencoder

DAE は AE の拡張であり，ノイズにより歪んだ信号 $\tilde{\mathbf{v}} = d(\mathbf{v})$（$d(\cdot)$ はノイズを重畳する関数）を入力とし，ノイズが除去された信号 $\mathbf{v}$ を出力する．AE よりも雑音に対して頑健な連想記憶の実現が期待できる．AE と同様に，パラメータの学習は誤差逆伝播法により行う．このときの目的関数は以下のように，$\tilde{\mathbf{v}}$ を入力した際の出力 $\mathbf{o}$ とクリーン信号 $\mathbf{v}$ との二乗誤差として定義される．

$$\begin{aligned} J(\theta) &= \frac{1}{2}\sum_{n}^{N} ||v_n - o_n||^2 \\ &= \frac{1}{2}\sum_{n=1}^{N} \left|\left| v_n - \sum_{m=1}^{M} w_{nm}^{(D)} f\left(\sum_{n=1}^{N} w_{mn}^{(E)} d(v_n) + b_m\right) - c_n \right|\right|^2 \end{aligned} \tag{2.45}$$

DAE は，中間層の層数を増やすことにより，表現能力を増やすことができる．しかし，中間層の層数が多いニューラルネットワークを誤差逆伝播法で学習する際には，“vanishing gradient” と呼ばれる現象により，学習がうまくできないという問題が指摘されている．この解決するための方法として，貪欲学習 [24] と呼ばれる学習方法が提案されている．Algorithm 2 に，貪欲学習のアルゴリズムを示す．1:〜10:の処理は，パラメータの初期値を推定するもので，pre-training とよばれる．このとき，RBM の代わりに tied weights の制約を加えた AE を用いることも可能である．また，11:の処理は，pre-training により推定されたパラメータを，誤差関数に基づき修正するこ

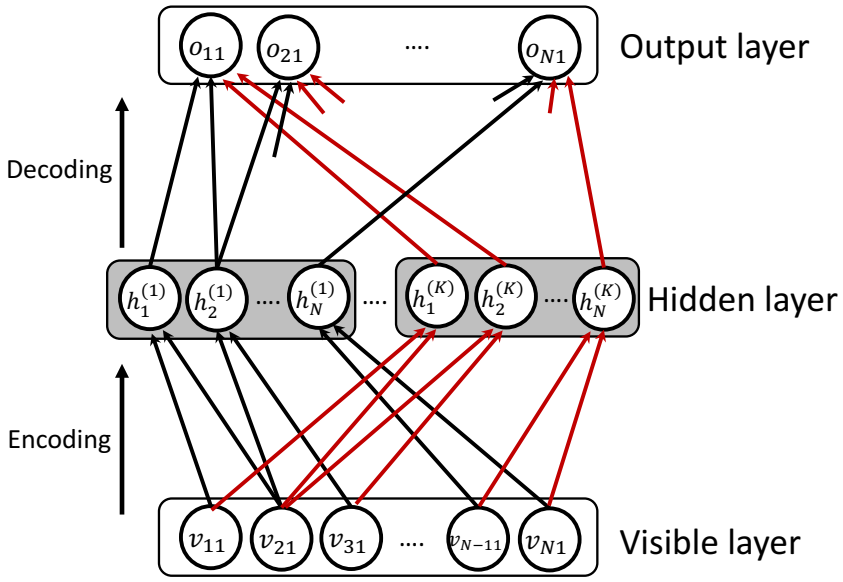



図 2.8: Convolutional autoencoder.

とより，fine-tuning とよばれる．

### 2.3.4 Convolutional autoencoder

入力パターンとして音声の時間周波数パターンのような 2 次元のデータを用いるとき，AE や DAE は，時間と周波数の関係を考慮しない．一方 CAE は，入力パターンを細かい局所パターンの組み合わせとして考える．そのため，倍音成分の時間変化などの局所的な変化を考慮しながら，音声のスペクトルを表現するのに．ふさわしい特徴を抽出されることが期待できる．

図 2.8 に CAE の構造を示す．AE と同様に，エンコード，デコードの処理からなると解釈することができる．エンコード処理ではまず，入力される $N_I \times M_I$ の 2 次元パターン $\mathbf{x} = \{x_{ij} | 0 \leq i \leq N_I - 1,\ 0 \leq j \leq M_I - 1\}$ に対して，$K$ 種類の $N_F \times M_F$ の二次元フィルタ $\mathbf{W}_k^{(E)} = \{W_k^{(E)} | 0 \leq k \leq K - 1\}$ を畳み込むことにより，$K$ 種類の応答 $\mathbf{h} = \{h_{ij}^k | 0 \leq i \leq (N_I + N_F - 1) - 1,\ 0 \leq j \leq (M_I + M_F - 1) - 1,\ 0 \leq k \leq K - 1\}$ を得る．$W_K^{(E)}$ は局所特徴 $\mathbf{h}^k = \{h_{ij}^k | 0 \leq i \leq N_I - 1,\ 0 \leq j \leq M_I - 1\}$ を抽出するフィルタを表す．例えば音声スペクトルを入力すると，局所特徴として倍音構造の上昇や下降などの時間変化を表す特徴が抽出される．デコードでは，$\mathbf{h}^k$ に対して，$K$ 種類の $N_F \times M_F$ の二次元フィルタ $\mathbf{W}_k^{(D)} = \{W_k^{(D)} | 0 \leq k \leq K - 1\}$ を，畳み込むことによりパターン $\mathbf{o} = \{o_{ij} | 0 \leq i \leq N_I - 1,\ 0 \leq j \leq M_I - 1\}$ を出力する．

CAE のパラメータである畳み込みフィルタは，入力と出力の二乗誤差を目的関数を用いて，誤

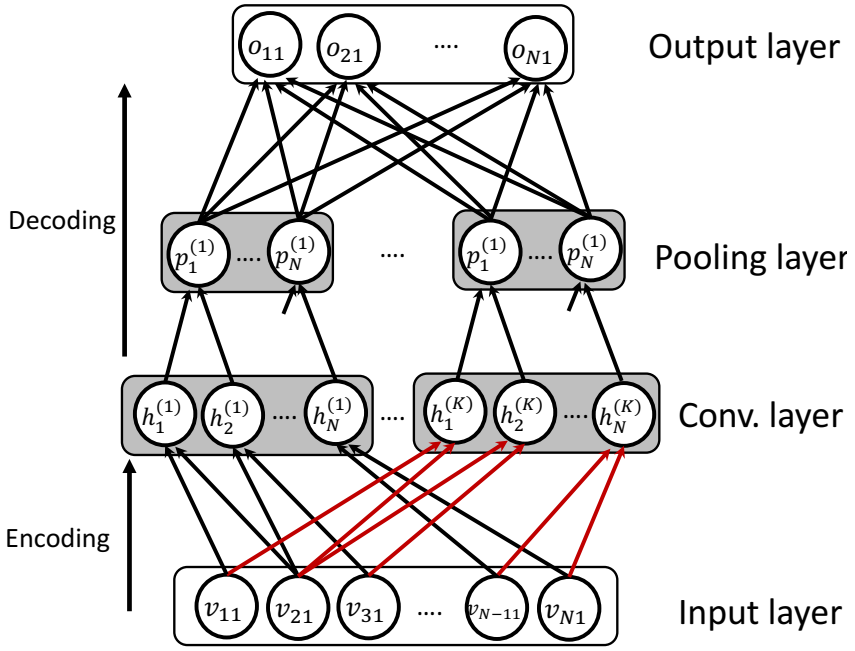



図 2.9: Convolutional neural network.

差逆伝播法により学習する．このとき，$\mathbf{h}^k$ に対して Max-pooling の処理を適用すると，$\mathbf{W}^{(E)}$ を効率的に学習するが可能となる．Max-pooling は，$\mathbf{h}^k$ を $N_p \times M_p$ 個の小領域に分割し，各領域における最大値はそのままにし，それ以外の値を 0 とすることにより実現される．

### 2.3.5 Convolutional neural network

図 2.9 に CNN の構造を示す．CNN は，音声認識の分野において高い性能を示すことが報告されている [71]．音声認識で用いる場合には，入力ベクトルに対してフィルタを畳み込み，その出力に対して Max-pooling を適用し，全接続のネットワークにより音素事後確率を出力するという構造が広く用いられる．音素事後確率の代わりに，音声スペクトルを出力として用いれば，音声の局所特徴を考慮しながら音声を推定するモデルを構築することが可能となる．詳細は，次章にて説明する．

# 第3章　連想記憶を用いた線形分離行列推定法

## 3.1　はじめに

聴感上の違和感が少なく音源信号を復元可能なブラインド音源分離の実現を目指し，線形分離フィルタに基づく枠組みにおいて，連想記憶を用いた分離行列の推定法を開発した．

独立成分分析（Independent component analysis: ICA）[39] や独立ベクトル分析 (Independent vector analysis: IVA) [40] などに代表される線形行列に基づく枠組みは，聴感上違和感を与える非線形歪が原理的に発生しないという利点がある．ICA や IVA では，音源が互いに独立であるという仮定（音源の独立性の仮定）に基づき，分離信号間の独立性が最大となるような分離行列を推定する．そのため，音源の独立性の仮定が成立しない場合には，分離性能が劣化する．例えば直接音と壁や床などから到来する反射音は互いに非独立であるため，そのような環境においては，音源の独立性を頼りに分離を行うことは難しい．

そこで，音源の独立性の代わりに，音源である音声らしさを陽に扱いながら分離行列を推定する枠組みを提案した [58]．類似した考えとして，音源を表現するのに適切な分布を用いて独立性を評価することで，音声らしさを考慮する方法 [48, 49, 50] がいくつか提案されている．しかし，分布を仮定する枠組みでは，より直接的に音声らしい分離音を出力することが難しい．

提案法は，分離行列を推定する際に，音声信号に類似する信号（参照信号）を導入する．参照信号は，事前に音声のスペクトルを学習させた連想記憶モデルにより推定する．参照信号を音源信号とみなし，分離行列により推定された分離信号と参照信号との誤差が最小となる分離行列を求める．すなわち，提案法では，線形分離フィルタの枠組みの中で再現しうる信号の中で，音声らしい信号が再現されることが期待できる．

本章では，提案法の枠組みとその有効性について述べる．ここでは特に，連想記憶モデルとして DAE，CNN，Denoising convolutional autoencoder（DCAE）について概観する．また，二話者同時発話音声の分離実験をおこない，提案法が IVA により生じる歪を低減する効果があることを示す．

## 3.2 手法概要

### 3.2.1 概要

音声の特徴を考慮した枠組みを考慮しながら分離行列を推定する枠組みを提案する．提案法の処理の流れを図 3.1 に，アルゴリズムを Algorithm 3 に示す．提案法では，以下の処理を繰り返すことにより，連想記憶により得られる分離信号のスペクトル $\mathbf{Y}[k,l]$ に対応する参照信号のスペクトル $\mathbf{S}[k,l]$ と，$\mathbf{Y}[k,l]$ との誤差が最小となる線形分離行列 $\mathbf{W}[k]$ を求める．

**参照信号推定** ：

分離行列 $\mathbf{W}[k]$ を用いて分離信号のスペクトル $\mathbf{Y}[k,l]$ をもとめ，事前に音声のスペクトルを学習した連想記憶モデルにより $\mathbf{Y}[k,l]$ から音源らしさを表すスペクトル $\hat{\mathbf{S}}[k,l]$（以降，参照信号とよぶ）を推定する．

**分離行列更新** ：

参照信号 $\hat{\mathbf{S}}[k,l]$ を音源信号とみなし，分離信号 $\mathbf{Y}[k,l]$ と $\hat{\mathbf{S}}[k,l]$ との誤差が最小となるように分離行列 $\mathbf{W}[k]$ を更新する．

連想記憶が音源信号を適切に推定することができるならば，最終的に得られる分離信号は音源信号を出力されることを期待することができる．またこの枠組みでは，ICA や IVA などの従来の線形分離行列の枠組みで広く用いられている音源信号の独立性の仮定を立てる必要性がないため，エコー信号が混入する等の音源の独立性のみで解くことが困難な状況でも分離が可能であると思われる．

以降では，参照信号を推定するための連想記憶モデルの構築方法，および，分離行列の更新方法について述べる．また，ICA や IVA などに代表される独立性に基づき分離行列を推定する枠組みと提案法との関係についても述べる．

### 3.2.2 参照信号推定

分離信号のスペクトルには，妨害音の消し残しや分離処理により生じた歪などのノイズが含まれているものと考えられる．そこで，分離信号のスペクトルに含まれるノイズを DAE の枠組みを用いて取り除くことにより，参照信号を推定する方法を検討した．本研究で用いる連想記憶モデルは，入力された分離信号のスペクトルから，音声らしさを表す特徴ベクトルを抽出するエンコード，抽出された特徴ベクトルから参照信号を復元するデコードの 2 つの処理により構成する．こ

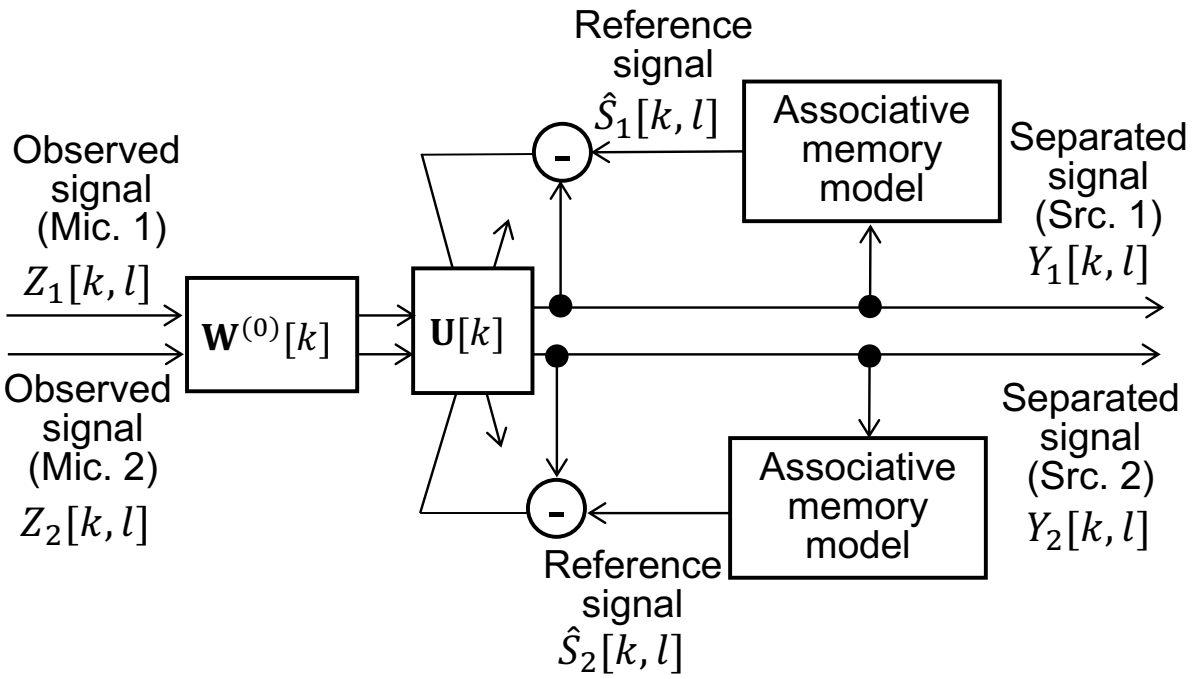



図 3.1: Schematic diagram of proposed method when two sources are estimated from two observations. System method consists of reference spectrum estimation and linear separation matrix update.

こでは，全接続型ニューラルネットワークに基づく連想記憶モデル（**Full-Full**），CNN に基づく連想記憶モデル（**Conv-Full**），そして提案する DCAE に基づく連想記憶モデル (**Conv-Conv**) の 3 種類の異なる連想記憶モデルを構築し，性能を比較した．

**Full-Full** の連想記憶モデルは，DAE で広く用いられているニューラルネットワークの構造である．音声を表現するのに不要な成分を取り除く効果が期待できるが，音声スペクトルの局所的なパターンを考慮していないという問題がある．**Conv-Full** の連想記憶モデルは，スペクトルを局所パターンの組み合わせとして考える．そのため，局所的に存在するようなノイズに対して頑健であることが期待できる．また，Max-pooling と組み合わせることにより，局所パターンを抽出するフィルタを効率的に学習することができる．しかし Max-pooling を用いることにより，局所パターンの時間周波数の位置情報が失われてしまう．そのため，元音声らしさが失われる可能性がある．**Conv-Conv** は，エンコード時に失われる局所パターンの位置情報を，デコード時に陽に扱う．そうすることで，元音声らしさを考慮した参照信号の推定を実現する．以降では，それぞれの構造について説明する．

**Algorithm 3** Algorithm for separation matrix optimization using associative memory.

**Require:** Observed signal $\mathbf{Z}[k,l]$

**Require:** Initial separation matrix obtained either by an existing linear filtering method (as described in Section 3.2.3) or a TF mask (as described in Section 4.2) $\mathbf{W}^{(0)}[k]$

**Require:** #epochs for reference signal update $N_R$, #epochs for filter update $N_M$, learning rate $\mu$

$\mathbf{M}^{(0)}[k] = \mathbf{E}$ ($\mathbf{E}$: identity matrix).
$\mathbf{Y}^{(0)}(k,l) = \mathbf{W}^{(0)}(k)\mathbf{Z}(k,l)$.
Estimate $\hat{\mathbf{S}}^{(0)}(k,l)$ from $\mathbf{Y}^{(0)}(k,l)$ using an associative memory model (AMM).
**for** $i = 0 : N_R$-1
    // Separation matrix optimization (Section 3.2.3)
    **for** $j = 0 : N_M$-1
        Calculate gradient $\mathbf{G}^{(j)}[k]$ using Eq. (3.17) given $\hat{\mathbf{S}}^{(i)}[k,l]$ and $\mathbf{Y}^{(i)}[k,l]$.
        $\mathbf{U}^{(j+1)}[k] = \mathbf{U}^{(j)}[k] - \mu\mathbf{G}^{(j)}[k]/||\mathbf{G}^{(j)}[k]||$.
    **end for**
    $\bar{\mathbf{U}}[k] = \mathbf{U}^{(N_M)}[k]$.
    $\mathbf{Y}^{(i+1)}[k,l] = \bar{\mathbf{U}}[k]\mathbf{W}^{(0)}\mathbf{Z}[k,l]$.
    // Reference signal estimation (Section 3.2.2)
    Estimate $\hat{\mathbf{S}}^{(i+1)}[k,l]$ from $\mathbf{Y}^{(i+1)}[k,l]$ using AMM.
    $\mathbf{U}^{(0)}[k] = \bar{\mathbf{U}}[k]$.
**end for**

**Output:** $\bar{\mathbf{W}}[k] = \bar{\mathbf{U}}[k]\mathbf{W}^{(0)}[k]$.

### DAE に基づく連想記憶モデル (Full-Full)

分離信号の対数パワースペクトルから，513 ビン ×10 フレームの時間周波数パターン **I** を，5 フレーム間隔で切り出す．**I** に対して平均を 0，分散を 1 とする標準化を適用したものを連想記憶の入力として用いる．同様の処理を，対応する目的信号の対数パワースペクトルに対しても行い，抽出された時間周波数パターン **O** を抽出する．こうして得られた **I** から **O** への写像関数を，図 3.2 に示すような構造により実現する．

エンコード処理では **I** を 5130 次元のベクトルに変換し，以下の式により，音声らしさを表す 2048 次元の特徴ベクトル **h** を抽出する．

$$\mathbf{h}_j = \begin{cases} \tanh\left(\mathbf{w}_1^{(e)}\mathbf{I} + \mathbf{b}_1^{(e)}\right) & j = 1 \\ \tanh\left(\mathbf{w}_j^{(e)}\mathbf{h}_{j-1} + \mathbf{b}_j^{(e)}\right) & j = 2 \cdots L-1 \end{cases} \tag{3.1}$$

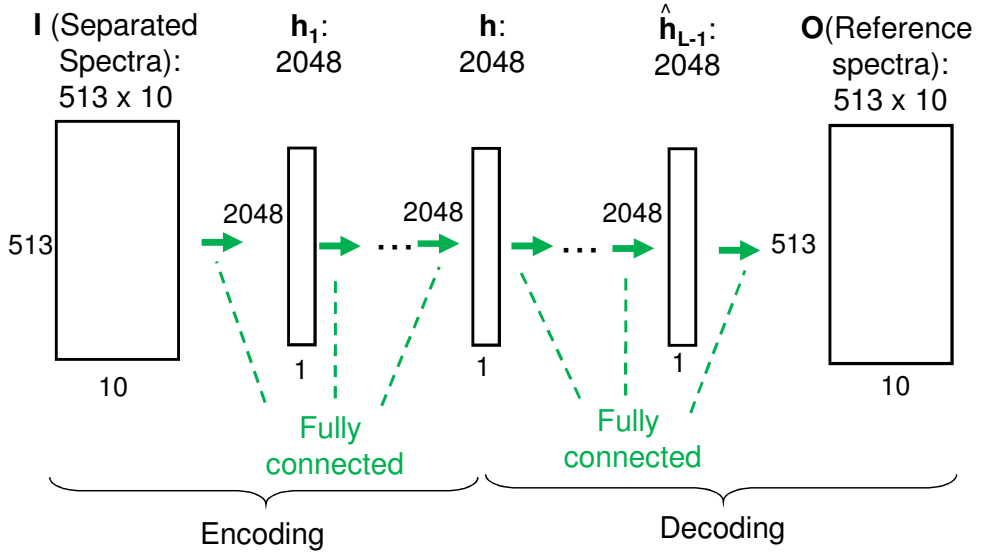



図 3.2: Architecture of the denoising autoencoder (DAE)-based associative memory model (AAM). $\mathbf{h}$ denotes a feature vector extracted during encoding.

$$\mathbf{h} = \tanh\left(\mathbf{w}_L^{(e)}\mathbf{h}_{L-1} + \mathbf{b}_L^{(e)}\right) \tag{3.2}$$

ここで $L$ は，$\mathbf{I}$ から $\mathbf{h}$ を抽出する関数を表現するための隠れ層の層数を表す．また $\mathbf{w}_j^{(e)}$ および $\mathbf{b}_j^{(e)}$ は，$j$ 番目の隠れ層における重みおよびバイアスを表す．

デコード処理では，下式により $\mathbf{h}$ から $\mathbf{O}$ を復元する．

$$\hat{\mathbf{h}}_j = \begin{cases} \tanh\left(\mathbf{w}_1^{(d)}\mathbf{h} + \mathbf{b}_1^{(d)}\right) & j = 1 \\ \tanh\left(\mathbf{w}_j^{(d)}\hat{\mathbf{h}}_{j-1} + \mathbf{b}_j^{(e)}\right) & j = 2 \cdots L-1 \end{cases} \tag{3.3}$$

$$\mathbf{O} = \mathbf{w}_L^{(d)}\hat{\mathbf{h}}_{L-1} + \mathbf{b}_L^{(d)} \tag{3.4}$$

ここで，$\mathbf{w}_j^{(d)}$ および $\mathbf{b}_j^{(d)}$ は，$j$ 番目の隠れ層における重みおよびバイアスを表す．こうして得られた $\mathbf{O}$ を 513 ビン $\times 10$ フレームの時間周波数パターンに変換し，重畳加算法により参照信号の対数パワースペクトルを計算する．

連想記憶のパラメータ $\{ \mathbf{w}_1^{(e)}, \cdots, \mathbf{w}_L^{(e)}, \mathbf{b}_1^{(e)}, \cdots, \mathbf{b}_L^{(e)}, \mathbf{w}_1^{(d)}, \cdots, \mathbf{w}_L^{(d)}, \mathbf{b}_1^{(d)}, \cdots, \mathbf{b}_L^{(d)} \}$ は，誤差逆伝播法 [73] によって決定する．

### CNN に基づく連想記憶モデル (Conv-Full)

**Full-Full** と同様の方法で，連想記憶の入力 $\mathbf{I}$ および出力 $\mathbf{O}$ を抽出し，図 3.3 に示すようなネットワークにより，$\mathbf{I}$ から $\mathbf{O}$ への写像関数を実現する．

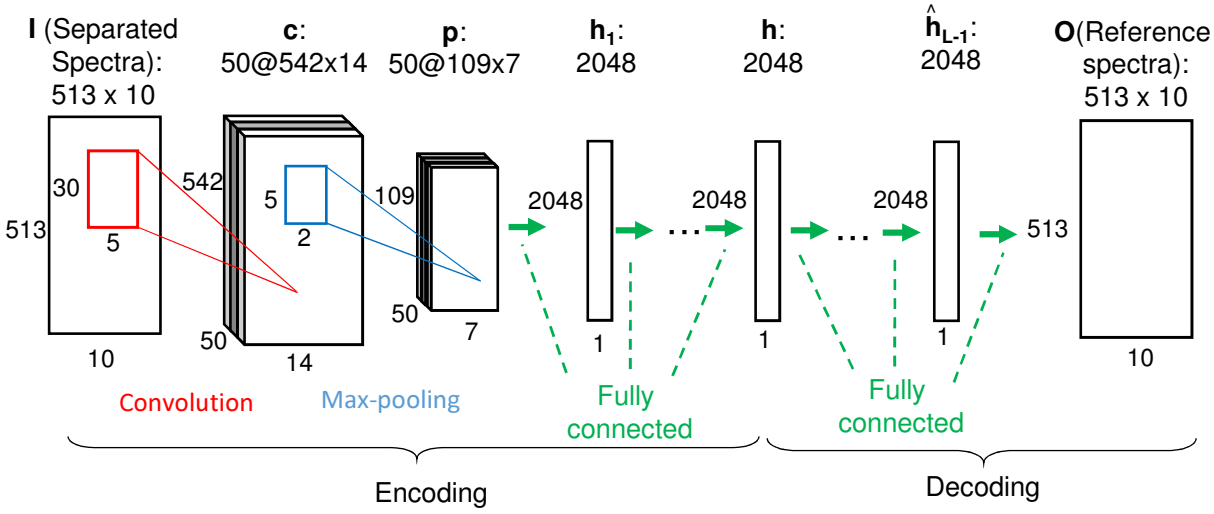



図 3.3: Architecture of the convolutional neural network (CNN)-based associative memory model (AAM). $\mathbf{c}$, $\mathbf{p}$ and $\mathbf{h}$ denote the feature map, down-sampled feature map and a feature vector extracted during encoding, respectively. $a$ in $a@b\times c$ denotes the number of filters, and $b$ and $c$ denote the sizes of the filter outputs in frequency and time, respectively. Note that location information is lost in the max-pooling process.

エンコーディング処理ではまず，$\mathbf{I}$ に対して 30 ビン ×5 ビンの大きさの 50 種類のフィルタ $\mathbf{w}_1 = \{\mathbf{w}_1^1, \cdots, \mathbf{w}_1^{50}\}$ を 1 フレーム，1 ビンずつずらしながら重畳することで，50 種類の特徴マップ $\mathbf{c} = \{\mathbf{c}^1, \cdots, \mathbf{c}^{50}\}$ を抽出する．この処理により，倍音構造の時間変化などの音声スペクトルの局所的な時間周波数パターンの特徴が抽出されることを期待する．$f$ 番目の特徴マップの抽出過程は以下のように表現される．

$$\mathbf{c}^f = \tanh\left(\mathbf{w}_1^f \otimes \mathbf{I} + \mathbf{b}_1^f\right), \tag{3.5}$$

ここで $\mathbf{c}^f$ は the $f$ 番目の特徴マップ，$\mathbf{w}_1^f$ および $\mathbf{b}_1^f$ はそれぞれ $f$ 番目の畳み込みフィルタおよびバイアスを示す．また $\otimes$ は，二次元畳み込みの操作を表す．次に，$\mathbf{c}^f$ を 5 ビン ×2 フレームの大きさの小領域に分割し，各領域における最大値を抽出することで，763 次元の特徴 $\mathbf{p}^f$ を抽出する．この処理は Max-pooling と呼ばれ，適用位置のずれに対して頑健なフィルタ $\mathbf{w}_1$ を効率的に学習することを可能とする．続けて，$\mathbf{p}^f$ をベクトル化し，それらを結合した 38150 次元のベクトル $\mathbf{p} = \{\mathbf{p}^1, \cdots, \mathbf{p}^{50}\}$ を作成し，以下の式により音声らしさを表す特徴 $\mathbf{h}$ を抽出する．

$$\mathbf{h}_j = \begin{cases} \tanh\left(\mathbf{w}_1^{(e)}\mathbf{p} + \mathbf{b}_1^{(e)}\right) & j = 1 \\ \tanh\left(\mathbf{w}_j^{(e)}\mathbf{h}_{j-1} + \mathbf{b}_j^{(e)}\right) & j = 2 \cdots L-1 \end{cases} \tag{3.6}$$

$$\mathbf{h} = \tanh\left(\mathbf{w}_L^{(e)}\mathbf{h}_{L-1} + \mathbf{b}_L^{(e)}\right) \tag{3.7}$$

ここで $\mathbf{w}_j^{(e)}$ および $\mathbf{b}_j^{(e)}$ は，$j$ 番目の隠れ層における重みおよびバイアスを表す

デコーディング処理では，下式により参照信号のスペクトルの局所パターン $\mathbf{O}$ を復元する．

$$\hat{\mathbf{h}}_j = \begin{cases} \tanh\left(\mathbf{w}_1^{(d)}\mathbf{h} + \mathbf{b}_1^{(d)}\right) & j = 1 \\ \tanh\left(\mathbf{w}_j^{(d)}\hat{\mathbf{h}}_{j-1} + \mathbf{b}_j^{(d)}\right) & j = 2 \cdots L-1 \end{cases} \tag{3.8}$$

$$\mathbf{O} = \mathbf{w}_L^{(d)}\hat{\mathbf{h}}_{L-1} + \mathbf{b}_L^{(d)} \tag{3.9}$$

ここで，$\mathbf{w}_j^{(d)}$ および $\mathbf{b}_j^{(d)}$ は，$j$ 番目の隠れ層における重みおよびバイアスを表す．こうして得られた $\mathbf{O}$ を用いて，重畳加算法により参照信号の対数パワースペクトルを計算する．

連想記憶のパラメータ $\{\ \mathbf{w}_1^1, \cdots, \mathbf{w}_1^{50}, \mathbf{b}_1^1, \cdots, \mathbf{b}_1^{50}, \mathbf{w}_1^{(e)}, \cdots, \mathbf{w}_L^{(e)}, \mathbf{b}_1^{(e)}, \cdots, \mathbf{b}_L^{(e)}, \mathbf{w}_1^{(d)}, \cdots, \mathbf{w}_L^{(d)}, \mathbf{b}_1^{(d)}, \cdots, \mathbf{b}_L^{(d)}\ \}$ は，**Full-Full** と同様に，誤差逆伝播法によって決定する．

この構造では，エンコーディング時にスペクトルの局所的な情報を陽に扱う．そのため，**Full-Full** と比較して音声の特徴を効果的に抽出できることが期待できる．一方で，max-pooling の処理により局所特徴の位置情報が失われてしまうため，参照信号は元音声らしさを失うという欠点がある．

### DCAE に基づく連想記憶モデル (Conv-Conv)

**Conv-Full** では，max-pooling によって局所特徴の位置情報が失われるという問題があった．そこで，max-pooling 時に選ばれた局所特徴の位置情報を保持しておき，その情報を元に参照信号をデコードする構造を開発した．図 3.4 に，この構造を示す．

エンコーディング処理では，**Conv-Full** と同様の方法で 2048 次元の特徴ベクトル $\mathbf{h}$ を抽出する．このとき，max-pooling で選択された位置情報 $\mathbf{i}$ を保持しておく．

デコーディング処理では，$\mathbf{p}$ に対応する 38150 次元の特徴 $\hat{\mathbf{p}}$ を，以下の式により復元する．

$$\hat{\mathbf{h}}_j = \begin{cases} \tanh\left(\mathbf{w}_1^{(d)}\mathbf{h} + \mathbf{b}_1^{(d)}\right) & j = 1 \\ \tanh\left(\mathbf{w}_j^{(d)}\hat{\mathbf{h}}_{j-1} + \mathbf{b}_j^{(d)}\right) & j = 2 \cdots L-1 \end{cases} \tag{3.10}$$

$$\hat{\mathbf{p}} = \mathbf{w}_L^{(d)}\hat{\mathbf{h}}_{L-1} + \mathbf{b}_L^{(d)} \tag{3.11}$$

ここで，$\mathbf{w}_j^{(d)}$ および $\mathbf{b}_j^{(d)}$ は，$j$ 番目の隠れ層における重みおよびバイアスを表す．次に $\hat{\mathbf{p}}$ を，50 個の 2 次元特徴 $\{\hat{\mathbf{p}}^1, \cdots, \hat{\mathbf{p}}^{50}\}$（$\hat{\mathbf{p}}^f \in \mathbb{R}^{109\times7}$）に分割する．こうして得られた $\hat{\mathbf{p}}^f$ に対してアップサンプリングを適用することで，$\mathbf{c}^f$ に対応する特徴 $\hat{\mathbf{c}}^f \in \mathbb{R}^{542\times14}$ を得る．アップサンプリングは，位置情報 $\mathbf{i}$ に基づき，Max-pooling で選択された位置には $\hat{\mathbf{p}}^f$ の値を，その他の位置には 0 を埋め合わせることで行う．最後に，$\hat{\mathbf{c}}^f$ に対して，30 ビン $\times$5 ビンの大きさの 50 種類のフィルタ

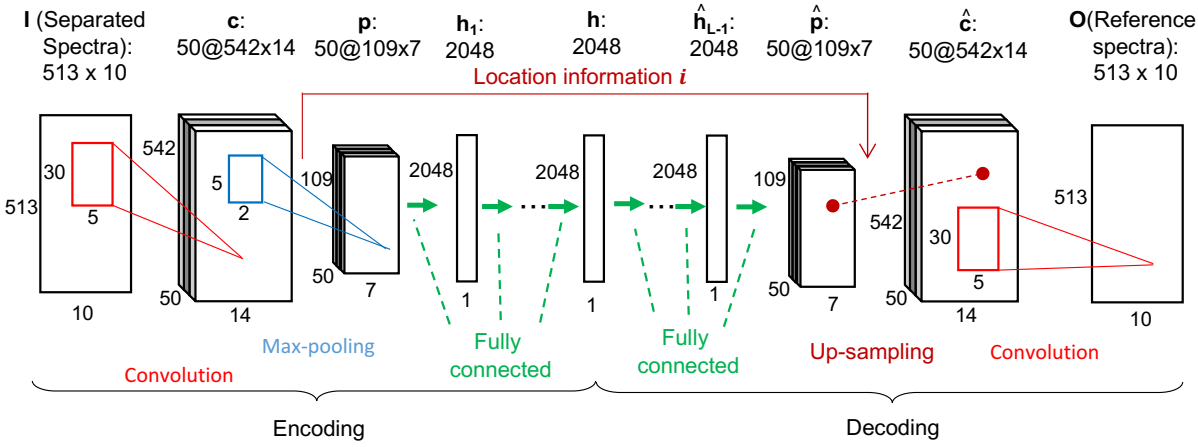



図 3.4: Architecture of the proposed denoising convolutional autoencoder (DCAE)-based associative memory model (AMM). $\mathbf{c}$ and $\hat{\mathbf{c}}$ denote the feature map and its estimate; $\mathbf{p}$ and $\hat{\mathbf{p}}$, the down-sampled feature map and its estimate; and $\mathbf{h}$, the feature vector extracted during encoding. In this architecture, $\hat{\mathbf{c}}$ is decoded using location information $\mathbf{i}$, which is lost during max-pooling in the encoding stage.

$\mathbf{w}_2 = \{\mathbf{w}_2^1, \cdots, \mathbf{w}_2^{50}\}$ を 1 フレーム，1 ビンずつずらしながら重畳することで，参照信号のスペクトルの時間周波数パターン $\mathbf{O}$ を決定する．この操作は以下の式のように表現できる．

$$\mathbf{O} = \sum_{f=1}^{50} \left( \mathbf{w}_2^f \otimes \hat{\mathbf{c}}^f + \mathbf{2}_4^f \right), \tag{3.12}$$

ここで，$\mathbf{w}_2^f$ および $\mathbf{b}_2^f$ は，$f$ 番目の畳み込みフィルタおよびバイアスを表す．こうして得られた $\mathbf{O}$ を用いて，重畳加算法により参照信号の対数パワースペクトルを計算する．

連想記憶のパラメータ $\{ \mathbf{w}_1^1, \cdots, \mathbf{w}_1^{50}, \mathbf{b}_1^1, \cdots, \mathbf{b}_1^{50}, \mathbf{w}_2^1, \cdots, \mathbf{w}_2^{50}, \mathbf{b}_2^1, \cdots, \mathbf{b}_2^{50}, \mathbf{w}_1^{(e)}, \cdots, \mathbf{w}_L^{(e)}, \mathbf{b}_1^{(e)}, \cdots, \mathbf{b}_L^{(e)}, \mathbf{w}_1^{(d)}, \cdots, \mathbf{w}_L^{(d)}, \mathbf{b}_1^{(d)}, \cdots, \mathbf{b}_L^{(d)} \}$ は，**Full-Full** と同様に，誤差逆伝播法によって決定する．

### 3.2.3 分離行列更新

$\mathbf{W}^{(0)}[k]$ および $\mathbf{Y}^{(0)}[k,l]$ をそれぞれ，既存の線形 BSS により初期化された分離行列およびその出力とする．ここでは，分離行列を更新するために，$\mathbf{Y}^{(0)}[k,l]$ を参照信号 $\hat{\mathbf{S}}[k,l]$ に近づける線形変換行列 $\mathbf{U}[k] \in \mathbb{C}^{N_s \times N_s}$ を考える．$\mathbf{U}[k]$ により変換されたスペクトルは，$\bar{\mathbf{Y}}[k,l]$ は以下のように書くことができる．

$$\bar{\mathbf{Y}}[k,l] = \mathbf{U}[k]\mathbf{Y}^{(0)}[k,l] = \mathbf{U}[k]\mathbf{W}^{(0)}[k]\mathbf{Z}[k,l]. \tag{3.13}$$

ここで $\bar{\mathbf{W}}[k] = \mathbf{U}[k]\mathbf{W}^{(0)}[k]$ とおくと，式（3.13）は $\bar{\mathbf{Y}}[k,l] = \bar{\mathbf{W}}[k]\mathbf{Z}[k,l]$ という線形変換で表現することができる．本研究では $\mathbf{U}[k]$ を求めた上で，$\mathbf{U}[k]$ と $\mathbf{W}^{(0)}[k]$ との積を計算することで，分離行列を更新することを試みる．

ここでは，最適な $\mathbf{U}[k]$ を求めるために，以下のコスト関数 $J[k]$ を設計した．

$$\begin{aligned} J[k] &= \frac{1}{N_l}\sum_{l=0}^{N_l-1}\left|\left|\log|\hat{\mathbf{S}}[k,l]|^2 - \log|\bar{\mathbf{Y}}[k,l]|^2\right|\right|^2 \\ &= \frac{1}{N_l}\sum_{l=0}^{N_l-1}\left|\left|\log|\hat{\mathbf{S}}[k,l]|^2 - \log|\mathbf{U}[k]\mathbf{Y}^{(0)}[k,l]|^2\right|\right|^2, \end{aligned} \tag{3.14}$$

ここで，$N_l$ はサンプル数を表す．式（3.14）は，参照信号と分離信号対数パワースペクトルの二乗誤差を表している．すなわち，$J[k]$ を最小にする $\mathbf{U}[k]$ は，分離信号の対数パワースペクトルを参照信号の対数パワースペクトルに近づける．$\mathbf{U}[k]$ の推定には，以下に示すような勾配法を用いた．

$$\mathbf{M}^{(0)}[k] = \mathbf{E}, \tag{3.15}$$

$$\mathbf{M}^{(n+1)}[k] \leftarrow \mathbf{M}^{(n)}[k] - \mu\frac{\mathbf{G}[k]}{||\mathbf{G}[k]||}, \tag{3.16}$$

$$\mathbf{G}[k] = \frac{\partial J[k]}{\partial \mathbf{U}^*[k]}, \tag{3.17}$$

$\mathbf{E}$，$\mu$，$n$，および $^*$ はそれぞれ，単位行列，学習係数，繰り返し回数および複素共役を表す．式（3.17）を求めるため，式（3.14）を以下のように書き換える．

$$\begin{aligned} J[k] &= \frac{1}{N_l}\sum_{l=0}^{N_l-1}\sum_{\imath=1}^{N_s}\left|\log|\hat{S}_i[k,l]|^2 - \log|\sum_{j=1}^{N_s}U_{ij}[k]Y_j^{(0)}[k,l]|^2\right|^2 \\ &= \frac{1}{N_l}\sum_{l=0}^{N_l-1}\sum_{\imath=1}^{N_s}\left|\log|\hat{S}_i[k,l]|^2 - \log|\bar{Y}_i^{(0)}[k,l]|^2\right|^2 \\ &= E\left(\sum_{\imath=1}^{N_s}\left|\log|\hat{S}_i[k,l]|^2 - \log|\bar{Y}_i^{(0)}[k,l]|^2\right|^2\right) \end{aligned} \tag{3.18}$$

ここで，$U_{ij}[k]$，$\hat{S}_i[k,l]$ および $Y_j^{(0)}[k,l]$ は，それぞれ $\mathbf{U}[k]$ の $(i,j)$ 番目の要素，$\hat{\mathbf{S}}_i[k,l]$ の $i$ 番目の要素，および $\hat{\mathbf{Y}}_j[k,l]$ の $j$ 番目の要素を表す．また $E(\cdot)$ は期待値を計算する関数を表す．式（3.18）を，$U_{ij}^*[k]$ で偏微分すると以下のようになる．

$$\begin{aligned} \frac{\partial J[k]}{\partial U_{ij}^*[k]} &= E\left(\sum_{l=0}^{N_l-1}\frac{\partial J[k]}{\partial \bar{Y}_i^*[k,l]}\frac{\partial \bar{Y}_i^*[k,l]}{\partial U_{ij}^*[k]}\right) \\ &= E\left(\frac{\partial}{\partial \bar{Y}_i^*[k,l]}\left(-2\log|\hat{S}_i[k,l]|^2\log|\bar{Y}_i[k,l]|^2\right.\right. \\ &\qquad \left.\left.+\left(\log|\bar{Y}_i[k,l]|^2\right)^2 + C\right)\frac{\partial \bar{Y}_i^*[k,l]}{\partial U_{ij}^*[k]}\right) \end{aligned}$$

$$= \quad -E\left(\frac{2}{\bar{Y}_i^*[k,l]}\left(\log|\hat{S}_i[k,l]|^2 - \log|\bar{Y}_i[k,l]|^2\right)\frac{\partial \bar{Y}_i^*[k,l]}{\partial U_{ij}^*[k]}\right) . \tag{3.19}$$

$C$ は $Y_j^{(0)*}[k,l]$ と無関係の項を表す．

また $\partial \bar{Y}_i^*[k,l]/\partial U_{ij}^*[k]$ は以下のように書くことが出来る．

$$\frac{\partial \bar{Y}_i^*[k,l]}{\partial U_{ij}^*[k]} = Y_j^{(0)*}[k,l] . \tag{3.20}$$

そこで式 (3.20) を式 (3.19) に代入し，以下のように勾配を計算する．

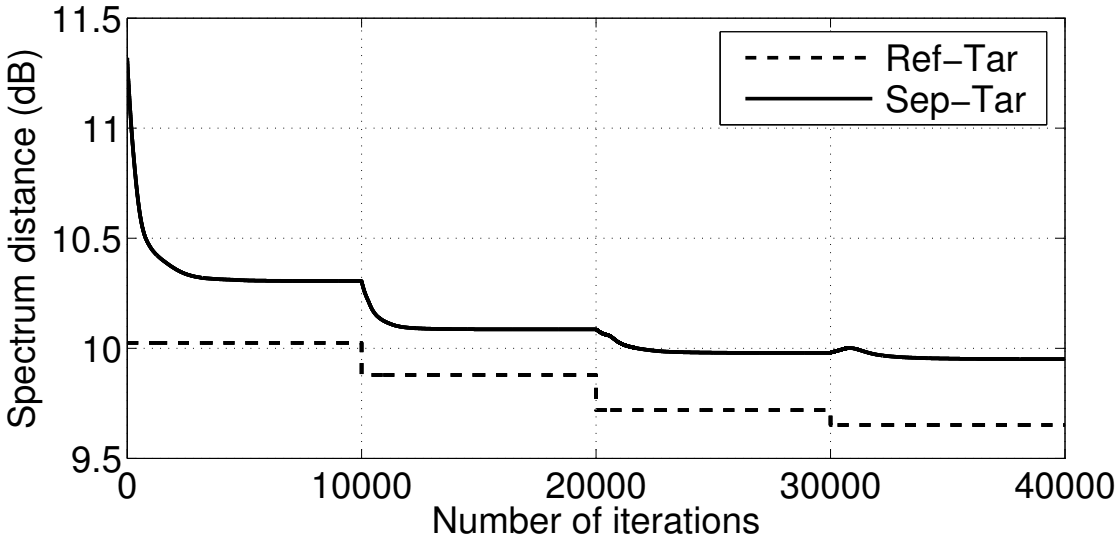
$$\frac{\partial J[k]}{\partial U_{ij}^*[k]} \quad = \quad -2E\left(\frac{Y_j^{(0)*}[k,l]}{\bar{Y}_i^*[k,l]}\left(\log|\hat{S}_i[k,l]|^2 - \log|\bar{Y}_i[k,l]|^2\right)\right) . \tag{3.21}$$

### 3.2.4 独立性に基づく方法と提案法との関係

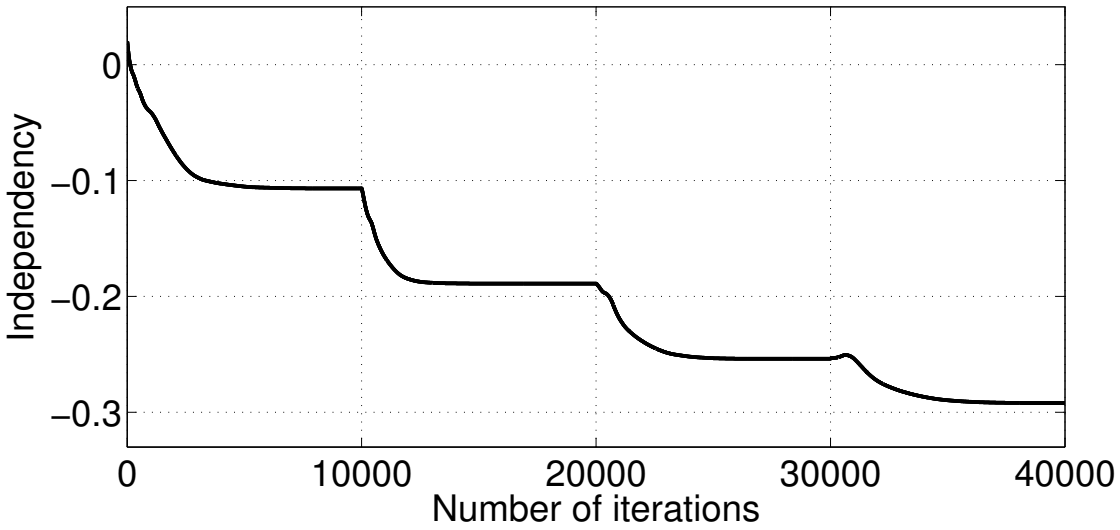
音源の独立性のみを考慮しながら音源分離を行う従来法に対する，音声らしさを考慮しながら音源分離を行う提案法の有効性を調査するために，二話者同時発話音声分離実験を行なった．

図 3.5(a) に分離行列更新に伴う分離信号または参照信号と音源信号とのスペクトル誤差の変化を，図 3.5(b) に分離信号間の独立性の変化を示す．分離信号間の独立性の計算には，音源の分布としてラプラス分布を仮定し，式（2.16）により求めた．実験では分離行列の初期値は IVA で決定しており，更新回数が 0 回のときの性能が従来法の性能を表す．分離行列は DCAE によって推定された参照信号を用いて逐次更新されるが，参照信号の更新は 10000 回分離行列が更新されるたびに行った．図では，参照信号が更新されるたびに音源信号に近づくにつれて，分離信号と音源信号とのスペクトル距離だけでなく，音源の独立性が改善されていく様子を確認することができる．また，参照信号が更新されるタイミングでスペクトル距離・独立性ともに大幅に改善され，参照信号が 4 回更新されたタイミングで分離行列の更新が収束している様子も確認できる．これらのことより，音源の独立性のみに基づき分離行列を最適化する従来の方法では局所解に陥ってしまうのに対し，提案法は参照信号を更新することで局所解の影響を取り除きながら，音源間の独立性を増加させる効果があるものと示唆する．

図 3.6 に，提案法における参照信号の更新に伴う SIR(signal-to-interference ratio) と SDR(signal-to-interference ratio) の変化を示す．SDR は分離処理によって生じた歪に対する目的音源のパワー比を，SIR は分離信号における目的音以外の成分に対する目的音源のパワー比を表す [74]．参照信号 $\hat{\mathbf{S}}^{(n)}$ は，連想記憶により目的音源以外の成分を取り除くように推定しているため分離信号 $\hat{\mathbf{Y}}^{(n)}$ よりも SIR が高くなるが，歪が生じてしまうため SDR が劣化してしまう．図では，参照信号の出




(a) Spectral distance




(b) Independency

図 3.5: Example of relation between cost function and number of iterations in separation matrix optimization. In (a), dashed line (Ref-Tar) represents spectrum distance between reference signal and target signal (i.e. ground-truth source signal); solid line (Sep-Tar) represents that between separated signal (i.e. estimated source signal) and target signal. In (b), cost function of Eq. (2.16) averaged over frequency bin is shown.

力をそのまま用いるのではなく分離行列の最適化に用いることで，SIR，SDR 共に改善されていく様子を確認することができる．

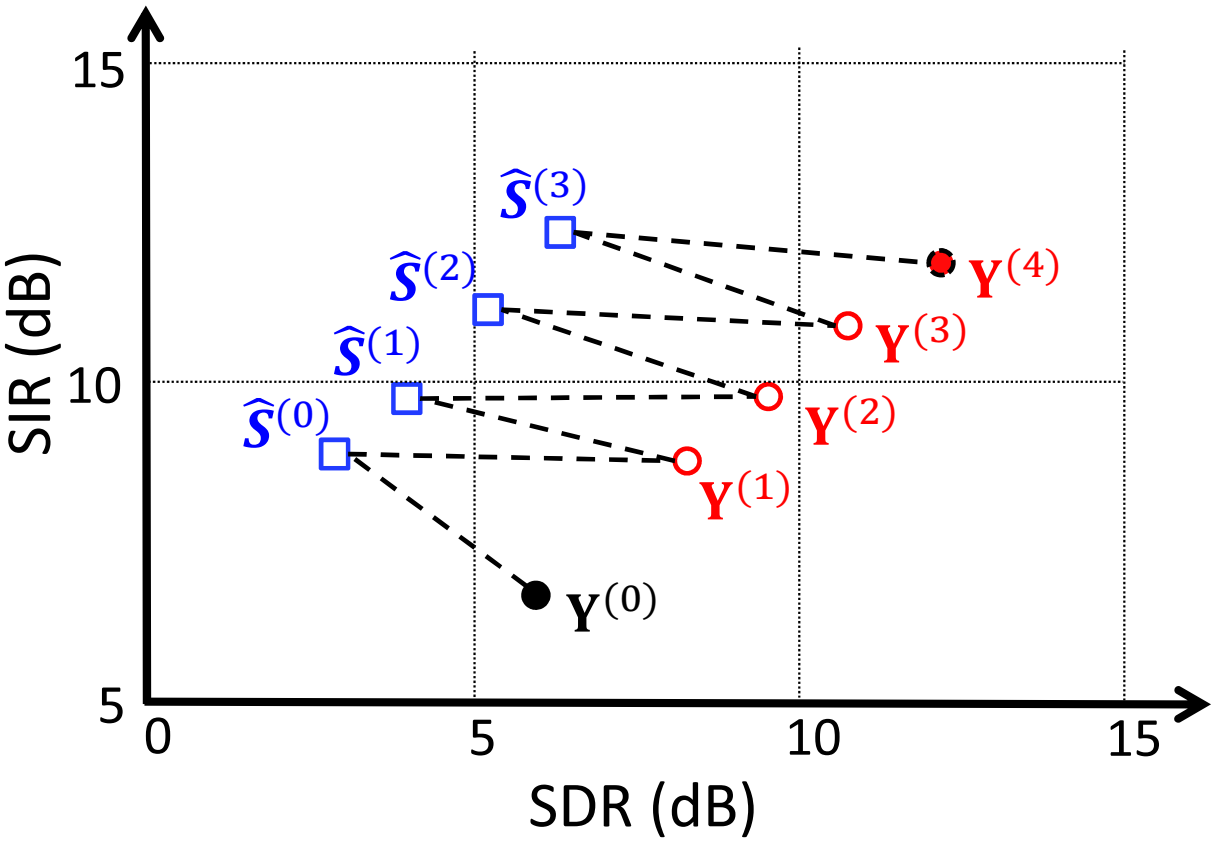



図 3.6: Example of relation between signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR). $\hat{\mathbf{S}}^{(n)}$ represents outputs of DCAE-based AMM in $n$-th reference signal estimation. $\mathbf{Y}^{(n)}$ represents outputs of separation matrix optimized using $\hat{\mathbf{S}}^{(n-1)}$.

## 3.3 音源分離実験

### 3.3.1 連想記憶モデルの学習

連想記憶の入力は，妨害音の消し残しや分離に生じた歪などのノイズが含まれることが予想される．一方で提案法により分離行列を更新していくとノイズが抑圧されて，ノイズの少ないパターンが入力されることが期待できる．また，連想記憶の出力はノイズを含まないパターンであることが期待される．したがって本研究で用いる連想記憶は，以下の性質を満たすことが望ましい．

- ノイズを含まないパターンを入力した場合は，入力パターンがそのまま出力される．
- ノイズを含むパターンを入力した場合は，ノイズが取り除かれたパターンが出力される．

そこで，連想記憶のパラメータは，クリーン信号を入力データ・教師データの双方に用いるデータ対（**clean-clean**），ノイズを含む分離信号を入力データ，ノイズが取り除かれたクリーン信号を教師データの用いるデータ対（**processed-clean**）の 2 種類のデータ対を用いて決定した．一般にニューラルネットワークの中間層の数が増えると，勾配消失（vanishing gradient）の問題により学習が困難となることが知られている．そこで本研究では，Stacking autoencoder に基づく貪欲学習を採用した [24]．このとき学習時のミニバッチサイズは 256，学習係数は初期値を 0.1 とし，new-bob 法により動的に制御した．以下に，**clean-clean** および **processed-clean** の詳細を示す．

**clean-clean**

クリーン信号の音声の対数パワースペクトルを入力データ・教師データ双方に用いる．ここでは，ATR 音素バランス文 [75] のセット B に含まれる 1,800 発話 (女性 4 話者 × 各話者 450 発話) を用いた．信号のサンプリング周波数は 16 kHz であり，STFT のフレーム長およびフレームシフトは，それぞれ 1024 サンプル（64 ms）および 256 サンプル（16 ms）とした．

**processed-clean**

ノイズを含む分離音声の対数パワースペクトルを入力データ，対応する目的信号の対数パワースペクトルを教師データに用いる．分離音声は，ATR 音素バランス文のセット B に含まれる女性 4 話者から得られる二話者同時発話音声に対して，補助関数法に基づく IVA [76] を適用することで作成した．4 話者から 12 組の話者対を作成し，9 話者対を学習セットに，残りの 3 話者対を開発セットとして用いた．開発セットは，連想記憶学習時の早期終了に用

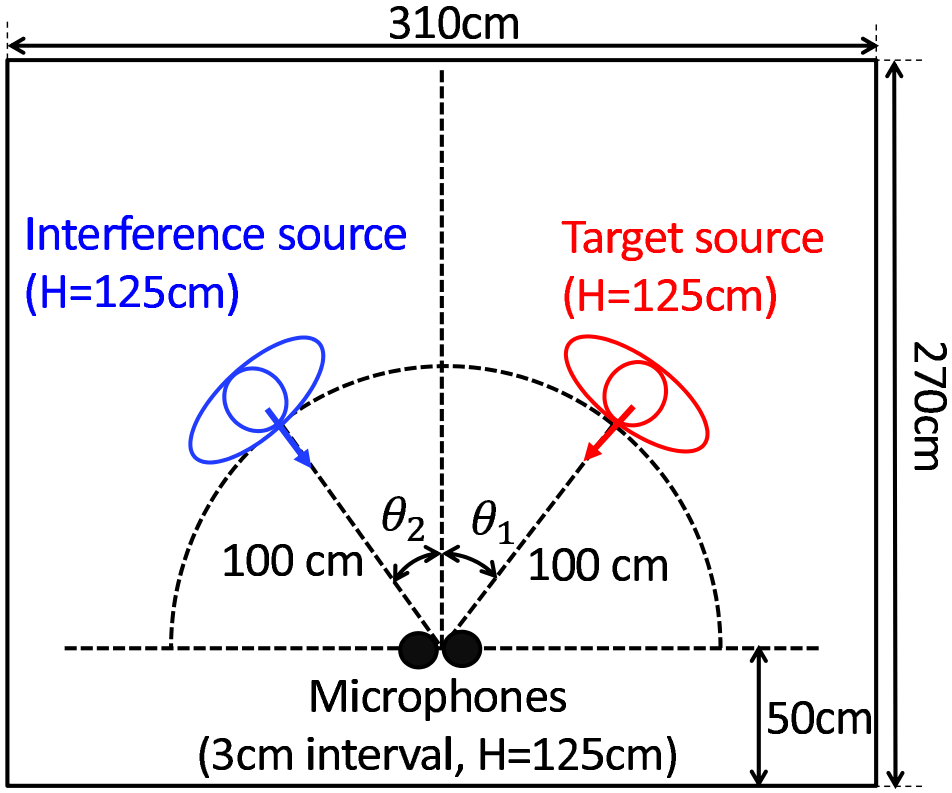



図 3.7: Experimental environment with two microphones and two sources, where $\theta_1$ and $\theta_2$ denote the directions of the target and interference sources, respectively.

表 3.1: Direction of target and interference sources.

| **data set** | **source direction of** $(\theta_1, \theta_2)$ |
|---|---|
| training | (-15,15), (-45,45), (-75,75), (-90,90) |
| development | (-60,60) |
| testing | (-30,30), (-30,0), (0,-30), (0,30), (30,0), (30,-30) |

いる．図 3.7 に示す環境において，各話者対が同時に発話することを想定し，ドライソースに遅延を加えることで，同時発話音声を合成した．このとき，音源の方向を表 3.1 に示す．学習セットでは話者毎に 50 発話を，開発セットでは話者毎に 52 発話をドライソースとして用いた．すなわち学習セットには，3,600 発話（9 話者対 × 2 話者 × 50 発話 × 4 方向）の分離音声を入力データ，対応する 3,600 発話のドライソースを教師データとして用いた．また開発セットには，312 発話（3 話者対 × 2 話者 × 52 発話 × 1 方向）の分離音声を入力データ，対応する 312 発話のドライソースを教師データとして用いた．

### 3.3.2 連想記憶モデルのパラメータ $L$ の決定

連想記憶モデルにおける全接続層の層数と提案法の性能との関係を調査し，最適な層数を決定した．

評価データは，シミュレーションにより生成した，インパルス応答は，図 3.7 に示す環境で収録した．このとき，音源位置 ($\theta_1$,$\theta_2$) は (-30,0) とした．本実験では，後部残響の影響を取り除くために，直接波到来時刻から初期反射到来時刻までの区間のみを切り出して用いた．日本語新聞読み上げコーパス（Japanese Newspaper Article Sentences: JNAS） [77] より無作為に選択した音声にインパルス応答を畳み込み，SNR が 0 dB となるように混合することで 30 発話（10 話者対 × 各組 3 発話）の二話者同時発話音声を作成した．この二話者同時発話音声に対して，以下の分離手法を適用し、その性能を評価した．

**IVA(Baseline)** ：IVA に基づく音源分離

**IVA-AMM(FF)-SMO** ：DAE に基づく連想記憶モデル（**FF**）を用いて推定した分離行列による分離

**IVA-AMM(CF)-SMO** ：CNN に基づく連想記憶モデル（**CF**）を用いて推定した分離行列による分離

**IVA-AMM(CC)-SMO** ：DCAE に基づく連想記憶モデル（**CC**）を用いて推定した分離行列による分離

提案する分離行列推定法（**IVA-AMM(FF)**，**IVA-AMM(CF)**，**IVA-AMM(CC)**）では，IVA により求められた分離行列を初期値として用いた．参照信号の更新回数の上限は 30 回，分離行列の更新回数の上限は 5000 回とした．また，式 (3.16) における学習係数 $\mu$ は初期値を 0.0001 とし，new-bob 法により動的に制御した．

分離性能の評価尺度としては，SIR（signal-to-interference ratio）および SDR（signal-to-distortion ratio）を用いた．SIR は分離信号に含まれる目的音源以外の成分に対する目的音源の成分の比を表し，SDR は分離処理により生じた歪に対する目的音源の成分の比を表す．以下に SIR，SDR の計算式を示す．

SIR [dB]:

$$\frac{1}{N_s N_k}\sum_{i=1}^{N_s}\sum_{k=0}^{N_k-1} 10\log_{10}\frac{\sum_{l=0}^{N_l-1}|S_i[k,l]|^2}{\sum_{j=1}^{N_s}\sum_{k=0}^{N_k-1}(1-\delta_{ij})|Y_{ij}[k,l]|^2} \tag{3.22}$$

表 3.2: Relation between separation performance and parameter of associative memory model $L$.

(a) SIR [dB]

| | $L$=0 | $L$=1 | $L$=2 | $L$=3 |
|---|---|---|---|---|
| **IVA(Baseline)** | 32.76 | — | — | — |
| **IVA-AMM(FF)-SMO** | — | **32.34** | 32.20 | 32.25 |
| **IVA-AMM(CF)-SMO** | **33.48** | 32.72 | 32.29 | 32.28 |
| **IVA-AMM(CC)-SMO** | **34.00** | 33.34 | 32.94 | 32.88 |

(b) SDR [dB]

| | $L$=0 | $L$=1 | $L$=2 | $L$=3 |
|---|---|---|---|---|
| **IVA(Baseline)** | 10.12 | — | — | — |
| **IVA-AMM(FF)-SMO** | — | **12.01** | 11.96 | 11.94 |
| **IVA-AMM(CF)-SMO** | 11.44 | 11.93 | **12.39** | 12.20 |
| **IVA-AMM(CC)-SMO** | 8.88 | 11.75 | **12.12** | 11.85 |

SDR [dB]:

$$\frac{1}{N_s N_k}\sum_{i=1}^{N_s}\sum_{k=0}^{N_k-1} 10\log_{10}\frac{\sum_{l=0}^{N_l-1}|S_i[k,l]|^2}{\sum_{l=0}^{N_l-1}\left(|S_i[k,l]|-|Y_{ii}[k,l]|\right)^2} \tag{3.23}$$

ただし，

$$Y_{ij}[k,l]=\sum_{m=1}^{N_m} W_{im}[k]H_{mj}[k]S_j[k,l] \tag{3.24}$$

$N_l$，$N_k$ は，それぞれフレーム数，周波数ビン数を表す．また，$\delta_{ij}$ は，$i=j$ のときに 1，$i\neq j$ のときに 0 を返す．

表 3.2 に，SIR および SDR と連想記憶モデルのパラメータ $L$ との関係を示す．**IVA-AMM(CC)-SMO** において，$L$=0 は Convolutional autoencoder を用いた場合と等価になる．$L$ が大きくなるほど，連想記憶モデルが深層化の影響が表れる．まず SIR の観点で比較すると，**IVA-AMM(FF)-SMO** および **IVA-AMM(CF)-SMO** は，**IVA** よりも劣化する傾向があるものの，**IVA-AMM(CC)-SMO** は，**IVA** の性能を上回る傾向がみられた．また，深層化するにつれて SIR は劣化する傾向がみられた．次に SDR の観点で比較すると，提案法はどの連想記憶を用いた場合においても，**IVA** の性能を上回る傾向がみられた．また，**IVA-AMM(FF)-SMO** は，深層化の効果がみられなかったものの，**IVA-AMM(CF)-SMO** および **IVA-AMM(CC)-SMO** においては，$L=2$ のときに SDR が最大となった．

以上のことを踏まえ，以降の実験では，**IVA-AMM(FF)-SMO** では $L=1$，**IVA-AMM(CF)-SMO** および **IVA-AMM(CC)-SMO** では $L=2$ を採用した．

### 3.3.3 無響環境における二話者同時発話音声分離の性能評価

図 3.7 に示すような環境を想定し，二話者同時発話音声に対する分離性能を調査した．

評価データは，シミュレーションにより生成した．インパルス応答は，図 3.7 に示す環境で収録した．このとき設定した音源位置 ($\theta_1$,$\theta_2$) を表 3.1 に示す．本実験では，後部残響の影響を取り除くために，直接波到来時刻から初期反射到来時刻までの区間のみを切り出して用いた．JNAS より無作為に選択した音声にインパルス応答を畳み込み，SNR が 0 dB となるように混合することで 180 発話（10 話者対 × 各組 3 発話 × 6 方向）の二話者同時発話音声を作成した．この二話者同時発話音声に対して，以下の分離手法を適用し、その性能を評価した．

**IVA(Baseline)** ：IVA に基づく音源分離

**IVA-AMM(FF)-SMO** ：DAE に基づく連想記憶モデルを用いて推定した分離行列による分離

**IVA-AMM(CF)-SMO** ：CNN に基づく連想記憶モデルを用いて推定した分離行列による分離

**IVA-AMM(CC)-SMO** ：DCAE に基づく連想記憶モデルを用いて推定した分離行列による分離

提案する分離行列推定法（**IVA-AMM(FF)**，**IVA-AMM(CF)**，**IVA-AMM(CC)**）では，IVA により求められた分離行列を初期値として用いた．参照信号の更新回数の上限は 30 回，分離行列の更新回数の上限は 5000 回とした．また，式 (3.16) における学習係数 $\mu$ は初期値を 0.0001 とし，new-bob 法により動的に制御した．分離性能の評価尺度は，3.3.2 と同様に，SIR および SDR を用いた．

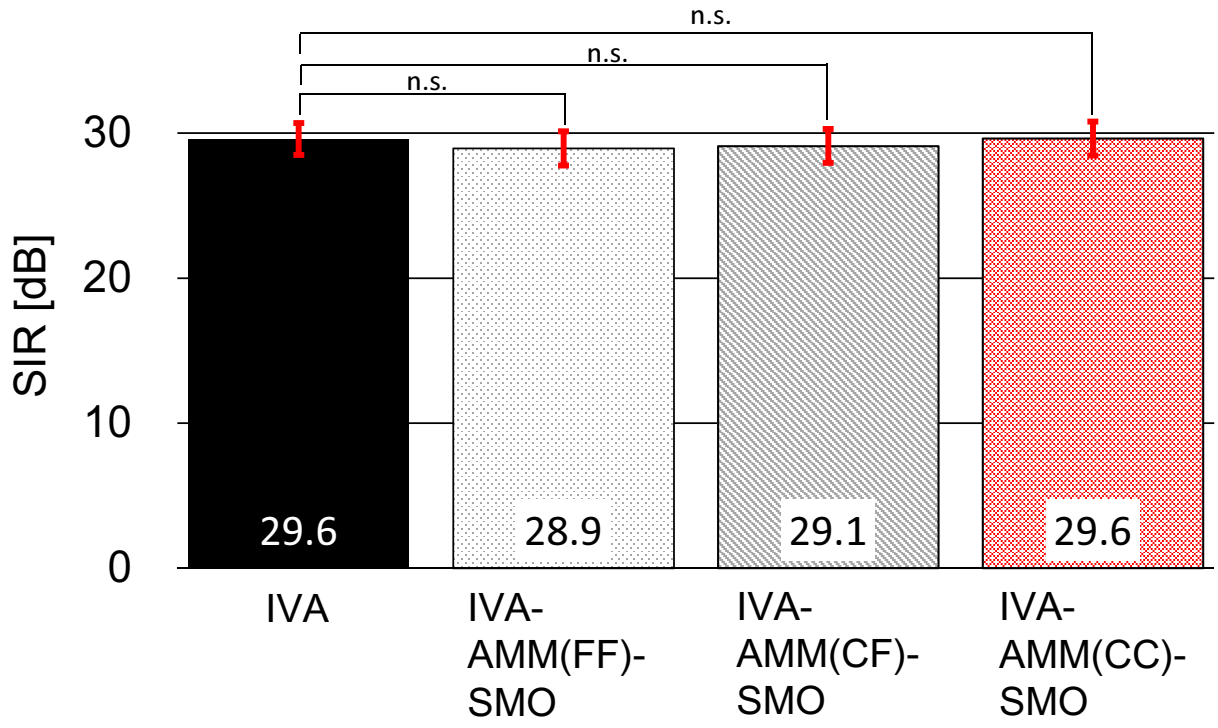
図 3.8(a) に，SIR を計算した結果を示す．連想記憶を用いて分離行列の補正を行った **IVA-AMM(FF)-SMO**，**IVA-AMM(CF)-SMO** および **IVA-AMM(CC)-SMO** は，補正を行わなかった **IVA** とほぼ同等の性能を示した．有意水準 10%で Welch の t 検定を行ったところ，各手法間で有意な差を確認することはできなかった．図 3.8(b) に，SDR を計算した結果を示す．連想記憶を用いて分離行列の補正を行った **IVA-AMM(FF)-SMO**，**IVA-AMM(CF)-SMO** および **IVA-AMM(CC)-SMO** は，補正を行わなかった **IVA** の性能を上回った．有意水準 1%で t 検定を行ったところ，**IVA** とその他の分離手法とで，有意な差を確認することができた．このことより，提案法は IVA により生じる歪を低減する効果があると言える．提案法において用いた連想記憶のモデルの違いによる効果を検証するため，有意水準 1%で t 検定を行ったところ，**IVA-AMM(FF)-SMO** と **IVA-AMM(CF)-SMO** の間で有意な差を確認することができたものの，その他のモデル間では有意な差を確認することができなかった．

表 3.3: Short time objective intelligibility (STOI) measure averaged over 360 utterances with their standard deviations.

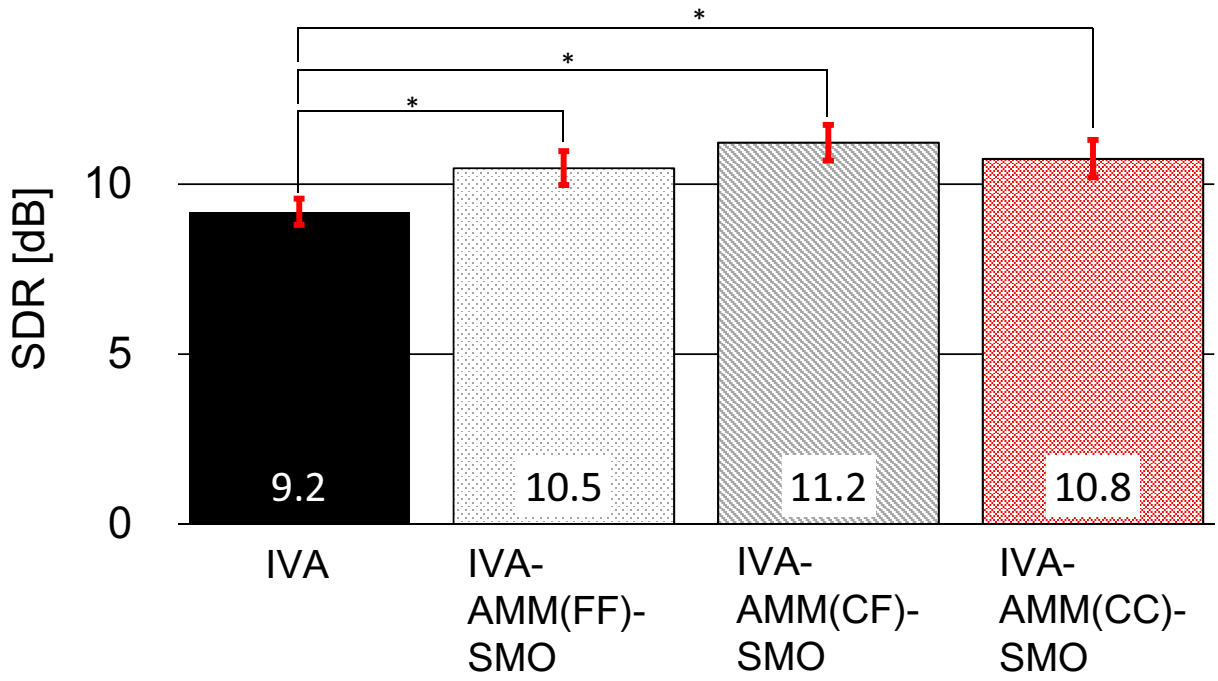
| BSS methods | mean | stddev |
|---|---|---|
| **IVA(Baseline)** | 0.94 | 0.11 |
| **IVA-AMM(FF)** | 0.82 | 0.09 |
| **IVA-AMM(CF)** | 0.81 | 0.10 |
| **IVA-AMM(CC)** | 0.83 | 0.11 |
| **IVA-AMM(FF)-SMO** | 0.94 | 0.11 |
| **IVA-AMM(CF)-SMO** | 0.94 | 0.12 |
| **IVA-AMM(CC)-SMO** | 0.92 | 0.12 |

この結果を，各手法により形成された指向特性の観点から考察する．図 3.9 に，**IVA** および **IVA-AMM(CC)-SMO** により形成された指向特性の例を示す．ここでは，0 ° 方向に妨害音源を配置した場合の例を示す．今回の条件においては，**IVA** および **IVA-AMM(CC)-SMO** ともに，0 ° 方向に死角を形成することができていることが確認できる．すなわち今回の実験条件では，**IVA** の時点で十分に死角を妨害音方向に向けることができており，連想記憶による補正を適用しても SIR の性能の改善に大きく寄与しなかったものであることが，考えられる．また，**IVA** では 1000 Hz 付近において歪が生じているのに対して，**IVA-AMM(CC)-SMO** ではその影響が緩和されている様子がみられる．このことからも，提案法は歪を低減する効果があることが確認できる．

最後に，連想記憶モデルの出力をそのまま使うのではなく，線形分離行列を推定する際の手がかりとして用いることの利点を，明瞭性の観点から述べる．明瞭性の尺度として主観評価値と相関の高いと言われている STOI(Short time objective intelligibility measure) [78] を用いた．表 3.3 に，STOI を計算した結果を示す．**IVA-AMM(FF)**，**IVA-AMM(CF)**，**IVA-AMM(CC)** はそれぞれ，DAE，CNN，CAE の出力を分離信号として用いた場合の結果を示す．連想記憶モデルの出力をそのまま用いた場合は，**IVA** に対して STOI の値が 0.13 程度劣化する．一方で，連想記憶モデルの出力を用いて推定した分離行列の出力は，**IVA** とほぼ同等の性能であり，0.02 程度の誤差であった．なお，有意水準 1%で t 検定を行ったところ，**IVA** と **IVA-AMM(CC)-SMO** とで有意な差を確認することはできなかった．以上のことより，連想記憶モデルの出力を分離信号として用いるよりも，分離行列の推定に用いた方が明瞭性の観点において有効であると言える．

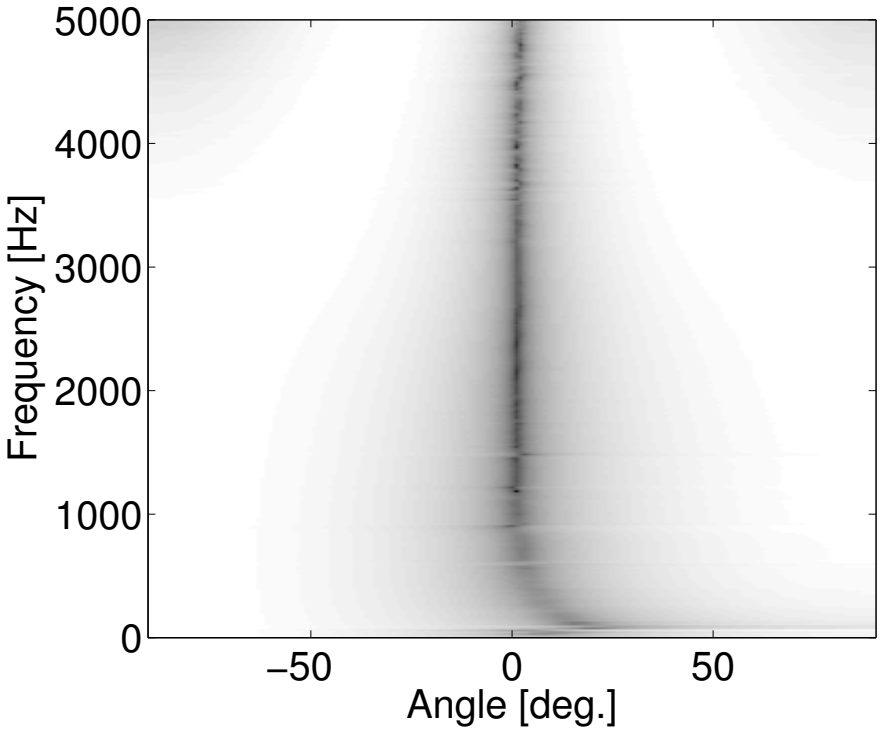



(a) SIR

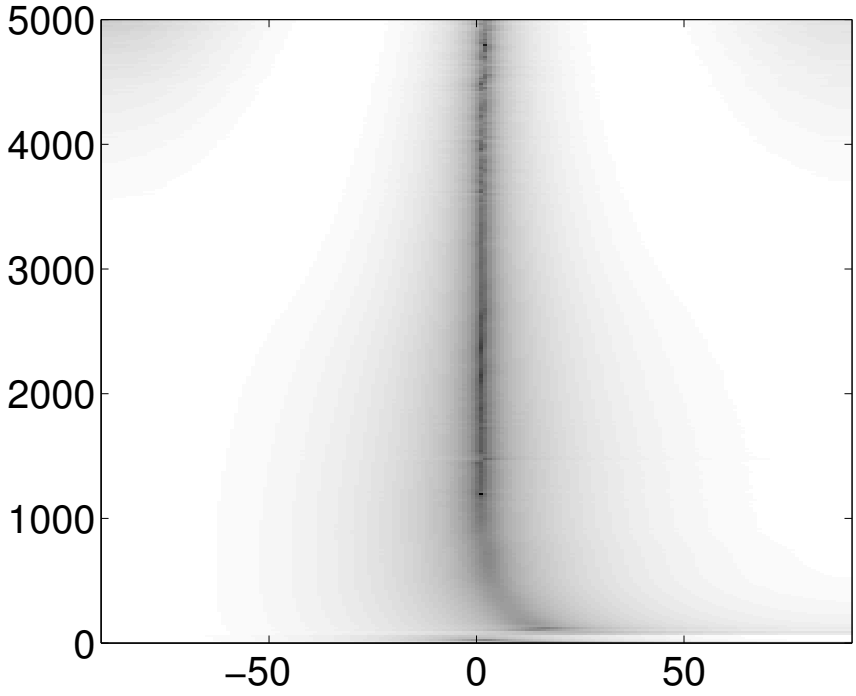



(b) SDR

図 3.8: Simultaneous speech separation performance with the proposed separation matrix optimization averaged over 180 utterance pairs along with their 99% confidence interval. * denotes significant difference of 1%; n.s. denotes no significant difference.




(a) **IVA**




(b) **IVA-AMM(CC)-SMO**

図 3.9: Example of directivity pattern of (a) separation matrix optimized with independent vector analysis and (b) separation matrix optimized with proposed method. In this case, disturbance is placed at the direction of 0 °.

### 3.3.4　エコー信号除去における分離性能の評価

音源に対して独立とは言い難い信号が含まれる場合における提案法の有効性を検証した．具体的には，図 3.7 に示すような環境において，目的音源として信号 $x(t)$ を，妨害音源として，$x(t)$ に遅延 $\tau$ を加えたエコー信号 $x(t+\tau)$ を同時に発音し，それぞれを分離した際の性能を調査した．

評価データはシミュレーションにより生成した．インパルス応答は 3.3.3 で用いたものと同じものを用いた．音源信号 $x(t)$ として，JNAS に含まれる 20 発話（女性 20 話者 $\times$ 各話者 1 発話）をとして用いた．また，エコー信号は，$x(t)$ の先頭に $\tau$ ms の無音区間を挿入することで生成した．ここでは $\tau$ が，50 または 25 となる場合の性能を調査した．$\tau$ が短くなればなるほど，音源信号とエコー信号との相関が高くなるため，分離が難しくなることが想定される．音源信号およびエコー信号にインパルス応答を畳み込み，SNR が 0 dB となるように重畳することで，混合信号 120 発話 (女性 20 話者 $\times$6 条件) を作成した．こうして得た混合信号に対して以下の分離手法を適用し，性能を調査した．

**IVA(Baseline)** ：IVA に基づく音源分離.

**IVA-AMM(FF)** ：DAE に基づく連想記憶モデルを DAE として用いたもの．入力として IVA の出力を与えたの出力を分離信号として用いる.

**IVA-AMM(CF)** ：CNN に基づく連想記憶モデルを DAE として用いたもの.

**IVA-AMM(CC)** ：DCAE に基づく連想記憶モデルを用いた音源分離.

**IVA-AMM(FF)-SMO** ：DAE に基づく連想記憶モデルを用いて推定した分離行列による分離.

**IVA-AMM(CF)-SMO** ：CNN に基づく連想記憶モデルを用いて推定した分離行列による分離.

**IVA-AMM(CC)-SMO** ：DCAE に基づく連想記憶モデルを用いて推定した分離行列による分離.

提案する分離行列推定法（**IVA-AMM(FF)-SMO**, **IVA-AMM(CF)-SMO**, **IVA-AMM(CC)-SMO**）では，IVA により求められた分離行列を初期値として用いた．参照信号の更新回数の上限は 30 回，分離行列の更新回数の上限は 5000 回とした．また，式 (3.16) における学習係数 $\mu$ は初期値を 0.0001 とし，new-bob 法により動的に制御した．

分離性能は，3.3.2 で説明した SIR および SDR，および，Vincent ら [74] が提案した signal-to-interference ratio (SIR)，signal-to-distortion ratio (SDR)，signal-to-artifacts ratio (SAR) により評価した．これらの尺度は，分離手法に依存せずに評価が可能であり，連想記憶の出力を分離

信号として用いる場合と，分離行列の推定に用いた場合とでの性能比較が可能となる．また，音源分離のコンペティションである SiSec [79] にて用いられている．以降では，これらの尺度の計算方法を説明する．まず，分離信号 $\hat{s}_i(t)$ を以下のように分解する．

$$\hat{s}_i(t) = \bar{s}_i(t) + e^{(spat)}(t) \quad + e^{(interf)}(t) + e^{(noise)}(t) + e^{(artif)}(t) \tag{3.25}$$

$\bar{s}(t)$ は目的音源の成分を表し，$e^{(spat)}(t)$，$e^{(interf)}(t)$，$e^{(noise)}(t)$，$e^{(artif)}(t)$ はそれぞれ，分離信号と音源信号との誤差成分（空間的な歪），妨害音源の成分，マイクロホンで観測されるノイズ成分，分離により生じるノイズ成分を表す．これらは，以下のように計算される．

$$\bar{s}_i(t) = \mathbf{P}_{s_i}\hat{s}_i(t) \tag{3.26}$$

$$e^{(spat)} = \hat{s}_i(t) - \bar{s}_i(t) \tag{3.27}$$

$$e^{(interf)} = \mathbf{P}_{\mathbf{s}}\hat{s}_i(t) - \bar{s}_i(t) \tag{3.28}$$

$$e^{(noise)} = \mathbf{P}_{\mathbf{s},\mathbf{n}}\hat{s}_i(t) - \mathbf{P}_{\mathbf{s}}\hat{s}_i(t) \tag{3.29}$$

$$e^{(artif)} = \hat{s}_i(t) - \mathbf{P}_{\mathbf{s},\mathbf{n}}\hat{s}_i(t) \tag{3.30}$$

$$\mathbf{P}_{s_i} = \Pi\{s_i\} \tag{3.31}$$

$$\mathbf{P}_{\mathbf{s}} = \Pi\{s_1, \cdots, s_{N_s}\} \tag{3.32}$$

$$\mathbf{P}_{\mathbf{s},\mathbf{n}} = \Pi\{s_1, \cdots, s_{N_s}, n_1, \cdots, n_{N_m}\} \tag{3.33}$$

$s_i$ および $n_i$ は，それぞれ，音源信号およびマイクロホンで観測されるのノイズを表す．また $\mathbf{P}_{\mathbf{a}} = \Pi\{a_1, a_2, \cdots, a_N\}$ は，ベクトル $\{a_1, a_2, \cdots, a_N\}$ で張られる部分空間への直交射影作用素を表す．このうち，$e^{(spat)}$ は音質の違いを表すものであり，音源分離性能には寄与しない成分であると考える．次にこれらの要素を用いて，SIR，SDR および SAR を以下の式により求める．

$$\text{SIR [dB]} := 10\log_{10}\frac{||\hat{s}_i(t) + e^{(spat)}(t)||^2}{||e^{(interf)}||^2} \tag{3.34}$$

$$\text{SDR [dB]} := 10\log_{10}\frac{||\hat{s}_i(t) + e^{(spat)}(t)||^2}{||e^{(interf)} + e^{(noise)} + e^{(artif)}||^2} \tag{3.35}$$

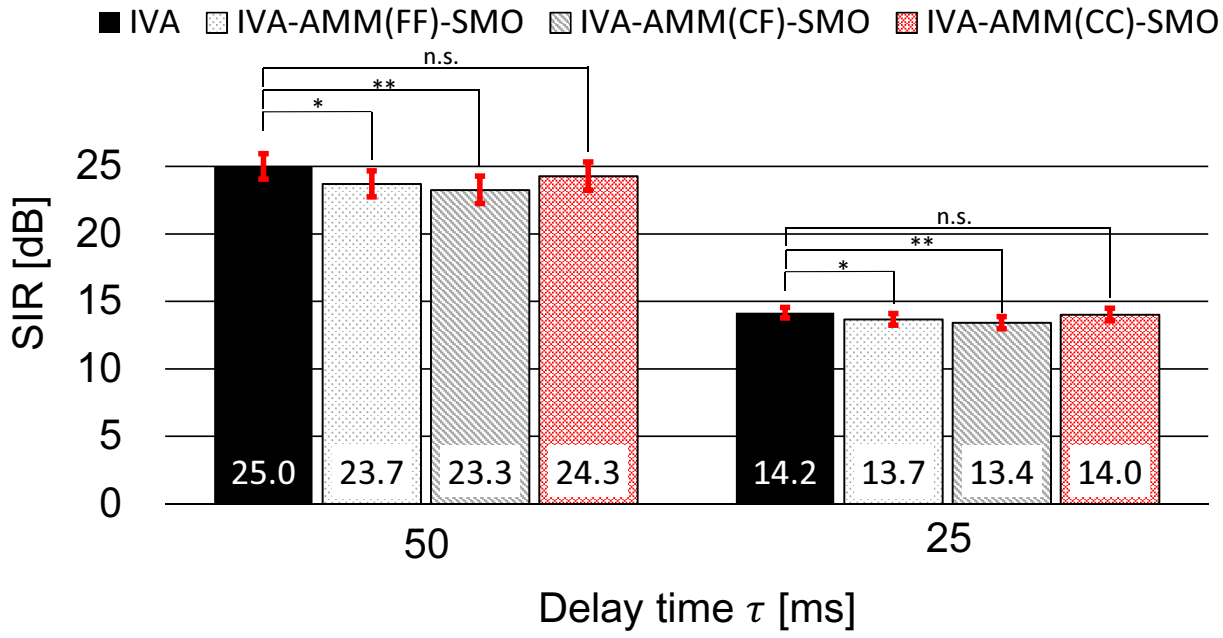
$$\text{SAR [dB]} := 10\log_{10}\frac{||\hat{s}_i(t) + e^{(spat)}(t) + e^{(interf)} + e^{(noise)}||^2}{||e^{(artif)}||^2} \tag{3.36}$$

すなわち，SIR は妨害音源に対する目的音源の成分の比率，SDR は目的音源以外の成分に対する目的音源の成分の比率，SAR は分離により生じた歪に対するその他の成分の比率をそれぞれ表す．

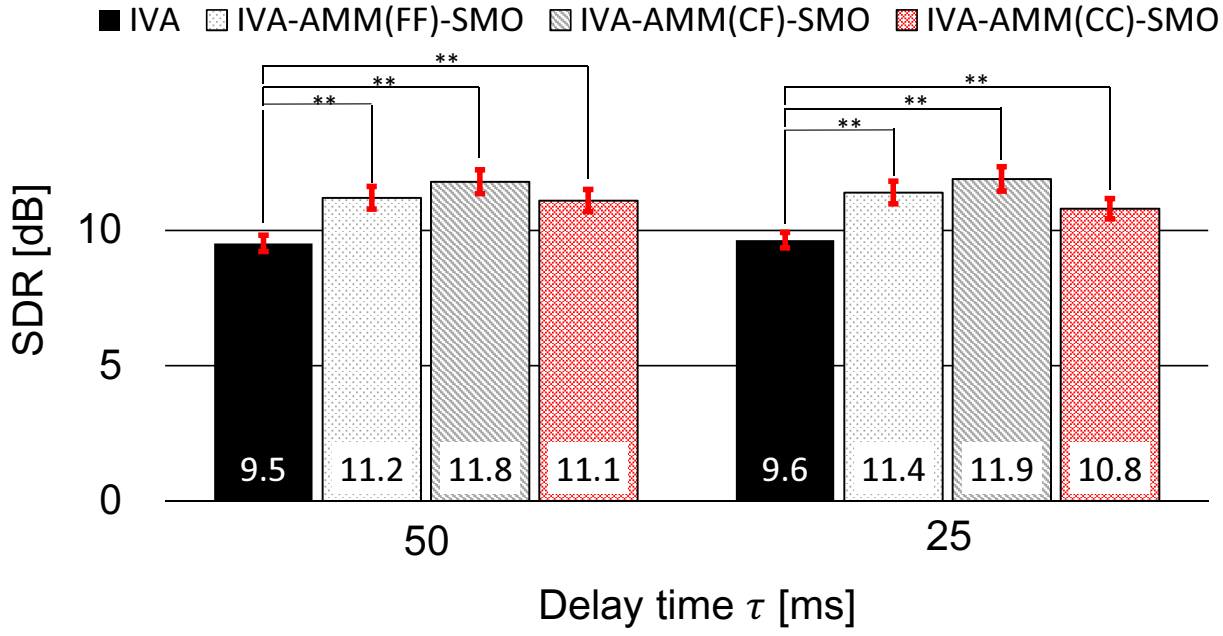
図 3.10 に評価データ 120 発話に対する SIR および SDR の平均値と 99 %信頼区間を示す．SDR の観点でみると，提案法により分離行列を推定した **IVA-AMM(FF)-SMO**，**IVA-AMM(CF)-SMO**，**IVA-AMM(CC)-SMO** は **IVA** の性能を上回った．このことより，提案法は音源が非独

立な場合においても，IVA により生じる歪を低減する効果があることが言える．SIR の観点でみると，**IVA-AMM(FF)-SMO** および **IVA-AMM(CF)-SMO** は，**IVA** の性能を下回った．一方で，**IVA-AMM(CC)-SMO** は **IVA** と同等の性能を示した．図 3.11 に，$\tau$=50 ms とした際の評価データ 120 発話に対する文献 [74] で定義される SIR，SDR および SAR の平均値と，99%信頼区間を示す．SIR の観点でみると，**IVA-AMM(FF)**， **IVA-AMM(CF)** は **IVA** と同等であったが，**IVA-AMM(CC)** は **IVA** の性能を上回った．しかし，SDR，SAR は **IVA** の性能を下回った．このことより，**IVA-AMM(CC)** を用いることにより，分離性能は改善するものの歪が多く発生してしまうことが分かる．一方で，線形分離行列を用いた **IVA-AMM(FF)-SMO**，**IVA-AMM(FF)-SMO** および **IVA-AMM(CC)-SMO** は **IVA** と比較して，SIR および SDR が改善する様子がみられる．これは，提案法による分離が歪の影響を低減しつつ，妨害音に対する分離性能を改善する効果があることを示している．

図 3.12 に，$\tau$=25 ms とした際の評価データ 120 発話に対する文献 [74] で定義される SIR，SDR および SAR の平均値と，99%信頼区間を示す．この条件では，$\tau$=50 ms の場合と比較して，エコー信号と音源信号との相関が高くなるため，分離が難しい．**IVA-AMM(FC)** および **IVA-AMM(CC)** は，**IVA** よりも高い SIR を示したが，SDR および SAR は劣化した．一方で，提案法により分離行列を更新した，**IVA-AMM(FC)-SMO** および **IVA-AMM(CC)-SMO** は，SIR，SDR，SAR すべての評価尺度において **IVA** の性能を上回った．以上の結果より，提案法は，音源が独立出ない場合においても，妨害音の抑圧性能が高く，歪の少ない分離を行うことが可能であるということができる．

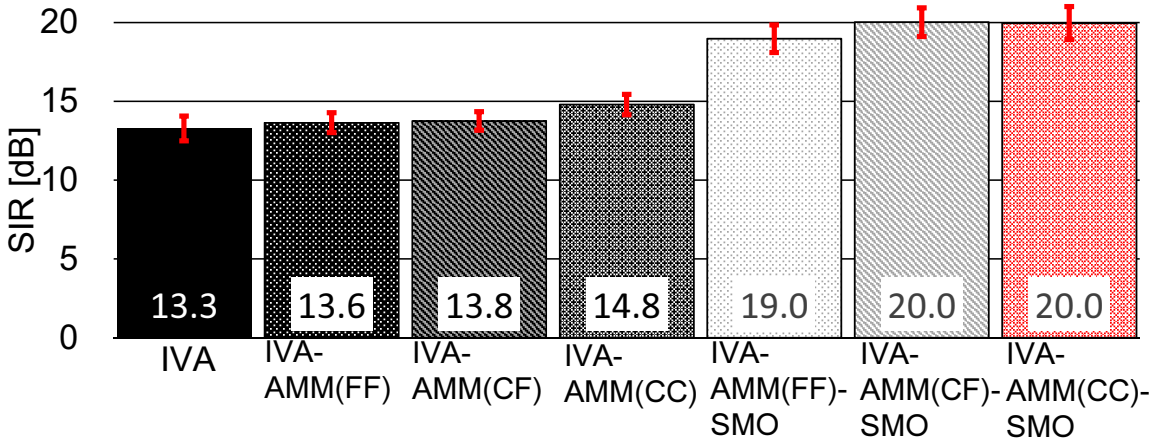



(a) SIR

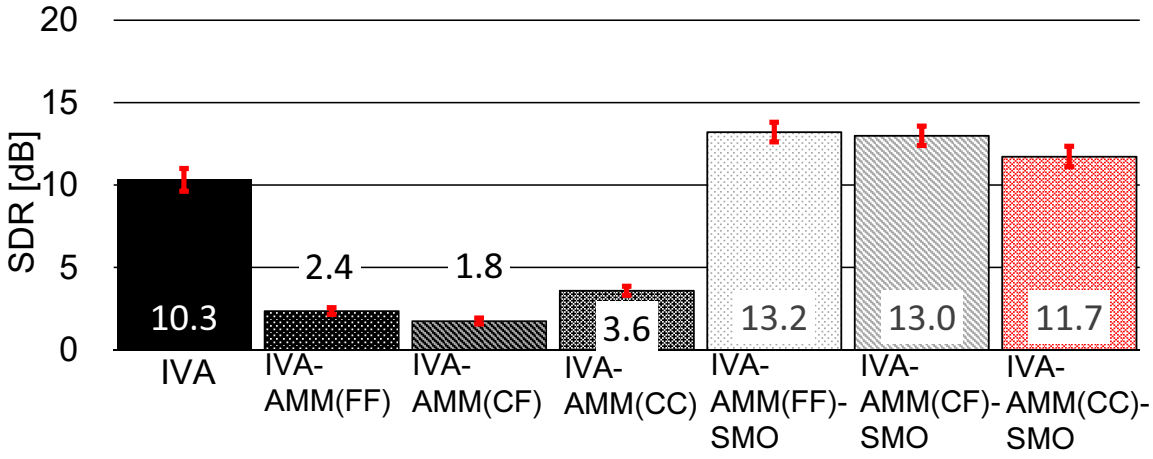



(b) SDR

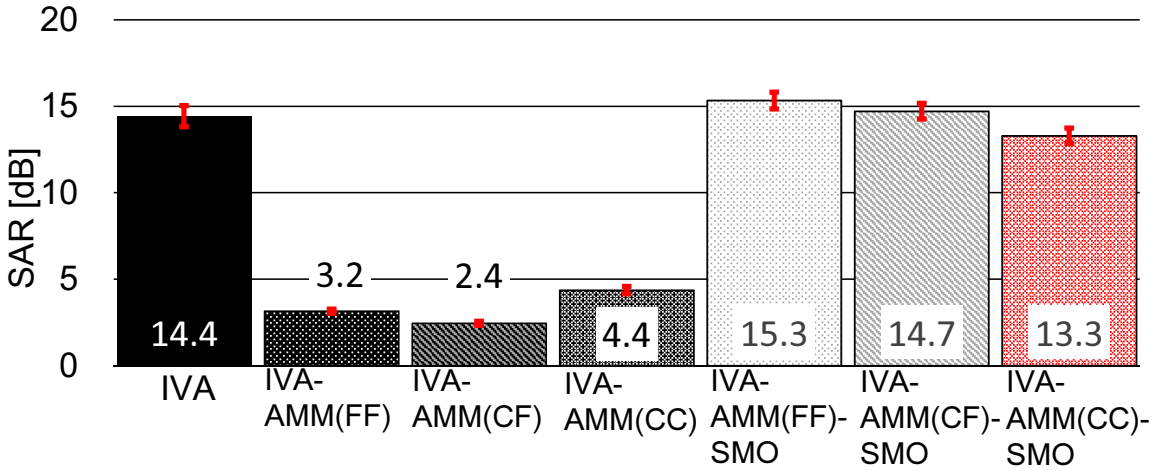
図 3.10: Echo canceling performance with the proposed separation matrix optimization averaged over 120 utterance pairs along with their 99% confidence interval. * denotes significant difference of 5 %; ** denotes significant difference of 1%; n.s. denotes no significant difference.




(a) SIR

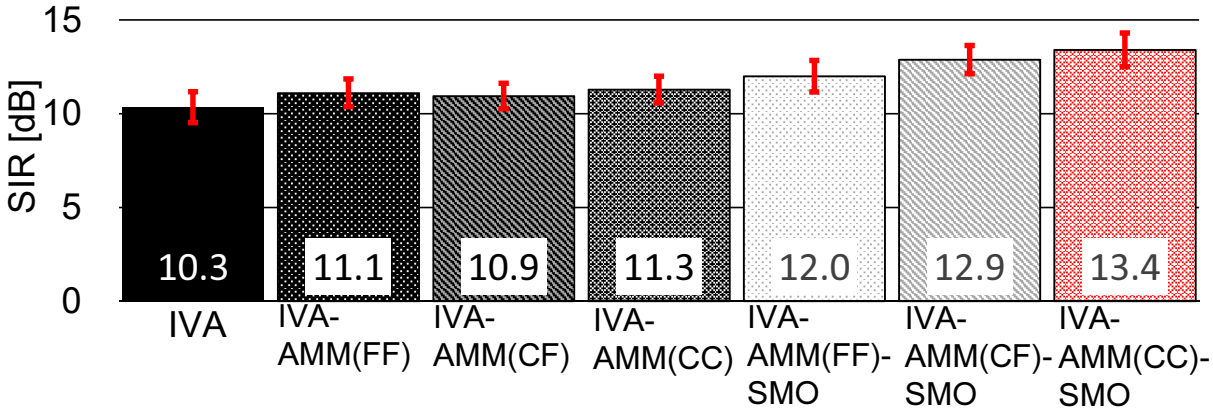



(b) SDR

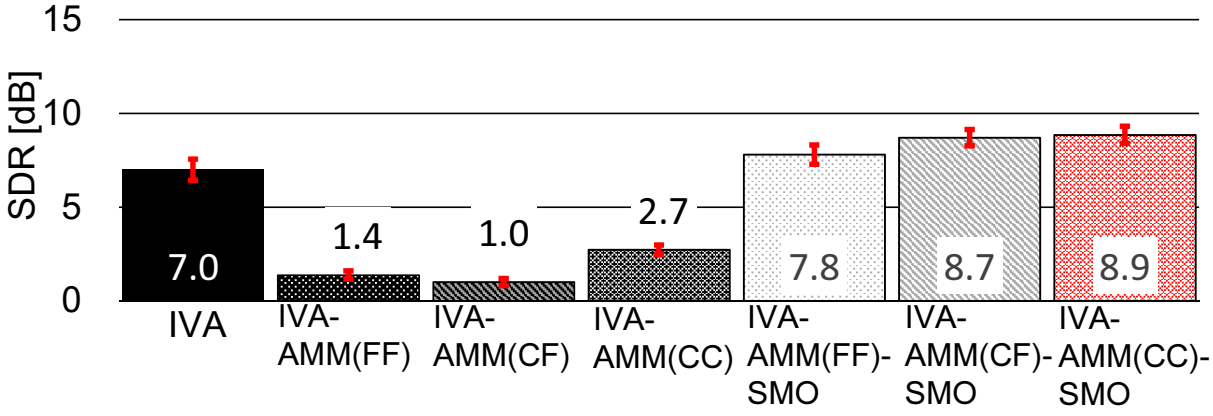



(c) SAR

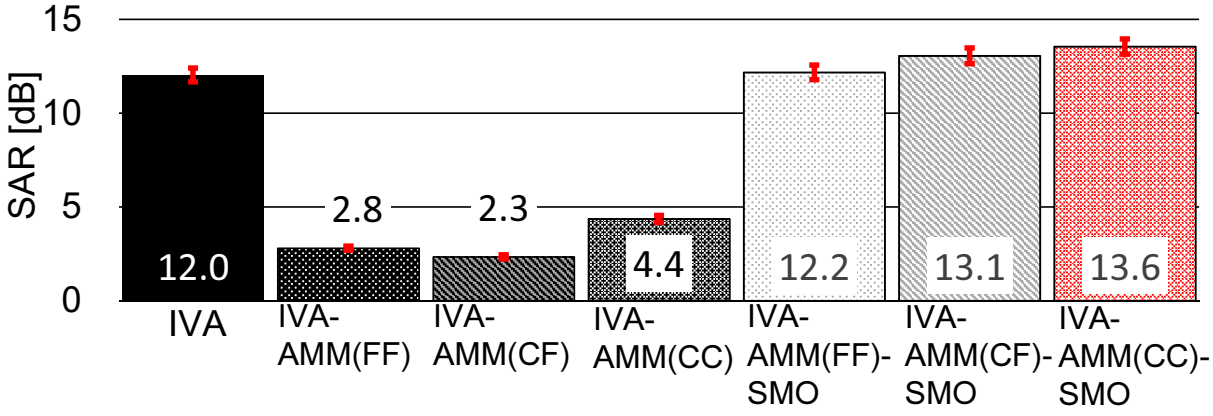
図 3.11: Echo canceling performance ($\tau$=50ms) with the proposed separation matrix optimization averaged over 120 utterances along with their 99% confidence interval. SIR, SDR and SAR are defined in the literature [74].




(a) SIR




(b) SDR




(c) SAR

図 3.12: Echo canceling performance ($\tau$=25ms) with the proposed separation matrix optimization averaged over 120 utterances along with their 99% confidence interval. SIR, SDR and SAR are defined in the literature [74].

### 3.3.5 残響・反射が存在する環境における分離性能の評価

3.3.3 では，反射を含まない環境での二話者同時発話音声の分離問題において，提案法が有効であることを確認した．また 3.3.4 では，エコー信号のように音源に対して非独立な信号を分離する際にも，提案法が有効であることを確認した．ここでは，残響や反射を含むような環境においても提案法が有効であるかどうかを検証した．

評価データはシミュレーションにより生成した．インパルス応答は，RWCP 実環境音声音響データベース（the Real Word Computing Partnership Sound Scene Database: RWCP-SSD）[80] に含まれるものを用いた．実験に用いたインパルス応答の収音環境の条件を，表 3.4 に示す．JNAS より無作為に抽出した女性話者が発した音声にインパルス応答を畳み込み，SNR が 0 dB となるように重畳することで 30 発話（10 話者対 × 各組 3 発話）の二話者同時発話音声を作成した．ここでは，同様の方法で，男性話者が発した二話者同時発話音声も 30 発話作成した．この二話者同時発話音声に対して，以下の分離手法を適用し，その性能を評価した．

**IVA(Baseline)** ：IVA に基づく音源分離

**IVA-AMM(FF)** ：DAE に基づく連想記憶モデル（**FF**）を DAE として用いたもの．入力として IVA の出力を与えた際の出力を分離信号として用いる．

**IVA-AMM(CF)** ：CNN に基づく連想記憶モデル（**CF**）を DAE をとして用いたもの．

**IVA-AMM(CC)** ：DCAE に基づく連想記憶モデル（**CC**）を DAE として用いたもの

**IVA-AMM(FF)-SMO** ：**IVA-AMM(FF)** を用いて推定した分離行列による分離

**IVA-AMM(CF)-SMO** ：**IVA-AMM(CF)** を用いて推定した分離行列による分離

**IVA-AMM(CC)-SMO** ：**IVA-AMM(CC)** を用いて推定した分離行列による分離

提案する分離行列推定法である **IVA-AMM(FF)-SMO**，**IVA-AMM(CF)-SMO**，**IVA-AMM(CC)-SMO** では，IVA により求められた分離行列を初期値として用いた．参照信号の更新回数の上限は 30 回，分離行列の更新回数の上限は 5000 回とした．また，式 (3.16) における学習係数 $\mu$ は初期値を 0.0001 とし，new-bob 法により動的に制御した．

分離性能は，Vincent らが提案した signal-to-interference ratio (SIR)，signal-to-distortion ratio (SDR)，signal-to-artifacts ratio (SAR) [74] により評価した．図 3.14 に評価データ 900 発話に対

表 3.4: Experimental setup for simultaneous speech separation in reverberant environment.

| | |
|---|---|
| Reverberation time (RT60) | 300 ms |
| Number of sources | 2 |
| Number of microphones | 2 |
| Distance between microphones to sources | 200 cm |
| Distance between microphones | 2.83 cm |
| Source direction of ($\theta_1$,$\theta_2$) | (-20,-40),(-20,-80),(-20,20).(-20,40),(-20,80),<br>(-40,-20),(-40,-80),(-40,20).(-40,40),(-40,80),<br>(-80,-20),(-80,-40),(-80,20).(-80,40),(-80,80),<br>(20,-20),(20,-40),(20,-80).(20,40),(20,80),<br>(40,-20),(40,-40),(40,-80).(40,20),(40,80),<br>(80,-20),(80,-40),(80,-80).(80,20),(80,40) |

する SIR，SDR および SIR の平均値と 99 %信頼区間を示す．ここでは，分離手法間の差を調べるために Welch の t 検定を行った．その結果を表 3.6 に示す.

連想記憶モデルの出力を分離信号として用いたとき，**IVA-AMM(FF)** と **IVA** はほぼ同等の SIR を示したが，**IVA-AMM(CF)** および **IVA-AMM(CC)** は，**IVA** よりも高い SIR を示した．このことより，スペクトルを局所的なパターンの組み合わせとして考慮した連想記憶モデルを用いることで，妨害音を抑圧することが可能であると言える．一方で，SDR と SAR の観点で比較すると，連想記憶モデルの出力は **IVA** よりも劣ってしまう傾向があることがわかった．すなわち，連想記憶により推定された参照信号は，線形分離行列の出力よりも歪が多く含まれてしまうため，参照信号をそのまま分離信号として用いることは難しいといえる．SDR、SIR の劣化原因として，連想記憶モデルにより推定されたスペクトルが統計処理による平滑化の影響を受けたものであることが挙げられる．連想記憶により出力されたスペクトルの周波数帯域毎の標準偏差を調べたところ，音源信号のスペクトルに比べて 3dB 程度低下していることが明らかとなった．このことから，連想記憶により出力されたスペクトルが平滑化の影響を受けているといえる.

連想記憶の出力を用いて分離行列を推定した提案法 **IVA-AMM(FF)-SMO**，**IVA-AMM(CF)-SMO** および **IVA-AMM(CC)-SMO** は，SIR，SDR および SAR すべての評価尺度において，**IVA** よりも高い値を示した．すなわち，提案法は反射や残響が含まれる環境においても，IVA により生じる歪の影響を低減しつつ，妨害音を抑圧することが可能な手法であると言える．提案法の中では，エンコード・デコード共にスペクトルの局所的な構造を考慮した連想記憶モデルを用いた **IVA-AMM(CC)-SMO** が最も高い性能を示した．このことより，提案法の枠組みでは連想記憶として **IVA-AMM(CC)** を用いることが効果的であるものと考えることができる.

表 3.5: Significant difference in results for simultaneous speech separation of female pair using IVA and the proposed linear BSS. **FF**, **CF** and **CC** denote **AMM(FF)**, **AMM(CF)** and **AMM(CC)**, respectively. *: significant difference of 5%; **: significant difference of 1%; n.s.: no significant difference; —: not tested.

(a) SIR

| | **IVA-FF** | **IVA-CF** | **IVA-CC** | **IVA-FF-SMO** | **IVA-CF-SMO** | **IVA-CC-SMO** |
|---|---|---|---|---|---|---|
| **IVA** | * | ** | ** | ** | ** | ** |
| **IVA-FF** | — | ** | ** | * | ** | ** |
| **IVA-CF** | ** | — | * | ** | n.s. | ** |
| **IVA-CC** | ** | ** | — | ** | * | ** |
| **IVA-FF-SMO** | ** | ** | ** | — | ** | ** |
| **IVA-CF-SMO** | ** | n.s. | * | ** | — | ** |
| **IVA-CC-SMO** | ** | ** | ** | ** | ** | — |

(b) SDR

| | **IVA-FF** | **IVA-CF** | **IVA-CC** | **IVA-FF-SMO** | **IVA-CF-SMO** | **IVA-CC-SMO** |
|---|---|---|---|---|---|---|
| **IVA** | ** | ** | ** | ** | ** | ** |
| **IVA-FF** | — | n.s. | ** | ** | ** | ** |
| **IVA-CF** | n.s. | — | ** | ** | ** | ** |
| **IVA-CC** | ** | ** | — | ** | ** | ** |
| **IVA-FF-SMO** | ** | ** | ** | — | ** | ** |
| **IVA-CF-SMO** | ** | ** | ** | ** | — | ** |
| **IVA-CC-SMO** | ** | ** | ** | ** | ** | — |

(c) SAR

| | **IVA-FF** | **IVA-CF** | **IVA-CC** | **IVA-FF-SMO** | **IVA-CF-SMO** | **IVA-CC-SMO** |
|---|---|---|---|---|---|---|
| **IVA** | ** | ** | ** | ** | ** | ** |
| **IVA-FF** | — | ** | ** | ** | ** | ** |
| **IVA-CF** | ** | — | ** | ** | ** | ** |
| **IVA-CC** | ** | ** | — | ** | ** | ** |
| **IVA-FF-SMO** | ** | ** | * * | — | ** | ** |
| **IVA-CF-SMO** | ** | ** | ** | ** | — | ** |
| **IVA-CC-SMO** | ** | ** | ** | ** | ** | — |

表 3.6: Significant difference in results for simultaneous speech separation of male pair using IVA and the proposed linear BSS. **FF**, **CF** and **CC** denote **AMM(FF)**, **AMM(CF)** and **AMM(CC)**, respectively. *: significant difference of 5%; **: significant difference of 1%; n.s.: no significant difference; —: not tested.

(a) SIR

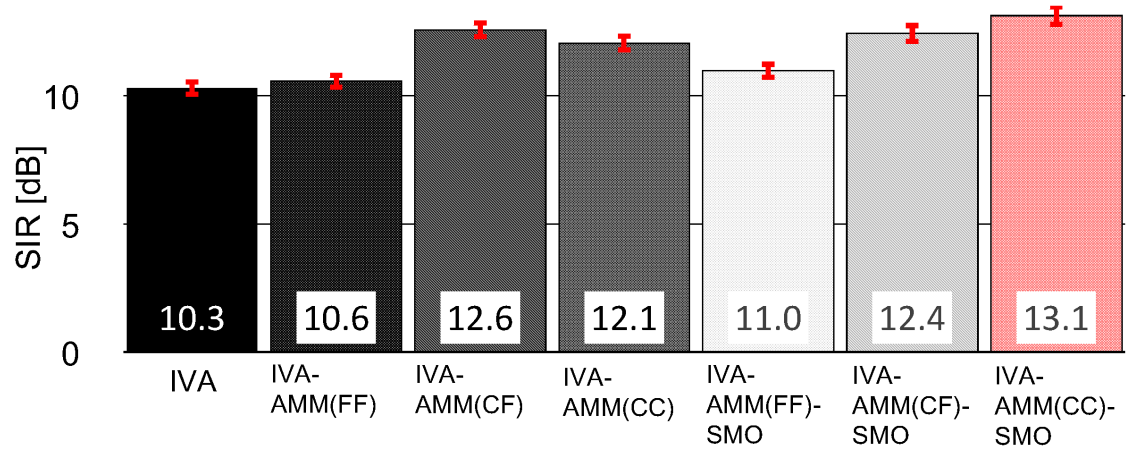
| | **IVA-FF** | **IVA-CF** | **IVA-CC** | **IVA-FF-SMO** | **IVA-CF-SMO** | **IVA-CC-SMO** |
|---|---|---|---|---|---|---|
| **IVA** | n.s. | ** | ** | ** | ** | ** |
| **IVA-FF** | — | ** | ** | ** | ** | ** |
| **IVA-CF** | ** | — | ** | n.s. | ** | ** |
| **IVA-CC** | ** | ** | — | ** | n.s. | ** |
| **IVA-FF-SMO** | ** | n.s. | ** | — | ** | ** |
| **IVA-CF-SMO** | ** | ** | n.s. | ** | — | ** |
| **IVA-CC-SMO** | ** | ** | ** | ** | ** | — |

(b) SDR

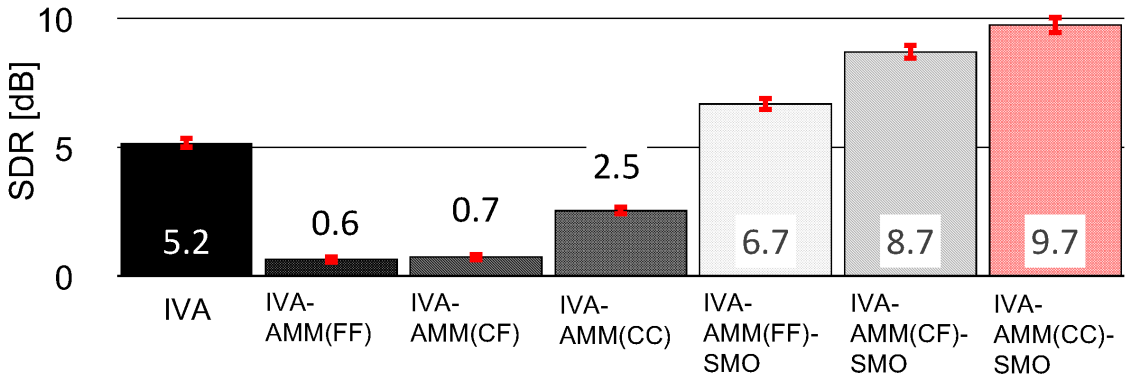
| | **IVA-FF** | **IVA-CF** | **IVA-CC** | **IVA-FF-SMO** | **IVA-CF-SMO** | **IVA-CC-SMO** |
|---|---|---|---|---|---|---|
| **IVA** | ** | ** | ** | ** | ** | ** |
| **IVA-FF** | — | ** | ** | ** | ** | ** |
| **IVA-CF** | ** | — | ** | ** | ** | ** |
| **IVA-CC** | ** | ** | — | ** | ** | ** |
| **IVA-FF-SMO** | ** | ** | ** | — | ** | ** |
| **IVA-CF-SMO** | ** | ** | ** | ** | — | ** |
| **IVA-CC-SMO** | ** | ** | ** | ** | ** | — |

(c) SAR

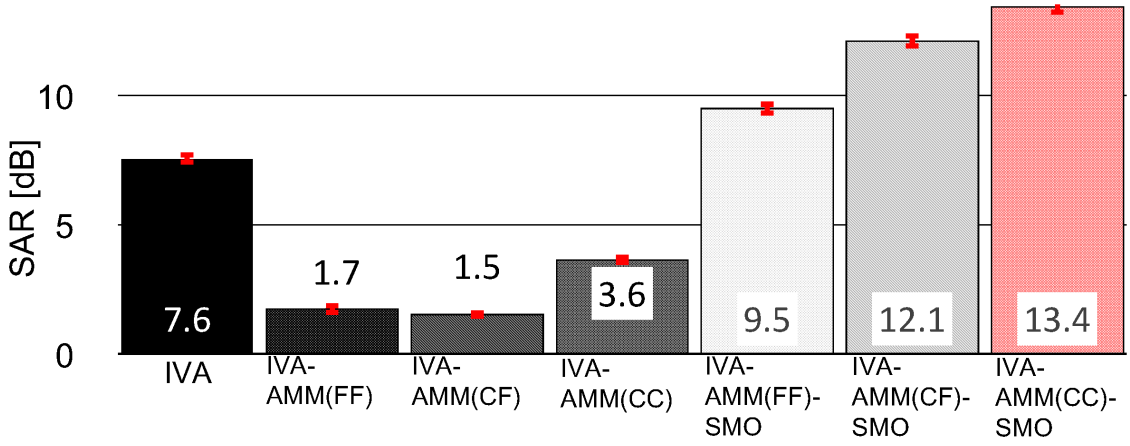
| | **IVA-FF** | **IVA-CF** | **IVA-CC** | **IVA-FF-SMO** | **IVA-CF-SMO** | **IVA-CC-SMO** |
|---|---|---|---|---|---|---|
| **IVA** | ** | ** | ** | ** | ** | ** |
| **IVA-FF** | — | ** | ** | ** | ** | ** |
| **IVA-CF** | ** | — | ** | ** | ** | ** |
| **IVA-CC** | ** | ** | — | ** | ** | ** |
| **IVA-FF-SMO** | ** | ** | * * | — | ** | ** |
| **IVA-CF-SMO** | ** | ** | ** | ** | — | ** |
| **IVA-CC-SMO** | ** | ** | ** | ** | ** | — |




(a) SIR

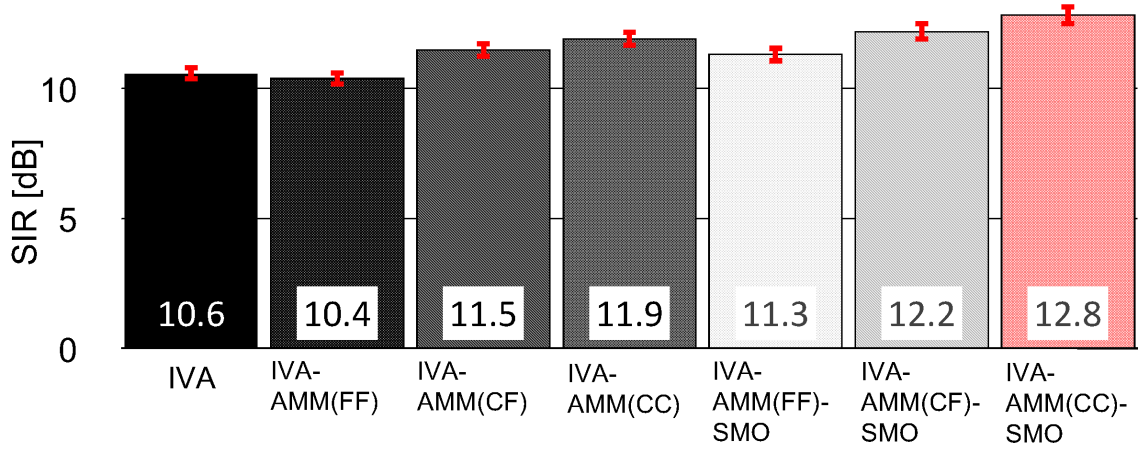



(b) SDR

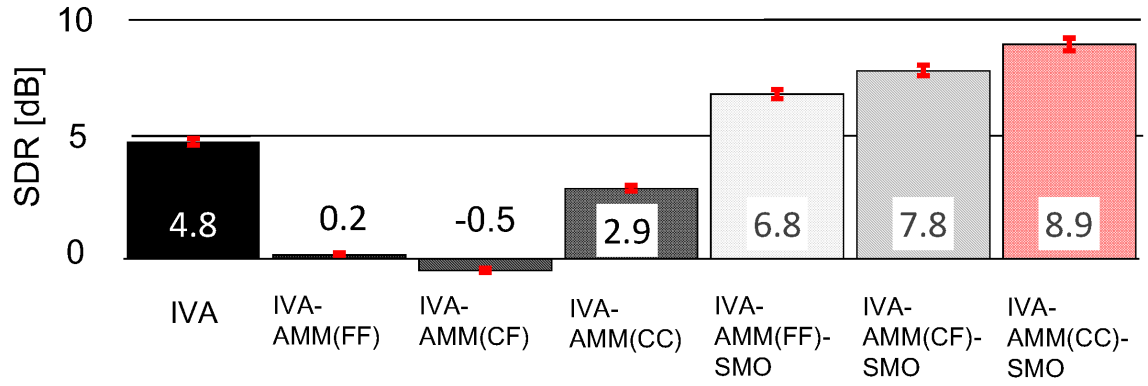



(c) SAR

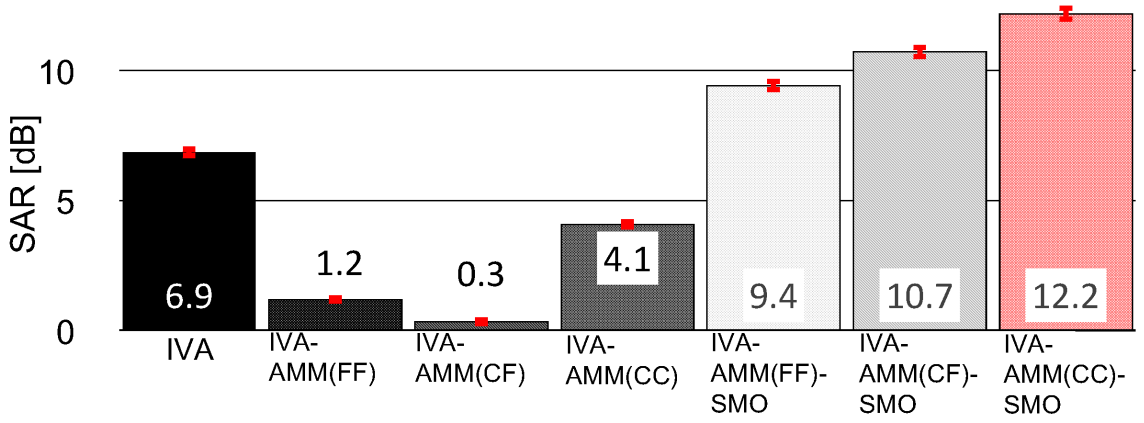
図 3.13: Evaluation of IVA and the proposed AMM-based linear BSS for female pair, where SDR, SIR and SAR were averaged over 900 utterances and the results are shown with their 99 % confidence intervals.




(a) SIR

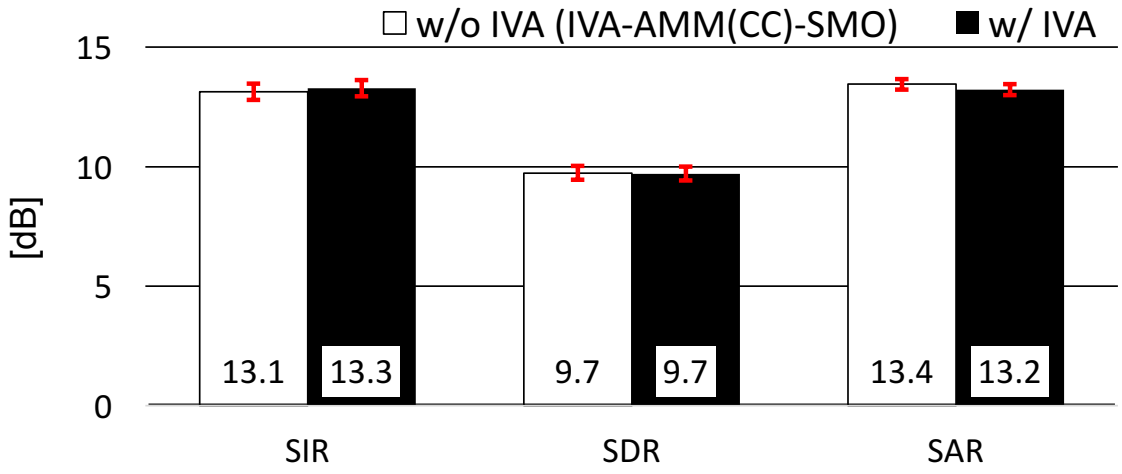



(b) SDR

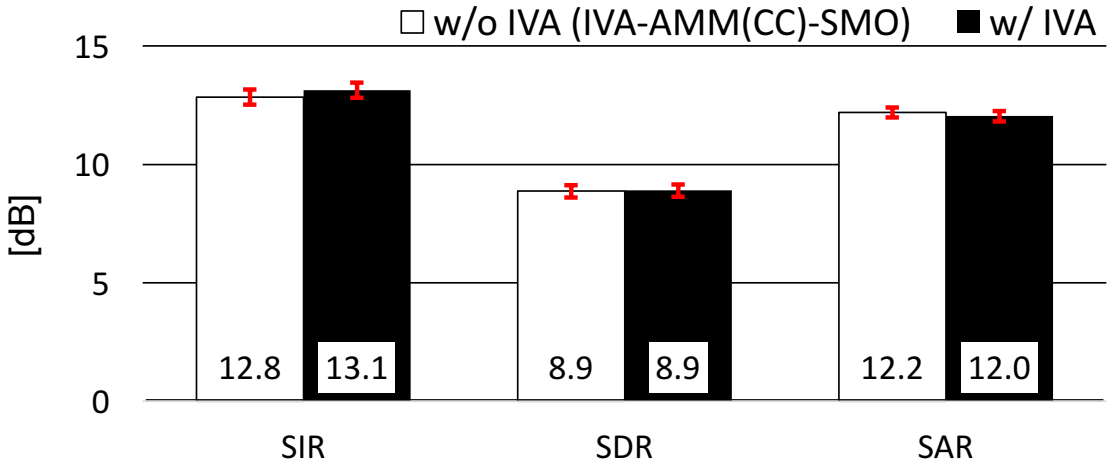



(c) SAR

図 3.14: Evaluation of IVA and the proposed AMM-based linear BSS for male pair, where SDR, SIR and SAR were averaged over 900 utterances and the results are shown with their 99 % confidence intervals.




(a) female




(b) male

図 3.15: Separation performance using the proposed separation matrix optimization method (**IVA-DCAE-SMO**) with and without IVA post-processing averaged over 900 utterances as well as their 99 % confidence intervals. **w/o IVA** is the same as **IVA-DCAE-SMO**. **w/ IVA** indicates that IVA was applied to the separation matrix optimized using **IVA-DCAE-SMO**.

3.2.4 節では，提案法による分離行列更新を行うことにより出力信号間の独立性が大きくなることを示した．その効果を定量的に確認するため，提案法により推定された分離行列を初期値としIVA を適用した際の分離性能を調査した．その結果，図 3.15 に示したように，IVA 適用の有無によって分離性能に大きな変化はみられなかった．このことより，提案法は分離信号が音声らしくなるようにするだけでなく，互いに独立となるように分離する効果があると考えることができる．

## 3.4 まとめ

妨害音に対する抑圧性能が高く歪が少ないブラインド音源分離の実現を目指して，音声のスペクトルを学習した連想記憶モデルを用いて線形分離行列を推定する方法を提案した.

ICA や IVA などに代表される線形分離行列を用いた手法は歪の少ない手法と知られている. しかし，分離行列の推定には音源の独立性の仮定が広く用いられており，独立性の仮定では音源である音声らしさが陽に扱われていない. また，音源に対して独立とは言い難い信号が混入した場合には，独立性に基づき分離することが難しいという課題がある. 一方提案法は，連想記憶モデルを用いることで音声らしさを陽に扱いながら分離行列を推定する. また，独立性の仮定は不要なため，独立とは言い難い信号が混入した場合でも分離が可能となる.

本章では二話者同時発話音声の分離実験により，提案法が独立性に基づき分離行列を推定した場合よりも歪が少なく妨害音に対する抑圧性能が高いことを示した．また，音源に対して非独立な信号としてエコー信号が重畳された信号に対する分離性能を調査し，提案法の有効性を確認した．提案法では分離された信号間の独立性を陽に扱わないが，分離された信号が互い独立になるようにする効果があることも示した.

# 第4章　時間周波数マスクと連想記憶に基づく線形BSSのタンデム接続型音源分離

## 4.1　はじめに

時間周波数マスクに基づくBSSの後段に連想記憶を用いた線形分離行列推定法を適用するタンデム接続型音源分離を開発し，歪の発生量および妨害音の分離性能が共に高い音源分離を実現した．

既存の音源分離手法は，線形処理に基づく方法，非線形処理に基づく方法，複数の手法のタンデム接続に基づく方法に分類することができる．非線形処理に基づく方法の代表例として，時間周波数マスクを用いた方法 [12] が挙げられる．この方法は，妨害音の成分を効果的に抑圧することが可能であるが，ミュージカルノイズなどの非線形歪が発生する．一方，独立成分分析（Independent component analysis: ICA) [39] や独立ベクトル分析（Independent vector analysis: IVA) [40] などに代表される線形処理に基づく方法は，非線形歪が原理的に発生しないという利点がある．しかし，強い反射音が存在する，残響がある環境など，条件においては分離性能が劣化する．タンデム接続に基づく方法は，それらの手法を組み合わせることでそれぞれの欠点を補い合うものである．例えば，ICAと時間周波数マスクを組み合わせた方法 [51, 52]，SMDP(Segrigation using multiple directivity pattern) [53] など，線形処理の後段に非線形処理を組み合わせることで線形処理の分離性能を改善することが可能となる．他には，時間周波数マスクの後段にICAを適用することにより，時間周波数マスクによって生じた歪の影響を低減することが可能であるという報告がある [55].

本章では，時間周波数マスクの利点である妨害音抑圧性能を反映しつつ，欠点である歪の影響を低減するタンデム接続型音源分離を提案する．線形処理としては，3章で検討したDCAEを用いた線形分離行列推定法を用いる．ここでは，時間周波数マスクと線形処理の組み合わせ方法として，非線形処理と線形処理の単純接続，線形処理のみで構成した方法の2つを検討した．非線形処理と線形処理の単純接続では，文献 [55] と同様に，まず，時間周波数マスクにより推定された分離信号に対して，IVAを適用することで分離行列を求める．そして，分離行列に対して連想記憶に基づく分離行列推定を適用する．時間周波数マスクにより妨害音を抑圧されるため，後段

の線形処理が改善されることが期待できる．また，時間周波数マスクの出力には非線形歪が含まれるが，この影響は後段の線形処理により低減されることを期待する．線形処理のみで構成した方法では，時間周波数マスクをそのまま用いるのではなく，同等の処理を線形分離行列で近似する．その分離行列を初期値として IVA を適用し，連想記憶に基づく分離行列推定を行う．この方法においては，観測信号から分離信号を求める処理が線形処理のみで構成されるため，非線形歪が原理的に発生しないという利点がある．

本手法の有効性を検証するために，残響が存在する環境における二話者同時発話音声の分離実験を行った．その結果，時間周波数マスクと線形分離行列推定を組み合わせるためには，時間周波数マスクをそのまま用いるのではなく，線形処理に近似した上で用いることが効果的であることを確認した．

## 4.2 手法概要

時間周波数と連想記憶に基づく分離行列推定を組み合わせたタンデム接続型音源分離を実現を目指し，非線形処理と線形処理の単純接続，非線形処理を線形処理に近似した上で線形処理と接続する 2 つ枠組みを検討した．以降では，これらの方法を概観する．

#### 非線形処理と線形処理の単純接続に基づく枠組み

図 4.1 に示すように，時間周波数マスクを適用した信号に対して線形分離行列を適用する．時間周波数マスクは妨害音源の抑圧性能が高いため，事前にそれらの信号を抑圧することで後段の線形分離行列の推定性能が改善されることを期待している．ここでは，時間周波数マスク $M_n[k,l]$ は 3.2.3 節で示した方法により求めるものとする．このとき，妨害音源が抑圧された信号 $\mathbf{Y}^{(BM)}[k,l]$=$[Y_1^{(BM)}[k,l],\cdots,Y_{N_s}^{(BM)}[k,l]]^T$ は下式により求めることができる．

$$\mathbf{Y}^{(BM)}[k,l] = \mathbf{M}[k,l] \circ \mathbf{Z}[k,l] \ , \tag{4.1}$$

$$\mathbf{M}[k,l] = [M_1[k,l],\cdots,M_{N_s}[k,l]] \ , \tag{4.2}$$

ここで，$\circ$ は要素積を表す．この枠組みでは，$\mathbf{Y}^{(BM)}[k,l]$ を観測信号とみなし，IVA を適用することで分離行列 $\mathbf{W}^{(0)}[k]$ を求める．そして，$\mathbf{W}^{(0)}[k]$ を線形分離行列の初期値と定め，3.2.3 節に示した方法により分離行列を補正する．前述のように，$\mathbf{Y}^{(BM)}[k,l]$ には非線形歪が含まれるが，分離行列を用いることでその影響が低減されることを期待している．

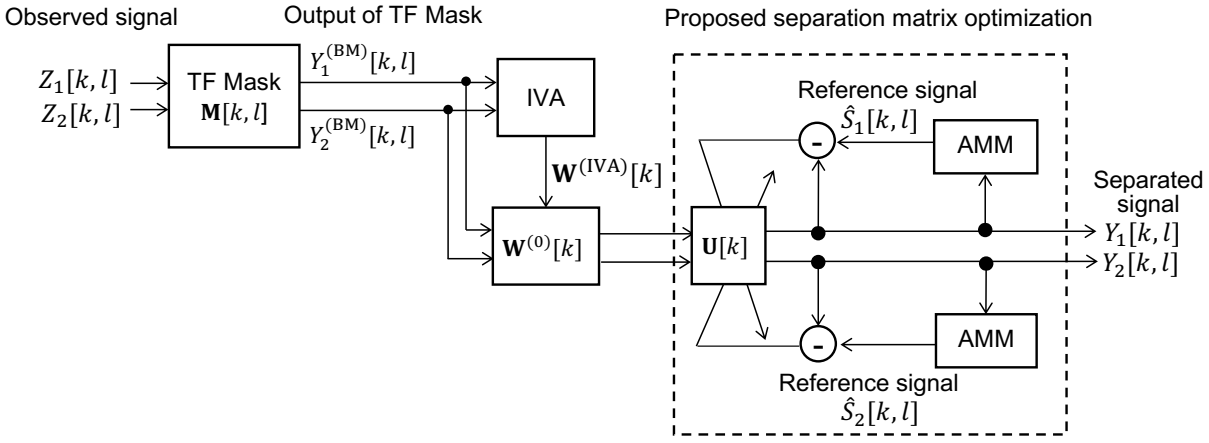



図 4.1: Tandem connectionist BSS based on the simple connection of TF masking and linear separation filtering. The outputs of the TF mask for $\mathbf{M}[k,l]$ were transformed by $\mathbf{W}^{(0)}[k]$ and $\mathbf{U}[k]$ to obtain the source estimates. IVA was applied to the outputs of the TF mask and yield separation matrix $\mathbf{W}^{(\mathrm{IVA})}[k]$. The matrix was used as the initial separation matrix $\mathbf{W}^{(0)}[k]$ and the proposed separation matrix optimization method was applied. Note that the AMM used for separation matrix updating was trained using the IVA output following TF masking IVA and its corresponding dry source in order to consider the effect of musical noise.

線形処理のみで構成した方法

図 4.2 に示すように，線形処理のみで構成する．先述の通り，時間周波数マスクを用いると非線形歪の影響が含まれてしまう．この枠組みでは，時間周波数マスクをそのまま用いるのではなく，出力が最も $\mathbf{Y}^{(BM)}[k,l]$ に近くなる線形変換行列 $\bar{\mathbf{P}}[k]$ を用いることにより，非線形歪の影響を低減されることを期待している．$\mathbf{Y}^{(BM)}[k] = [\mathbf{Y}^{(BM)}[k,0], \cdots, \mathbf{Y}^{(BM)}[k,N_l-1]]^T$，$\mathbf{Z}(k) = [\mathbf{Y}^{(BM)}[k,0], \cdots, \mathbf{Y}^{(BM)}[k,N_l-1]]^T$（$N_l$ はフレーム数）をそれぞれ，時間周波数マスクによる推定値および観測信号の時系列パターンとすると，時間周波数マスクを近似する分離行列 $\bar{\mathbf{P}}[k]$ は，以下の条件を満たす．

$$\bar{\mathbf{P}}[k] = \arg\min_{\mathbf{P}[k]} ||\mathbf{Y}^{(BM)}[k] - \mathbf{W}[k]\mathbf{Z}[k]||^2 \ . \tag{4.3}$$

式 (4.3) を満たす最適解は以下のように計算することができる．

$$\bar{\mathbf{P}}[k] = \mathbf{Y}^{(BM)}[k]\mathbf{Z}[k]^H \left(\mathbf{Z}[k]\mathbf{Z}[k]^H\right)^{-1} \ . \tag{4.4}$$

$\bar{\mathbf{P}}[k]$ を初期値として IVA を適用し，3.2.3 節に示した方法により分離行列を補正する．分離行列の初期値として，$\bar{\mathbf{P}}[k]$ をそのまま用いることも可能であるが，SIR および SDR の観点でみると

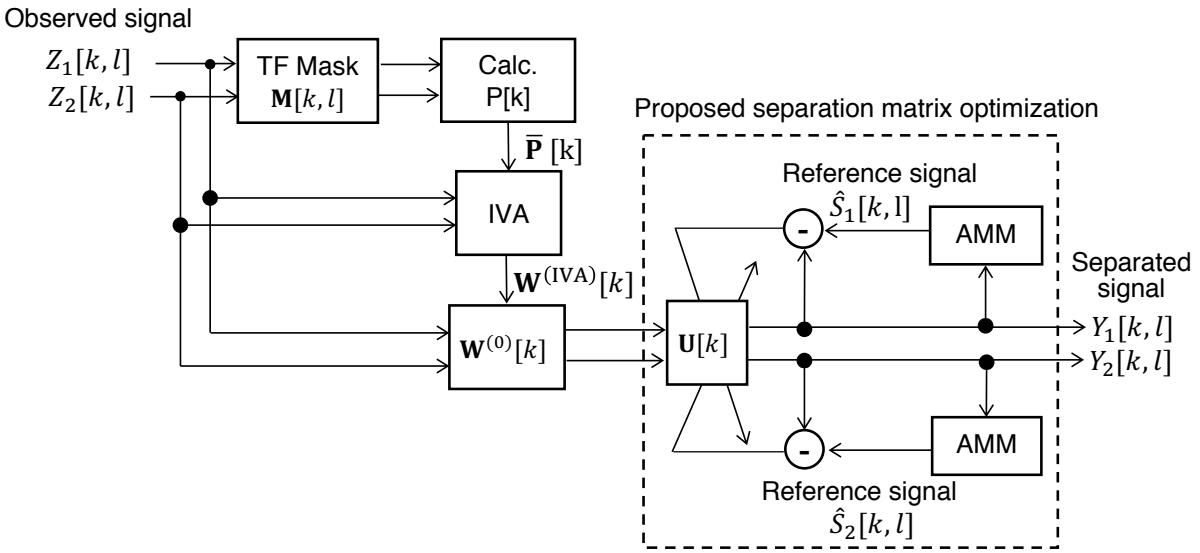



図 4.2: A tandem connectionist framework only comprising linear filtering. The observed signals are transformed by $\mathbf{W}^{(0)}[k]$ and $\mathbf{U}[k]$ to obtain the source estimates. In contrast to the framework depicted in Fig. 4.1, the TF mask outputs are not employed for linear filtering but instead they are used for calculating the initial values of the separation matrix $\bar{\mathbf{P}}[k]$. $\bar{\mathbf{P}}[k]$ is used as the initial separation matrix for IVA, followed by the proposed separation matrix optimization method. Note that the AMM used for separation matrix updating is trained with the signals separated by IVA and its corresponding dry source.

IVA を適用した方が性能が良いことが小規模データセットで確認された．以降の性能評価実験では，$\bar{\mathbf{P}}[k]$ を IVA により更新した分離行列 $\mathbf{W}[k]$ を $\mathbf{W}^{(0)}[k]$ として用いた結果のみを示す．

## 4.3 分離実験

### 4.3.1 連想記憶モデルの学習

連想記憶のパラメータは，クリーン信号を入力データ・教師データの双方に用いるデータ対（**clean-clean**），ノイズを含む分離信号を入力データ，ノイズが取り除かれたクリーン信号を教師データの用いるデータ対（**processed-clean**）の 2 種類のデータ対を用いて決定した．学習には，Stacking autoencoder に基づく貪欲学習を採用し，学習時のミニバッチサイズは 256，学習係数は初期値を 0.1 とし，new-bob 法により動的に制御した．以下に，**clean-clean** および **processed-clean** の詳細を示す．

**clean-clean**

クリーン信号の音声の対数パワースペクトルを入力データ・教師データ双方に用いる．ここ

表 4.1: Experimental setup with simulated environment. In testing set, two positions are selected from $\theta = (-80, -40, -20, 20, 40, 80)$.

| **Data set** | **Source direction** $(\theta_1, \theta_2)$ [deg] | **RT60** $T_{60}$ [ms] | **Mic. interval** $x$ [cm] | **Mic.-Source distance** $d$ [cm] |
|---|---|---|---|---|
| Training | (-15,15), (-45,45), (-75,75), (-90,90) | 0 | 3.00 | 100 |
| Development | (-60,60) | 0 | 3.00 | 100 |
| Testing | ${}_6P_2$ | 300 | 2.83 | 200 |

では，ATR 音素バランス文のセット B に含まれる 1,800 発話 (女性 4 話者 × 各話者 450 発話) を用いた．信号のサンプリング周波数は 16 kHz であり，STFT のフレーム長およびフレームシフトは，それぞれ 1024 サンプル（64 ms）および 256 サンプル（16 ms）とした．

**processed-clean**

ノイズを含む分離音声の対数パワースペクトルを入力データ，対応する目的信号の対数パワースペクトルを教師データに用いる．分離音声は，ATR 音素バランス文のセット B に含まれる女性 4 話者から得られる二話者同時発話音声に対して，補助関数法に基づく IVA を適用することで作成した．なお，図 4.1 に示した非線形処理と線形処理の単純接続における連想記憶では，非線形歪の影響を考慮するために，IVA の前段に時間周波数マスクを適用したときの分離音声を学習に用いた．4 話者から 12 組の話者対を作成し，9 話者対を学習セットに，残りの 3 話者対を開発セットとして用いた．図 4.3 に示す環境において，各話者対が同時に発話することを想定し，ドライソースに遅延を加えることにより，同時発話音声を合成した．このとき，音源の方向を表 4.1 に示す．学習セットでは話者毎に 50 発話を，開発セットでは話者毎に 52 発話をドライソースとして用いた．すなわち学習セットには，3,600 発話（9 話者対 × 2 話者 × 50 発話 × 4 方向）の分離音声を入力データ，対応する 3,600 発話のドライソースを教師データとして用いた．また開発セットには，312 発話（3 話者対 × 2 話者 × 52 発話 × 1 方向）の分離音声を入力データ，対応する 312 発話のドライソースを教師データとして用いた．

### 4.3.2 評価尺度

分離音声の音質，音声認識精度の観点で評価を行った．音質としては，文献 [74] で定義される SIR および SDR を用いた．また，音声認識精度を調べるために，Deep neural network(DNN) に基づく連続音素認識を行い，その音素誤り率を計算した．音素認識を行う際の音響特徴量は，

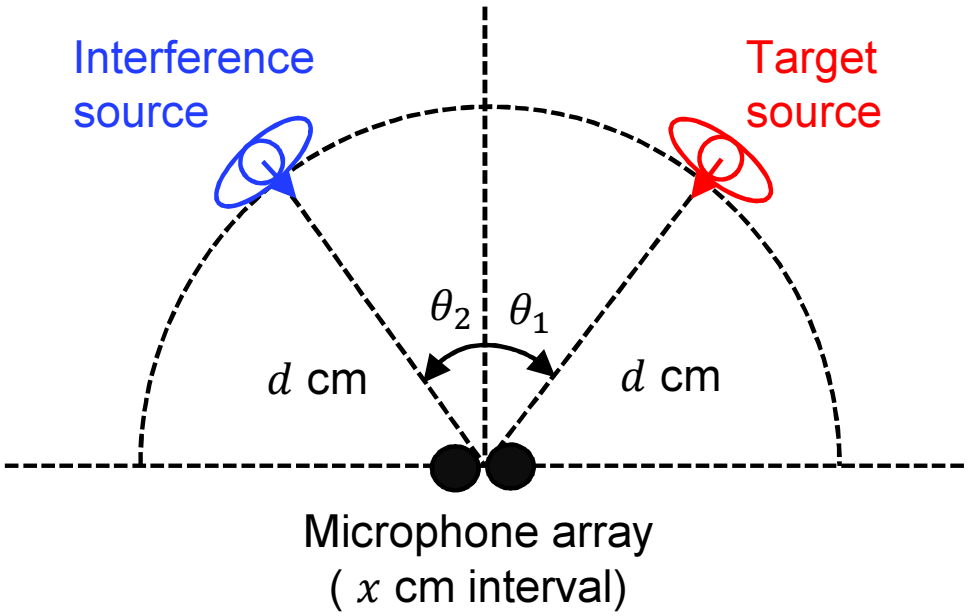



図 4.3: Experimental environment with two microphones and two sources. $x$ denotes the distance between microphones; $d$ denotes the distance between the microphone and sources; $\theta_1$ and $\theta_2$ denote the directions of the target source and interference source.

MFCC+ΔMFCC+ΔΔMFCC の 39 次元を用い，話者の変動を取り除くために，fMLLR による話者適応学習を行った．音響モデルは，ASJ-JNAS に含まれるクリーン音声 41,396 発話から学習したトライフォンの DNN-HMM モデルを用いた．このとき DNN 隠れ層は 5 層，各層のノード数は 1024 とした．言語モデルは，ASJ-JNAS に含まれる 20,000 発話から学習した音素バイグラムを用いた．デコーダには，Kaldi[81] を用いた．

### 4.3.3 実験結果

連想記憶に基づく分離行列推定法と時間周波数マスクを組み合わせた効果を調べるため，以下に示す分離手法を比較した．

**IVA** ：IVA に基づく音源分離

**IVA-AMM-SMO** ：第 3 章で説明した分離行列推定法による分離．**IVA** で求めた分離行列を初期値とし、連想記憶モデル DCAE による参照信号を用いて更新した．

**Mask** ：時間周波数マスクを用いた非線形 BSS [12].

**Mask-IVA** ：**Mask** の出力を観測信号とみなし IVA を適用した BSS.

**Mask-IVA-AMM-SMO** ：非線形処理と線形処理の単純接続に基づく枠組み． **Mask-IVA** によって求められた分離行列を初期値とし，DCAE の出力を参照信号として線形分離行列を更新する．

**Mask-Lin-IVA** ：**Mask** を初期値として，IVA を適用した BSS.

**Mask-Lin-IVA-AMM** ：DCAE に基づく連想記憶モデルを DAE として用いたもの．入力として **Mask-Lin-IVA** 与えた場合の出力を分離信号として用いる．

**Mask-Lin-IVA-SMO** ：**Mask-Lin-IVA** を分離行列の初期値，線形処理のみで構成した枠組み．**Mask-Lin-IVA-AMM** を参照信号として分離行列を更新した提案法

図 4.4 に提案法を用いたタンデム接続型音源分離の分離性能を示す．横軸が SDR，縦軸が SIR を示しており，本研究では両軸が共に高い分離手法の実現を目指している．前章にて提案した **IVA-AMM-SMO** は，**IVA** により求められた線形分離行列に対して，分離信号が音声らしくなるように補正を行う．この方法により，**IVA** よりも高い SDR，SIR が得られる様子が確認できる．本章では，**IVA-AMM-SMO** の前段に **Mask** を組み込む 2 種類のタンデム接続型音源分離 **Mask-IVA-AMM-SMO** および **Mask-Lin-IVA-AMM-SMO** を検討した．**Mask** の後段に **IVA** を適用した **Mask-IVA** は SIR，SDR の観点で **Mask** の性能を上回った．これは，IVA に基づく線形処理が時間周波数マスクによって生じてしまう非線形歪みの影響を低減しつつ，分離性能も改善することができることを示している．**Mask-IVA** により求められた分離行列を連想記憶モデルを用いて更新した **Mask-IVA-AMM-SMO** は **Mask-IVA** の性能をさらに上回り，今回比較した手法の中で最も高い SIR を示した．ところが，**Mask** を前段に適用しなかった **IVA-AMM-SMO** と比較すると，SDR は 2dB 程度劣化してしまった．このことより，時間周波数マスクと提案法の単純接続は，妨害音の抑圧効果を改善することができるものの，歪の影響を考慮する必要があるということがあるといえる．時間周波数マスクと同等の処理を線形分離行列で近似した **Mask-Lin-IVA** と提案法を組み合わせた **Mask-Lin-IVA-AMM-SMO** は，SIR，SDR ともに **IVA-AMM-SMO** および **Mask** を上回った．これは，時間周波数マスクを線形分離行列で近似することで時間周波数マスクの歪の影響を低減することができるのみならず，提案法の妨害音の抑圧性能および歪の発生量を改善することができたことを示している．提案する連想記憶を用いた分離行列推法は勾配法に基づくため，分離性能は分離行列の初期値に依存する．また，**IVA-AMM-SMO** の初期値として用いた **IVA** と **Mask-Lin-IVA-AMM-SMO** の初期値として用いた **Mask-Lin-IVA** を比較す

表 4.2: Phoneme error rate (%) averaged over 30 source directions.

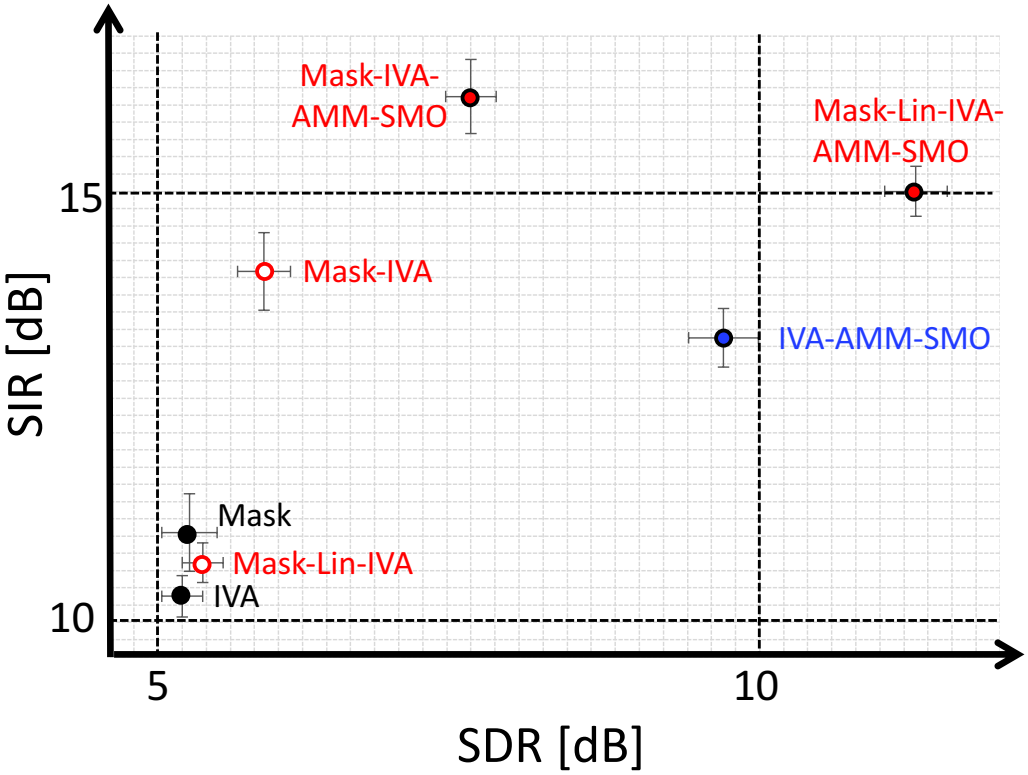
| BSS method | Female | Male | Average |
|---|---|---|---|
| **IVA** | 33.4 | 38.8 | 36.1 |
| **IVA-AMM-SMO** | 29.5 | 36.3 | 32.9 |
| **Mask** | 31.5 | 37.3 | 34.4 |
| **Mask-IVA** | 31.2 | 37.1 | 34.2 |
| **Mask-IVA-AMM-SMO** | 31.7 | 39.2 | 35.4 |
| **Mask-Lin-IVA** | 29.6 | 35.3 | 32.4 |
| **Mask-Lin-IVA-AMM-SMO** | **25.2** | **32.3** | **28.8** |

ると，SIR および SDR が若干改善している．これらのことから，**Mask-Lin-IVA-AMM-SMO** は，**IVA-AMM-SMO** よりも良い初期値が得られたため性能が改善したのではないかと考える．

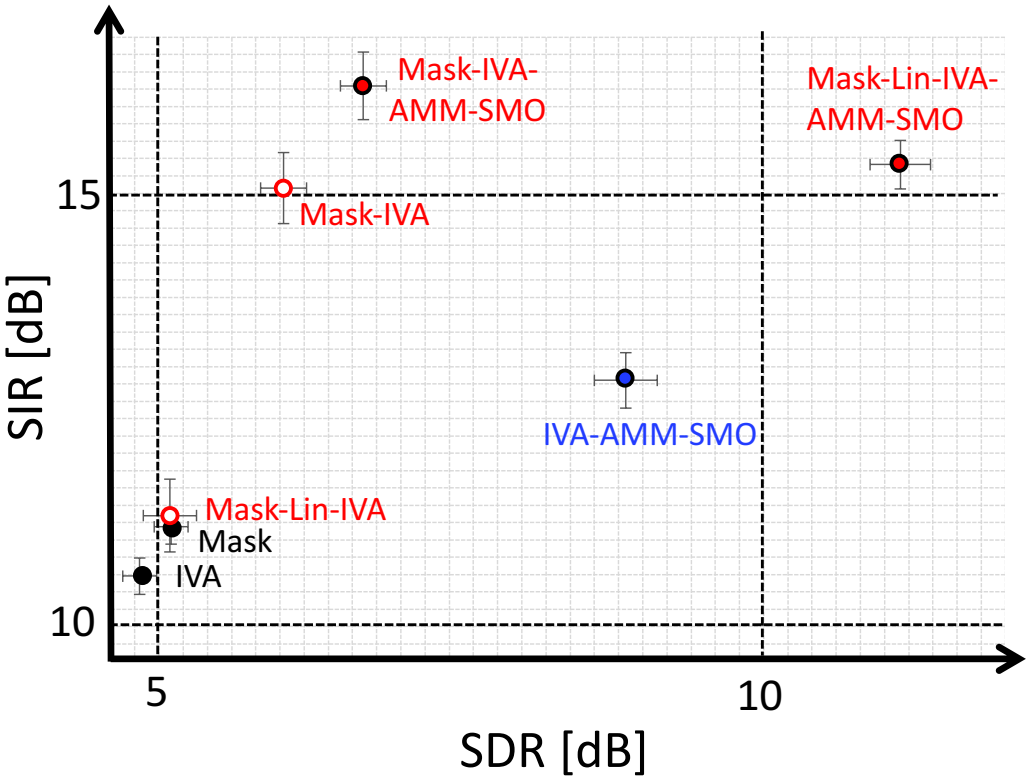
表 4.2 に音素誤り率の比較結果を示す．**IVA-AMM-SMO** は，**IVA** よりも高い認識性能を示した．このことより，音声らしさを考慮して分離行列を更新する提案法の枠組みは音声認識の観点でも有効であると考えられる．**IVA-AMM-SMO** の前段に **Mask** を単純に接続した **Mask-IVA-AMM-SMO** は性能が劣化してしまったが，線形処理のみで構成した **Mask-Lin-IVA-AMM-SMO** は性能が改善した．**Mask-IVA-AMM-SMO** は後段の線形分離行列により非線形歪の影響を低減することを期待しているが，完全には取り除くことが難しいため音響モデルとのミスマッチが発生してしまう．一方，**Mask-lin-IVA-SMO** は非線形歪が原理的に発生しないため，音響モデルとのミスマッチも少なく，音声認識性能が改善したものと考える．

## 4.4 まとめ

連想記憶に基づく線形分離行列推定法の性能を改善することを目指し，前段に時間周波数マスクを組み込んだタンデム接続型音源分離の枠組みを検討した．本枠組みでは，線形分離行列の推定に不要な成分を時間周波数マスクを用いて事前に抑圧することで，後段の分離行列推定が容易になることを期待している．時間周波数マスクをそのまま用いると非線形歪の影響を取り除くことが難しいため，時間周波数マスクと同等の処理を線形分離行列で近似したものを用いる方法を提案した．二話者同時発話音声の分離実験により，提案するタンデム接続型音源分離により連想記憶を用いた線形分離行列推定法の分離性能および歪の発生量が改善することを確認した．また，音声認識の観点においても提案法が有効であることも示した．




(a) female




(b)male

図 4.4: Evaluation of the proposed tandem connectionist framework for female and male pairs, where SDR and SIR were averaged over 900 utterances and the results are shown with their 99 % confidence intervals.

# 第5章　結論

聴感上違和感の少ない BSS の実現を目指し，音源である音声らしさを考慮しながら分離行列を推定する枠組みを提案した．

既存の BSS は，時間周波数マスクに基づく手法などに代表される非線形 BSS と，ICA や IVA などに代表される線形 BSS とに大別することができる．非線形 BSS は，妨害音を効果的に抑圧することが可能であるものの，ミュージカルノイズといった非線形歪が発生するため，分離音の自然性は低くなる．一方線形 BSS は，非線形 BSS よりも妨害音に対する抑圧性能が劣るものの，線形分離行列を用いることにより，非線形歪が原理的に発生しないという利点がある．本研究では以下の 2 つのアプローチにより，自然性および妨害音に対する抑圧性能の高い BSS の実現を目指した．

- 線形 BSS により求められた線形分離行列に対して，音声のスペクトルを学習した連想記憶モデルを用いて分離行列を更新する枠組みを適用
- 時間周波数マスクを用いた非線形 BSS と連想記憶モデルを用いた分離行列推定の枠組みを組み合わせたタンデム接続型音源分離

第 2 章では，提案法を説明する上で重要な基礎技術について説明した．最初に音信号処理の基礎として，時間波形から時間周波数表現であるスペクトルを求める方法および，スペクトルから時間波形を合成する方法について説明した．次に，既存の BSS 手法について概観した．線形 BSS の枠組みとして ICA および IVA を，非線形 BSS の枠組みとして時間周波数マスクを用いた方法について概観した．最後に既存の連想記憶モデルとして，RBM，AE，DAE，CAE，CNN を紹介した．

第 3 章では，連想記憶モデルを用いた分離行列推定法を提案し，性能評価を行った結果について報告した．従来の線形 BSS では，音源の独立性に基づき分離行列の推定を行うことが多い．一方，提案法は事前に音源である音声を学習させた連想記憶を用いることにより，音声らしさを考慮しながら分離行列を推定する．二話者同時発話音声の分離実験を行い，従来の線形分離行列推定法と比較して高精度な分離が可能であることを確認した．また，提案法は音源の独立性の仮定

を置く必要がないため，エコー信号のように音源に対して独立でない信号に対しても分離が可能であることを，エコー除去実験により示した．

第 4 章では，提案した分離行列推定の枠組みと時間周波数マスクを用いた非線形 BSS を統合する方法について検討した．時間周波数マスクを用いて分離行列の推定に不要な成分を事前に抑圧することで，後段の線形分離行列を用いた方法の性能を改善されることが期待できる．しかし，時間周波数マスクを用いると非線形歪が発生するという欠点があった．そこで，時間周波数マスクをそのまま用いるのではなく，同等の分離処理を線形分離行列で近似した上で用いる方法を検討した．二話者同時発話音声の分離実験の結果，提案法は，時間周波数マスクにより生じる歪の影響を低減しつつ，時間周波数マスクと同等の妨害音の抑圧効果があることを確認した．

最後に今後の検討課題と今後の展望について述べる．本研究で提案した分離行列推定法は，性能および計算コストの面でさらなる改善を行う余地がある．まず提案法の性能は，連想記憶モデルのスペクトル推定精度に依存する．そのため，Deep stacking network (DSN) [82]，Very deep convolutional network [83] などのより複雑なネットワーク構造の利用，マルチタスク学習 [84] など，連想記憶モデルのスペクトル推定精度を改善するための取り組みが必要であると考える．また，提案法では分離行列を推定する際に最急降下法を用いているため，収束までに時間がかかる．計算コストの少ない最適化方法を検討する必要がある．

本研究では，連想記憶を用いて線形分離行列を推定する枠組みを提案したが，残響抑圧フィルタやシングルチャネル雑音除去の後処理への応用も可能である．提案法により音声らしさを陽に扱うことで，歪が少なく高精度に音声を再現することができるものと期待する．

# 謝辞

本論文は，筆者が早稲田大学大学院基幹理工学研究科情報理工学専攻博士後期課程において，知覚情報システム研究室で行った研究をまとめたものです．

本研究を進めるにあたり，数多くのご指導およびご助言を下さいました早稲田大学理工学術院 小林哲則教授に心より厚く感謝申し上げます．

ご多忙の中にも関わらず快く副査を引き受けて下さり，本論文について有益なご助言を下さいました早稲田大学理工学術院 松山泰男名誉教授，同学術院 甲藤二郎教授，筑波大学大学院 牧野昭二教授に深く感謝いたします．

本研究に対する御助言だけでなく，原稿執筆や学会発表のご指導をして下さいました早稲田大学基幹理工学研究科 小川哲司准教授に心より感謝いたします．また，本研究に関してご議論いただきました沖電気工業株式会社研究開発センタ 矢頭隆氏，同研究開発センタ 片桐一浩氏，同研究開発センタ 藤枝大博士に心よりお礼申し上げます．

本研究を進めるにあたり，早稲田大学知覚情報システム研究室の皆様に多大の御協力を頂きました．早稲田大学・豊橋技科大学　新田恒雄名誉教授には，本研究とは別の側面で，研究の進め方や研究者としての生き方について御助言を頂きました．早稲田大学助手 俵直弘博士，卒業生 白石洋平氏には，研究に関する議論だけでなく，研究室運営業務のサポートをして頂きました．博士後期課程 峰村今朝明氏には，社会人としての御助言を頂きました．研究室後輩の皆様のおかげで，充実した研究室生活を送ることができました．ここに感謝の意を表します．

最後に，博士後期課程への進学を認め経済的・精神的に支援して下さった祖母，両親，そして最期まで応援して下さった祖父に重ねて厚く謝意を表します．

# 参考文献


[1] T. Higuchi, H. Kameoka, Unified approach for audio source separation with multichannel factorial HMM and DOA mixture model, in: Proc. EUSIPCO 2015, 2015, pp. 2088–2092.

[2] N. Tanaka, T. Ogawa, K. Akagiri, T. Kobayashi, Development of zonal beamformer and its application to robot audition, in: Proc. EUSIPCO2010, 2010, pp. 1529–1533.

[3] Y. Haraguchi, S. Miyabe, H. Saruwatari, K. Shikano, T. Nomura, Source-oriented localization control of stereo audio signals based on blind source separation, in: Proc. EUSIPCO2008, 2008, pp. 177–180.

[4] E. C. Cherry, Some experiments on the recognition of speech, with one and with two ears, J. Acoust. Soc. of Am. 25 (5) (1953) 975–979.

[5] B. D. V. Veen, K. M. Buckley, Beamforming: a versatile approach to spatial filtering, ASSP Magazine, IEEE 5 (2) (1988) 4–24.

[6] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, K. Shikano, Blind source separation combining independent component analysis and beamforming, EURASIP Journal on Applied Signal Processing 2003 (2003) 1135–1146.

[7] V. A. N. Barroso, J. M. F. Moura, Maximum likelihood beamforming in the presence of outliers, in: Proc. ICASSP1991, 1991, pp. 1409–1412 vol.2.

[8] O. L. F. III, An algorithm for linearly constrained adaptive array processing, Proc. of the IEEE 60 (8) (1972) 926–935.

[9] L. J. Griffiths, C. W. Jim, An alternative approach to linearly constrained adaptive beamforming, Antennas and Propagation, IEEE Transactions on 30 (1) (1982) 27–34.

[10] M. S. Pedersen, J. Larsen, U. Kjems, L. C. Parra, A survey of convolutive blind source separation methods, Springer handbook on Speech Processing and Speech Communication.

[11] https://sisec.inria.fr.

[12] H. Sawada, S. Araki, S. Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment, IEEE Trans. ASLP 19 (3) (2011) 516–527.

[13] S. Araki, T. Nakatani, H. Sawada, Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation, in: Proc. ICASSP2010, 2010, pp. 5–8.

[14] A. Alinaghi, W. Wang, P. J. B. Jackson, Integrating binaural cues and blind source separation method for separating reverberant speech mixtures, in: Proc. ICASSP2011, 2011, pp. 209–212.

[15] Y. Tu, J. Du, Y. Xu, L. Dai, C.-H. Lee, Deep neural network based speech separation for robust speech recognition, in: Proc ICSP2014, 2014, pp. 532–536.

[16] Y. H. Tu, J. Du, L.-R. Dai, C. H. Lee, Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition, in: Proc ICASSP2015, 2015, pp. 61–65.

[17] Ö. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Trans. on Signal Proces. 52 (2004) 1830–1847.

[18] T. Melia, S. Rickard, C. Fearon, Histogram-based blind source separation of more sources than sensors using a DUET-ESPRIT technique, in: Proc. EUSIPCO2005, 2005, pp. 1–4.

[19] S. Araki, H. Sawada, R. Mukai, S. Makino, Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors, Elsevier Signal Process. 87 (8) (2007) 1833–1847.

[20] Y. Izumi, N. Ono, S. Sagayama, Sparseness-based 2ch BSS using the EM algorithm in reverberant environment, in: Proc. WASPAA2007, 2007, pp. 147–150.

[21] N. Ito, S. Araki, T. Nakatani, Permutation-free clustering of relative transfer function features for blind source separation, in: Proc. EUSIPCO2015, 2015, pp. 409–413.

[22] N. Madhu, C. Breithaupt, R. Martin, Temporal smoothing of spectral masks in the cepstral domain for speech separation., in: Proc. ICASSP2008, 2008, pp. 45–48.

[23] Y. Ansai, S. Araki, S. Makino, T. Nakatani, T. Yamada, A. Nakamura, N. Kitawaki, Cepstral smoothing of separated signals for underdetermined speech separation., in: Proc. ISCAS2010, 2010, pp. 2506–2509.

[24] G. Hinton, L. Deng, D. Yu, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, G. Dahl, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Processing Magazine 29 (6) (2012) 82–97.

[25] A. Narayanan, D. Wang, Ideal ratio mask estimation using deep neural networks for robust speech recognition, in: Proc. ICASSP, 2013, pp. 7092–7096.

[26] X. L. Zhang, D. L. Wang, Multi-resolution stacking for speech separation based on boosted DNN, in: Proc.INTERSPEECH2015, 2015.

[27] H. Zhang, X. Zhang, S. Nie, G. Gao, W. Liu, A pairwise algorithm for pitch estimation and speech separation using deep stacking network, in: Proc. ICASSP2015, 2015, pp. 246–250.

[28] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proc. ICML2008, 2008, pp. 1096–1103.

[29] S. Moon, J.-N. Hwang, Coordinated training of noise removing networks, in: Proc. ICASSP1993, Vol. 1, 1993, pp. 573–576 vol.1.

[30] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, A. Ng, Recurrent neural networks for noise reduction in robust ASR, in: Proc.INTERSPEECH2012, 2012.

[31] X. Lu, Y. Tsao, S. Matsuda, C. Hori, Speech enhancement based on deep denoising autoencoder, in: Prop. INTERSPEECH 2013, 2013, pp. 436–440.

[32] P. Brakel, D. Stroobandt, B. Schrauwen, Bidirectional truncated recurrent neural networks for efficient speech denoising, in: Proc. INTERSPEECH2013, 2013, p. 5.

[33] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, S. Kuroiwa, Reverberant speech recognition based on denoising autoencoder., in: Proc. INTERSPEECH2013, 2013, pp. 3512–3516.

[34] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, An experimental study on speech enhancement based on deep neural networks, Signal Processing Letters, IEEE 21 (1) (2014) 65–68.

[35] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, C.-H. Lee, Robust speech recognition with speech enhanced deep neural networks, in: Proc. INTERSPEECH2014, 2014, pp. 616–620.

[36] P.-S. Huang, M. Kim, M. H. Johnson, P. Smaragdis, Deep learning for monaural speech separation, in: Proc. ICASSP2014, 2014, pp. 1562–1566.

[37] B. Xia, C. Bao, Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification, Speech Commun. 60 (2014) 13 – 29.

[38] T. T. Vu, B. Bigot, E. S. Chng, Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition, in: Proc. ICASSP2016, 2016, pp. 499–503.

[39] P. Smaragdis, Blind separation of convolved mixtures in the frequency domain, Neurocomputing 22 (1-3) (1998) 21–34.

[40] T. Kim, H. T. Attias, S.-Y. Lee, T.-W. Lee, Blind source separation exploiting higher-order frequency dependencies, IEEE Trans. ASLP 15 (1) (2007) 70–79.

[41] S. Amari, S. C. Douglas, A. Cichocki, H. H. Yang, Multichannel blind deconvolution and equalization using the natural gradient, in: Signal Processing Advances in Wireless Communications, First IEEE Signal Processing Workshop on, 1997, pp. 101–104.

[42] A. J. Bell, T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, NEURAL COMPUTATION 7 (1995) 1129–1159.

[43] H. S. T. Kawamura, T. Nishikawa, A. Lee, K. Shikano, Blind source separation based on a fast-convergence algorithm combining ICA and beamforming, IEEE Trans. on ASLP 14 (2) (2006) 666–678.

[44] H. Sawada, R. Mukai, S. Araki, S. Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation, Speech and Audio Processing, IEEE Transactions on 12 (5) (2004) 530–538.

[45] N. Murata, S. Ikeda, A. Ziehe, An approach to blind source separation based on temporal structure of speech signals, Neurocomputing 41 (2001) 1–24.

[46] K. Matsuoka, Minimal distortion principle for blind source separation, in: Proc. SICE 2002., Vol. 4, 2002, pp. 2138–2143.

[47] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, F. Itakura, Evaluation of blind signal separation method using directivity pattern under reverberant conditions, in: Proc ICASSP2000, Vol. 5, 2000, pp. 3140–3143.

[48] E. Moulines, J. Cardoso, E. Gassiat, Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models, in: Proc. ICASSP1997, Vol. 5, 1997, pp. 3617–3620.

[49] I. Lee, J. Hao, T.-W. Lee, Adaptive independent vector analysis for the separation of convoluted mixtures using EM algorithm, in: Proc. ICASSP2008, 2008, pp. 145–148.

[50] Y. Liang, J. Harris, S. M. Naqvi, G. Chen, J. A. Chambers, Independent vector analysis with a generalized multivariate gaussian source prior for frequency domain blind source separation, Signal Processing 105 (2014) 175–184.

[51] D. Kolossa, R. Orglmeister, Nonlinear postprocessing for blind speech separation 3195 (2004) 832–839.

[52] H. Sawada, S. Araki, R. Mukai, S. Makino, Blind extraction of dominant target sources using ICA and time-frequency masking, IEEE Trans on ASLP 14 (6) (2006) 2165–2173.

[53] T. Isa, T. Sekiya, T. Ogawa, T. Kobayashi, Source separation using multiple directivity patterns produced by ica-based bss, in: Proc EUSIPCO2006, 2006, pp. 1–5.

[54] D. S. Williamson, Y. Wang, D. Wang, Reconstruction techniques for improving the perceptual quality of binary masked speech, The Journal of the Acoust. Soc. of America 136 (2) (2014) 892–902.

[55] S. Araki, S. Makino, A. Blin, R. Mukai, H. Sawada, Underdetermined blind separation for speech in real environments with sparseness and ICA, in: Proc ICASSP2004, Vol. 3, 2004, pp. iii–881–4.

[56] T. Higuchi, N. Ito, T. Yoshioka, T. Nakatani, Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise, in: Proc. ICASSP2016, 2016, pp. 5210–5214.

[57] J. Heymann, L. Drude, R. Haeb-Umbach, Neural network based spectral mask estimation for acoustic beamforming, in: Propc. ICASSP2016, 2016, pp. 196–200.

[58] M. Omachi, T. Ogawa, T. Kobayashi, M. Fujieda, K. Katagiri, Separation matrix optimization using associate memory model for blind source separation, in: Proc. EUSIPCO 2015, 2015.

[59] M. Omachi, T. Ogawa, T. Kobayashi, Associative memory model-based linear filtering and its application to tandem connectionist blind source separation, IEEE/ACM Trans. on ASLP.

[60] D. Scherer, A. Müller, S. Behnke, Evaluation of pooling operations in convolutional architectures for object recognition., in: Proc. ICANN, Vol. 6354, 2010, pp. 92–101.

[61] D. Griffin, J. S. Lim, Signal estimation from modified short-time fourier transform, Acoust., Speech and Signal Process., IEEE Trans. on 32 (2) (1984) 236–243.

[62] S. Roucos, A. Wilgus, High quality time-scale modification for speech, in: Proc. ICASSP'85, Vol. 10, 1985, pp. 493–496.

[63] S. Amari, Natural gradient works efficiently in learning, Neural Comput. 10 (2) (1998) 251–276.

[64] N. Ono, S. Miyabe, Auxiliary-function-based independent component analysis for super-gaussian sources, in: Proc. LVA/ICA, 2010, pp. 165–172.

[65] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood for incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39 (1977) 1–38.

[66] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences 79 (8) (1982) 2554–2558.

[67] D. H. Ackley, G. E. Hinton, T. J. Sejnowski, A learning algorithm for boltzmann machines, Cognitive Science 9 (1) (1985) 147–169.

[68] G. E. Hinton, A practical guide to training restricted boltzmann machines., in: Neural Networks: Tricks of the Trade (2nd ed.), Vol. 7700 of Lecture Notes in Computer Science, Springer, 2012, pp. 599–619.

[69] P. Baldi, K. Hornik, Neural networks and principal component analysis: Learning from examples without local minima, Neural Networks 2 (1) (1989) 53 – 58.

[70] J. Masci, U. Meier, D. Ciresan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, in: Proc. ICANN, 2011, pp. 52–59.

[71] L. Deng, O. A. Hamid, Y. Dong, A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion, in: Proc. ICASSP2013, 2013, pp. 6669–6673.

[72] G. E. Hinton, Training products of experts by minimizing constrastive divergence, Neural computation 14 (2002) 1771–1800.

[73] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.

[74] E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation, Audio, Speech, and Language Processing, IEEE Transactions on 14 (4) (2006) 1462–1469.

[75] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano, ATR Japanese speech database as a tool of speech recognition and synthesis, Speech Commun. 9 (4) (1990) 357 – 363.

[76] N. Ono, Stable and fast update rules for independent vector analysis based on auxiliary function technique, in: Proc. WASPAA2011, 2011, pp. 189–192.

[77] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, J. Acoust. Soc. Jpn 20 (3) (1999) 199–206.

[78] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in: Proc. ICASSP2010, 2010, pp. 4214–4217.

[79] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, N. Duong, The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges, Elsevier Signal Process. 92 (8) (2012) 1928–1936.

[80] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, in: Proceedings of 2nd ICLRE, 200, pp. 965–968.

[81] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, G. Stemmer, The kaldi speech recognition toolkit, in: IEEE 2011 workshop, 2011.

[82] L. Deng, B. Hutchinson, D. Yu, Parallel training of deep stacking networks, in: Proc Interspeech2012, 2012.

[83] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556.

[84] R. Caruana, Multitask learning: A knowledge-based source of inductive bias, in: Proc. of the Tenth International Conference on Machine Learning, 1993, pp. 41–48.

# 研究業績

## 本論文に関連する研究業績

### 論文

[1] M. Omachi, T. Ogawa and T. Kobayashi, “Associative memory model-based linear filtering and its application to tandem connectionist blind source separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 3, pp. 637–650, Mar 2017.

### 講演

[2] M. Omachi, T. Ogawa, T. Kobayashi, M. Fujieda, K. Katagiri, “Separation matrix optimization using associative memory model for blind source separation,” Proc. EUSIPCO2015, pp.1103–1107, Sep. 2015.

[3] 大町基, 小川哲司, 小林哲則, 藤枝大, 片桐一浩. “連想記憶に基づく線形分離行列推定を用いたタンデム接続型音源分離,” 日本音響学会春季講演論文集, pp.21–24, Mar. 2016 (第 13 回学生優秀発表賞受賞).

[4] 大町基, 小川哲司, 小林哲則, 藤枝大, 片桐一浩. “連想記憶に基づくブラインド音源分離のエコーキャンセリングへの応用,” 日本音響学会秋季講演論文集, pp.593–596, Sep. 2015.

[5] 大町基, 小川哲司, 小林哲則, 藤枝大, 片桐一浩. “連想記憶を用いた線形分離行列推定法に関する検討,” 音学シンポジウム 2015（第 107 回 音楽情報科学研究会）, May 2015.

[6] 大町基, 小川哲司, 小林哲則, 藤枝大, 片桐一浩. “連想記憶と線形分離フィルタを用いたブラインド音源分離,” 第 105 回 音声言語情報処理研究会, Feb. 2015.

[7] 大町基, 小川哲司, 赤桐建三, 小林哲則, “指向性を付与したマルチチャンネルウィーナフィルタを前段に持つ音源分離方式の検討,” 日本音響学会春季講演論文集, pp.937–940, Mar. 2013.

[8] 大町基, 小川哲司, 小林哲則, “天井設置型マイクロホンアレイにおいて残響が音声の分離・認識性能に与える影響,” 日本音響学会秋季講演論文集, pp.791–792, Sep. 2012.

## その他の研究業績

### 講演

[1] Y. Kubota, M. Omachi, T. Ogawa, T. Kobayashi and T. Nitta, “Effect of frequency weighting on MLP-based speaker canonicalization,” Proc. INTERSPEECH2014, pp.2987–2991, Sep. 2014.

[2] 久保田雄一, 大町基, 小林哲則, 新田恒雄, “話者正準化を用いた連続音声認識における改良,” 日本音響学会春季講演論文集, pp.1–2, Mar. 2015.

[3] 久保田雄一, 大町基, 小川哲司, 小林哲則, 新田恒雄, “MLP を用いた話者正準化に基づく音声認識の検討,” 第 102 回 音声言語情報処理研究会, Jul. 2014.

[4] 久保田雄一, 大町基, 小川哲司, 小林哲則, 新田恒雄, “標準話者母音スペクトルへの変換に基づく話者正準化,” 日本音響学会春季講演論文集, pp.77–78, Mar. 2014.

[5] 大町基, 岩田和彦, 小林哲則, “実環境における距離感変換音声の評価,” 日本音響学会春季講演論文集, pp.343–344, Mar. 2012.

[6] 大町基, 岩田和彦, 小林哲則, “距離感を変換した合成音声の評価,” 日本音響学会秋季講演論文集, pp.385–388, Sep. 2011.

[7] 大町基, 岩田和彦, 小林哲則, “音声における距離感の基本周波数パターン変換方法の検討,” 日本音響学会春季講演論文集, pp.421–422, Mar. 2011.

[8] 大町基, 岩田和彦, 小林哲則, “音声における距離感の変換方法の評価,” 日本音響学会秋季講演論文集, pp.339–440, Sep. 2010.

[9] 大町基, 岩田和彦, 小林哲則, “音声における距離感の変換方法の検討,” 日本音響学会春季講演論文集, pp.295–296, Mar. 2010.

[10] 大町基, 岩田和彦, 小林哲則, “距離感を与える音声の特徴分析と合成,” 電子情報通信学会技術研究報告, SP2009-89, pp.159–163, Dec. 2009.