一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2017-104,SIP2017-113,SP2017-87 (2018-03)

独立深層学習行列分析に基づく多チャネル音源分離の実験的評価

北村　大地†　角野　隼斗†　高宗　典玄†　高道慎之介†　猿渡　洋†
小野　順貴††

† 東京大学　〒113–8656 東京都文京区本郷 7–3–1
†† 首都大学東京　〒191–0065 東京都日野市旭が丘 6–6

あらまし　本稿では，新しい教師あり多チャネル音源分離手法である独立深層学習行列分析（IDLMA）を提案する．IDLMA は，従来のブラインド音源分離の独立低ランク行列分析と，近年発展している教師あり学習のディープニューラルネットワーク（DNN）を融合したアルゴリズムであり，独立成分分析を起源とする統計的独立性に基づく信号分離理論の正当な教師あり拡張手法である．本手法では，DNN を用いて音源の時間周波数構造をモデル化しつつ，観測信号の空間的な混合モデルをブラインドに推定することができる．音楽信号を用いた評価実験では，IDLMA が従来の DNN に基づく多チャネル音源分離手法よりも高速かつ高精度な音源分離を実現できることを示す．
キーワード　多チャネル音源分離，独立成分分析，ディープニューラルネットワーク

Experimental Evaluation of Multichannel Audio Source Separation Based on IDLMA

Daichi KITAMURA†, Hayato SUMINO†, Norihiro TAKAMUNE†, Shinnosuke TAKAMICHI†,
Hiroshi SARUWATARI†, and Nobutaka ONO††

† The University of Tokyo　7–3–1 Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
†† Tokyo Metropolitan University　6–6 Asahigaoka, Hino-shi, Tokyo 191–0065, Japan

Abstract　In this paper, we propose a new informed multichannel audio source separation called independent deeply learned matrix analysis (IDLMA). IDLMA is a unified algorithm of conventional blind source separation, independent low-rank matrix analysis, and a supervised learning method based on deep neural networks (DNN) and can be interpreted as a natural informed extension of the independence-based source separation theory. Although a source model is estimated by pre-trained sourcewise DNN, a spatial model can blindly be estimated by statistical independence between sources. The experiment using music signals shows the efficacy of IDLMA compared with the conventional DNN-based techniques.
Key words　multichannel audio source separation, independent component analysis, deep neural networks

1. はじめに

ブラインド音源分離（blind source separation: BSS）とは，音源位置や混合系が未知の条件で，観測された信号のみから混合前の音源信号を推定する技術である．優決定条件（音源数 ≦ 観測チャネル数）における BSS では，独立成分分析（independent component analysis: ICA）[1] に基づく手法及びその拡張手法が主流である．特に，独立ベクトル分析（independent vector analysis: IVA）[2]–[6] と非負値行列因子分解（nonnegative matrix factorization: NMF）[7], [8] を融合した手法である独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[9]–[12] は，従来手法よりも高精度な音源分離を実現している．ILRMA は，音源間の統計的独立性と各音源の時間周波数構造の低ランク性を仮定した BSS であり，時間周波数の低ランクな共変構造を NMF でモデル化することで，周波数領域 ICA [13]–[16] におけるパーミュテーション問題 [17]（音源毎の分離フィルタの順序が周波数間で整合できない問題）を回避しつつ，周波数毎の分離行列を推定する．

劣決定条件（音源数 > 観測チャネル数）における多チャネル BSS では，多チャネル観測信号の時間周波数成分を多変量複素ガウス分布の分散及び相関行列でモデル化する手法 [18]（以後，Duong 法と呼ぶ）が有名である．Duong 法では，各音源

のパワースペクトログラムに対応する時変な分散（音源モデル）と，空間的な伝達系に対応する時不変な相関行列（空間相関行列，空間モデル）を周波数毎に推定するが，周波数領域 ICA と同じく推定後にパーミュテーション問題を解決しなければならない．その後，Duong 法は音源モデルの推定に NMF を導入した多チャネル NMF（multichannel NMF: MNMF）[19], [20] へと発展し，音源の時間周波数構造の低ランク性に基づいてパーミュテーション問題を回避する推定法が登場した．この MNMF と ILRMA は空間相関行列のランクに関する制約を除いて本質的に等価なアルゴリズムである [9]．但し，Duong 法や MNMF は混合系を空間モデルとして推定するのに対して ILRMA は分離系を推定するアルゴリズムであり，最適化の観点では MNMF よりも ILRMA の方が高速であり，パラメタの初期値に対しても頑健であることが実験的に示されている [9]．MNMF や ILRMA のように NMF を音源モデルとして用いるだけでなく，より一般に，任意の単一チャネル音源分離手法を用いて，分離結果のいわば「お手本」となるようなスペクトログラムを求めておき，それを音源モデルとして用いて分離行列を高精度に推定する研究も行われており [21]，こうした研究からも多チャネル音源分離における音源モデルの重要性が示唆される．

一方，近年はディープニューラルネットワーク（deep neural networks: DNN）に基づく音源分離手法も広く研究されるようになり，単一チャネル信号を対象とした手法 [22]– [24] や多チャネル信号を対象とした手法 [25], [26] が数多く提案されている．音声や楽器音のように分離対象となる音源の学習データが大量に用意可能な状況では，DNN に基づく音源情報（対象音源の音色やその時間変化等）のモデル化は有効であり，数多くの文献で高精度な音源分離を実現している．しかし，多チャネル信号で観測できる空間情報（音源位置，マイクロホン位置，部屋の形状，残響時間等の膨大な物理要因に依存する情報）を DNN で学習し，汎化性の高いモデルを構築することは通常不可能である．このような状況に対して，Duong 法と DNN を組み合わせた手法（以後，Duong+DNN 法と呼ぶ）が提案された [27]．本手法では，音源情報のモデル化には事前学習した DNN を用い，混合系に対応する空間情報は Duong 法の空間相関行列としてモデル化及びブラインドに推定することで，高精度な教師あり多チャネル音源分離を達成している．依然として空間情報をブラインドに推定するという観点では非常に合理的なアプローチであるが，フルランクの空間相関行列の推定アルゴリズムは従来の Duong 法と同様であるため，比較的大きい計算コストや初期値に対する分離性能の不安定性が問題となる．

本稿では，優決定条件かつ分離対象音源の学習データが得られるという条件下で，混合系ではなく分離系を高速に推定するアルゴリズムを提案し，最適化における改良アルゴリズムの有効性を実験的に評価する．本手法で用いる学習データとは「女性音声」や「ギター」等，特定の音源クラスに属するデータのことであり，観測信号に混合されている音源と全く同じ音源のサンプルでなくても良い．提案手法は，ILRMA における NMF 低ランク音源モデルを，対象音源を強調する DNN 音源モデルに置き換えた手法であり，教師あり音源モデルに基づいて分離系を高速かつブラインドに推定する．従って，ILRMA と MNMF の関係と同様に，提案手法は混合系を推定する Duong+DNN 法の双対な手法である．また，提案手法の空間モデルの最適化アルゴリズムには音源の更新順に関する任意性があり，この順番に応じて分離性能が大きく変化してしまう現象を実験的に確認する．この問題の解決手法として，より高精度な音源分離が達成される更新順を自動的に選択する手法を新たに提案する．音楽信号を用いた音源分離実験では，提案手法が従来手法よりも分離性能及び計算時間に関して優れていることを示す．

2. 従 来 手 法

2.1 定 式 化

音源数と観測チャネル数をそれぞれ N 及び M とし，時間領域の音源信号 $s_n(\tau)$，観測信号 $x_m(\tau)$，及び分離信号 $y_n(\tau)$ をそれぞれ短時間 Fourier 変換（short-time Fourier transform: STFT）して得られる複素時間周波数成分を次式で表す．

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

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

$$\boldsymbol{y}_{ij} = (y_{ij,1},\ \cdots,\ y_{ij,N})^{\mathrm{T}} \in \mathbb{C}^{N\times 1} \tag{3}$$

ここで，τ，$i=1,\cdots,I$，$j=1,\cdots,J$，$n=1,\cdots,N$，$m=1,\cdots,M$ はそれぞれ離散時間，周波数ビン，時間フレーム，音源，及びチャネルのインデクスを表し，$^{\mathrm{T}}$ は転置を表す．さらに，各信号の複素スペクトログラム行列を $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$，$\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$，及び $\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$ で表す．これらの行列の要素はそれぞれ $s_{ij,n}$，$x_{ij,m}$，及び $y_{ij,n}$ に一致する．混合系が線形時不変であり，時間周波数領域での複素瞬時混合で表現できると仮定すると，周波数毎の時不変な複素混合行列 $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1}\ \cdots\ \boldsymbol{a}_{i,N}) \in \mathbb{C}^{M\times N}$（$\boldsymbol{a}_{i,n} = (a_{i,n1},\cdots,a_{i,nM})^{\mathrm{T}}$ は各音源のステアリングベクトル）が定義でき，観測信号を次式で表現できる．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{4}$$

この混合モデルは，Duong 法における各音源の空間相関行列のランクが 1 という制約（ランク 1 空間モデル [18]）に対応し，時不変混合系の残響時間が STFT の窓長よりも十分短い場合に成立する．この時，$M=N$ かつ $\boldsymbol{A}_i$ が正則であれば，分離フィルタ $\boldsymbol{w}_{i,n} = (w_{i,n1},\cdots,w_{i,nM})^{\mathrm{T}}$ で構成される分離行列 $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1}\ \cdots\ \boldsymbol{w}_{i,N})^{\mathrm{H}} \in \mathbb{C}^{N\times M}$ が存在し，分離信号を次式で表現できる．

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{5}$$

ここで，$^{\mathrm{H}}$ はエルミート転置を示す．優決定条件では，式 (5) 中の分離行列 $\boldsymbol{W}_i$ を全周波数において推定することが最終的な目標となる．本稿では，以後，決定的な系 ($M=N$) を考える．

2.2 独立低ランク行列分析（ILRMA）

ILRMA [9]– [12] は優決定条件 BSS であり，空間モデル（分離行列 $\boldsymbol{W}_i$）と音源モデル（NMF に基づく各音源の低ランク時間周波数構造）を同時に推定する最適化問題である．音源スペクトログラム $\boldsymbol{Y}_n$ の生成モデルとして，次式の分布を仮定する．

$$p(\boldsymbol{Y}_n) = \prod_{i,j} p(y_{ij,n})$$
$$= \prod_{i,j} \frac{1}{\pi r_{ij,n}} \exp\left(-\frac{|y_{ij,n}|^2}{r_{ij,n}}\right) \quad (6)$$
$$r_{ij,n} = \sum_k t_{ik,n} v_{kj,n} \quad (7)$$

ここで，$p(y_{ij,n})$ は分散 $r_{ij,n}$ の原点対称な複素ガウス分布であり i，j，及び n に関して独立である．分散 $r_{ij,n}$ が時間と周波数に依存して変動するため，行列の生成モデル $p(\boldsymbol{Y}_n)$ は非ガウスな分布である．また，$t_{ik,n} \geqq 0$ 及び $v_{kj,n} \geqq 0$ はそれぞれ NMF の基底及びアクティベーションであり，$k = 1, \cdots, K$（$K \ll \min(I, J)$）は基底のインデクスを示す．すなわち，$r_{ij,n}$，$t_{ik,n}$，及び $v_{kj,n}$ を要素に持つ非負行列をそれぞれ $\boldsymbol{R}_n \in \mathbb{R}_{\geq 0}^{I \times J}$，$\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I \times K}$，及び $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{K \times J}$ とすると，式 (7) はパワースペクトログラム $|\boldsymbol{Y}_n|^{\cdot 2}$ を低ランク音源モデル（モデルパワースペクトログラム）$\boldsymbol{R}_n = \boldsymbol{T}_n \boldsymbol{V}_n$ で近似していることに対応する．ここで，ベクトルや行列に対する絶対値記号とドット付きの指数は要素毎の絶対値と指数乗をとった行列を表す．

式 (6) に基づく観測信号の負対数尤度は次式で与えられる．

$$\mathcal{L} \stackrel{\mathrm{c}}{=} -2J \sum_i \log|\det \boldsymbol{W}_i| + \sum_{i,j,n} \left(\frac{|y_{ij,n}|^2}{r_{ij,n}} + \log r_{ij,n}\right) \quad (8)$$

ここで，$\stackrel{\mathrm{c}}{=}$ は定数項を除いて等しいことを示し，$y_{ij,n} = \boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}$ である．式 (8) を最小化することで，$\boldsymbol{W}_i$，$\boldsymbol{T}_n$，及び $\boldsymbol{V}_n$ を全て推定できる．式 (8) の第一項及び第二項は，時変ガウス分布を仮定した IVA [6] のコスト関数に一致し，第二項及び第三項は板倉斎藤擬距離に基づく NMF（Itakura–Saito NMF: ISNMF）[28] のコスト関数に一致する．従って，$\boldsymbol{W}_i$ の更新には反復射影法（iterative projection: IP）[5], [6]，$\boldsymbol{T}_n$ 及び $\boldsymbol{V}_n$ の更新には ISNMF の乗算更新則 [28], [29] を交互に適用することで，全変数を容易に最適化できる．

Figure 1 (a) に ILRMA による音源分離の原理図を示す．混合前の音源のパワースペクトログラム（分散行列）$|\boldsymbol{S}_n|^{\cdot 2}$ は，混合信号の分散行列 $|\boldsymbol{X}_m|^{\cdot 2}$ よりも基本的に低ランクであることから，分離信号の分散行列を NMF で低ランクにモデル化（$|\boldsymbol{Y}_n|^{\cdot 2} \approx \boldsymbol{R}_n = \boldsymbol{T}_n \boldsymbol{V}_n$）して分離行列 $\boldsymbol{W}_i$ の最適化に反映させることで，パーミュテーション問題を解決している．

3. 提 案 手 法

3.1 動　　機

優決定条件 BSS では，音源間の独立性仮定の下で推定される空間モデル（分離行列）がパーミュテーション問題 [17] を起こさないために，何らかの仮定を導入した音源モデルが不可欠である．例えば，IVA は同一音源の周波数成分の共変性（時間周波数領域のグループスパース構造）を仮定し，ILRMA は同一音源の時間周波数成分の低ランク共変構造を仮定及び推定している．他にも，グループスパース構造とスパース構造の組み合わせや，低ランク構造とスパース構造の組み合わせ等の音源モデルがパーミュテーション問題の回避に有効である [30]．このような統計的，あるいは構造的性質の仮定（音源モデル）がその音源に対して適切であれば，パーミュテーション問題を起こすことなく空間モデルが最適化でき，高精度な音源分離が達成される．しかし，音声信号やボーカル信号等に対する低ランク構造仮定等，不適切な音源モデルを用いた場合には，パーミュテーション問題が解決されず分離性能は劣化する．音源の性質を陽に仮定せずとも適切な音源モデル $\boldsymbol{R}_n$ を推定できれば理想的であるが，BSS の枠組みでは非常に困難な問題である．

分離対象音源の十分な学習データが用意できる場合は，その音源に対して適切な音源モデルを構築することは比較的容易である．特に，教師あり学習において大きな成果を上げている DNN に基づく手法は，音源分離問題に対してもその有効性が多くの文献で示されている（例えば [22]–[27] 等）．一方，空間的な伝達系は，音源位置やマイクロホン位置，部屋の形状，残響時間等膨大な物理要因に依存することから，それらを網羅する学習データを用意することは非現実的である．従って，音源モデルには学習済の DNN を用い，空間モデルは従来通りブラインドに推定する手法が合理的である．DNN 音源モデルと Duong 法による空間相関行列の推定を組み合わせた Duong+DNN 法 [27] は，そのような手法の先駆けであるが，Duong 法に基づく空間相関行列の推定の計算コスト及び不安定性が解決すべき問題である．また Duong+DNN 法は，空間相関行列の更新回数に応じて異なる DNN 音源モデル（パラメタ初期化用 DNN，空間モデル更新後用 DNN 等）を導入している．これらの DNN 音源モデルの学習には，Duong 法による音源分離途中の信号を学習データとして集める必要があるため，多大なコストが要求される．

本研究では，優決定条件を対象とした効率的な教師あり音源分離の確立を目標とし，音源間の統計的独立性に基づくブラインドな空間モデル推定と DNN に基づく教師有り音源モデルを組み合わせた手法を提案する．提案手法は，空間モデル推定の過程で，DNN 音源モデルにより推定される時間周波数の分散行列 $\boldsymbol{R}_n$ を活用することから，以後，独立深層学習行列分析（independent deeply learned matrix analysis: IDLMA）と呼ぶ．IDLMA で用いる音源モデルは，Duong+DNN 法とは異なり，空間モデルの反復回数によらず共通の DNN 音源モデルを用いる．また，複数の音源の分散（$\boldsymbol{R}_1, \cdots, \boldsymbol{R}_N$）を出力する一つの DNN ではなく，混合信号から一つの音源の分散 $\boldsymbol{R}_n$ を出力する DNN を N 個利用することで，DNN 音源モデルの再利用性（ポータビリティ）を確保する．

3.2 独立深層学習行列分析（IDLMA）

3.2.1 処理の概要

Figure 1 (b) に IDLMA による音源分離の原理図を示す．IDLMA は ILRMA と同様に，式 (6) の生成モデルに基づいて音源モデル $\boldsymbol{R}_n$ 及び空間モデル $\boldsymbol{W}_i$ を推定する．このとき，混合信号から n 番目の音源の分散行列 $\boldsymbol{R}_n$（モデルパワースペクトログラム）を推定する DNN を事前に学習しておく．これを DNN_n とする．例えば，ボーカル信号とその他の雑多な音源が混合した信号を入力とし，ボーカル信号のみの分散行列を出力するように DNN を学習することで，ボーカル信号を強調する

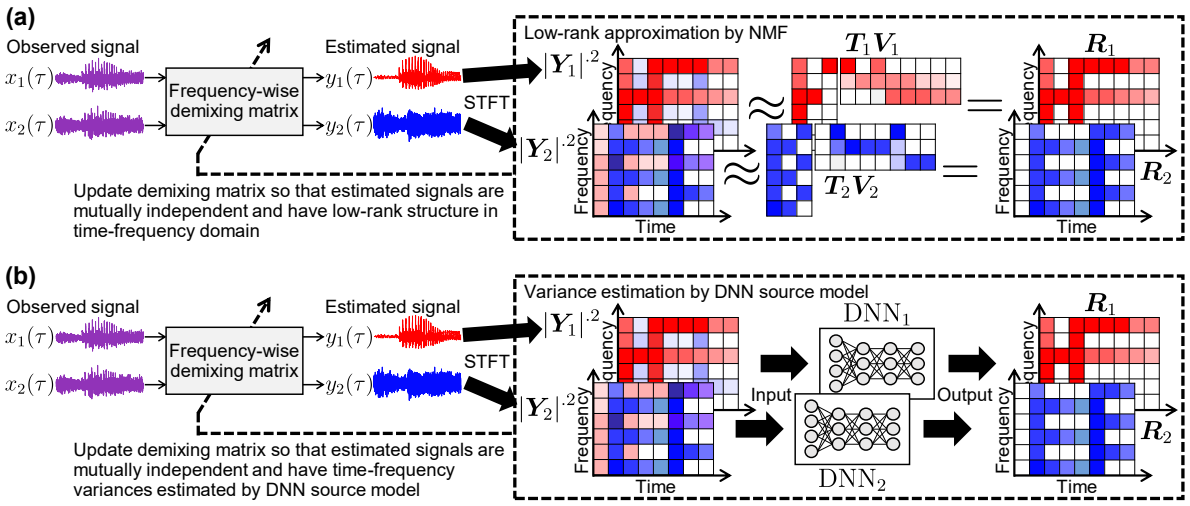

Fig. 1　Principle of source separation based on (a) ILRMA and (b) IDLMA.

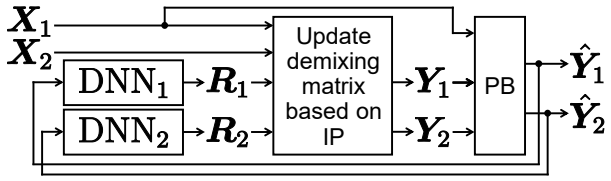

Fig. 2　Process flow of IDLMA in two sources case, where first channel is used as reference for PB.

DNN 音源モデルが得られる．このような特定の音源を強調する DNN を全音源（$\mathrm{DNN}_1, \cdots, \mathrm{DNN}_N$）に対して学習することで，低ランク性やスパース性等の陽なモデルではなく学習データから得た適切な音源モデルを構築でき，より高精度な分散行列 $\boldsymbol{R}_n$ 及び分離行列 $\boldsymbol{W}_i$ の推定に活用できる．

IDLMA の処理の流れを Fig. 2 に示す．分離行列 $\boldsymbol{W}_i$ は観測信号 $\boldsymbol{X}_m$ と推定分散行列 $\boldsymbol{R}_n$ を用いて IP によって更新され，暫定的な分離信号 $\boldsymbol{Y}_n$ が得られる．空間モデルの推定には周波数毎のスケールの任意性があるため，$\boldsymbol{Y}_n$ に対してリファレンスチャネルを用いたプロジェクションバック法（projection-back technique: PB）[31] を適用し，スケール補正後の分離信号 $\hat{\boldsymbol{Y}}_n \in \mathbb{C}^{I\times J}$ を得る．反復最適化の初期は分離が不十分であるため，$\hat{\boldsymbol{Y}}_n$ は他の音源成分が多く残留している．これを混合信号とみなし，音源モデル DNN_n に入力することで，より分離が進んだ推定分散行列 $\boldsymbol{R}_n$ が得られ，これを再び分離行列 $\boldsymbol{W}_i$ の更新に用いる．このプロセスを繰り返すことで，より高精度な $\boldsymbol{W}_i$ が推定される．なお，分散行列 $\boldsymbol{R}_n$ の初期値には，観測信号のリファレンスチャネルを DNN_n に入力したときの出力をそのまま用いることができる．なお，IDLMA の最終的な出力は $\hat{\boldsymbol{Y}}_n$ であり，これは時不変線形空間フィルタ $\boldsymbol{w}_{i,n}$ の出力であることから，IDLMA は ILRMA と同様に歪みの少ない分離信号が得られる．このような出力は，例えば音声認識システムにとって望ましいという利点がある [32]．

3.2.2　IP に基づく分離行列の更新と PB

まず，音源モデル（分散 $r_{ij,n}$）が与えられたと仮定する．この下で，各音源が独立になるように分離行列 $\boldsymbol{W}_i$ を更新する．分離フィルタ $\boldsymbol{w}_{i,n}$ は IP [5], [6] を用いて次式で更新できる．

$$\boldsymbol{U}_{i,n} = \frac{1}{J}\sum_j \frac{1}{r_{ij,n}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} \tag{9}$$

$$\boldsymbol{w}_{i,n} \leftarrow \left(\boldsymbol{W}_i\boldsymbol{U}_{i,n}\right)^{-1}\boldsymbol{e}_n \tag{10}$$

$$\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n}\left(\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{U}_{i,n}\boldsymbol{w}_{i,n}\right)^{-\frac{1}{2}} \tag{11}$$

ここで，$\boldsymbol{e}_n \in \mathbb{R}_{\{0,1\}}^{N\times 1}$ は n 番目の要素が 1，他の要素が 0 のベクトルである．分離フィルタ $\boldsymbol{w}_{i,n}$ 更新後は，分離信号を $y_{ij,n} \leftarrow \boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij}$ として更新する．このように推定される分離信号 $\boldsymbol{Y}_n$ は，周波数毎のスケールが揃っていないため，リファレンスチャネルのスケールに合わせる PB を適用する．

$$\hat{y}_{ij,n} = \left[\boldsymbol{W}_i^{-1}\left(\boldsymbol{e}_n \circ \boldsymbol{y}_{ij}\right)\right]_{m_{\mathrm{ref}}} \tag{12}$$

ここで，$\hat{y}_{ij,n}$ は $\hat{\boldsymbol{Y}}_n$ の要素，$\circ$ は要素毎の積，$[\cdot]_m$ は m 番目の要素値，m_{ref} はリファレンスチャネルのインデクスをそれぞれ表す．以上により得られる分離信号 $\hat{\boldsymbol{Y}}_n$ を，音源モデル DNN_n に入力する．

3.2.3　DNN 音源モデルを用いた分散の推定

各音源の DNN は，分離対象となる音源信号の学習データ $\tilde{\boldsymbol{S}}_n \in \mathbb{C}^{I\times J}$ とその他の音源信号の学習データ $\tilde{\boldsymbol{S}}_{n'}$（$n' \neq n$）を任意の比率で混合した信号 $\tilde{\boldsymbol{X}} \in \mathbb{C}^{I\times J}$ の振幅値 $|\tilde{\boldsymbol{X}}|^{\cdot 1}$ を入力とし，混合前の音源信号の振幅値 $|\tilde{\boldsymbol{S}}_n|^{\cdot 1}$ を予測するように学習する．$|\tilde{\boldsymbol{X}}|^{\cdot 1}$ を入力した DNN_n の出力信号（$|\tilde{\boldsymbol{S}}_n|^{\cdot 1}$ の推定値）を $\boldsymbol{D}_n = \mathrm{DNN}_n(|\tilde{\boldsymbol{X}}|^{\cdot 1}) \in \mathbb{R}_{\geq 0}^{I\times J}$ と表すと，DNN_n を学習するための損失関数には，板倉斎藤擬距離の分子と分母に微小値 δ_1 を加えた次式を用いる [27]．

$$L = \frac{1}{IJ}\sum_{i,j}\left(\frac{|\tilde{s}_{ij,n}|^2+\delta_1}{d_{ij,n}^2+\delta_1} - \log\frac{|\tilde{s}_{ij,n}|^2+\delta_1}{d_{ij,n}^2+\delta_1} - 1\right) \tag{13}$$

ここで，$\tilde{s}_{ij,n}$ 及び $d_{ij,n}$ はそれぞれ $\tilde{\boldsymbol{S}}_n$ 及び $\boldsymbol{D}_n$ の要素である．式 (13) を最小化する DNN を学習することは，生成モデル (6) に基づく分散 $r_{ij,n}$ の最尤推定と等価 [28] であることから，DNN_n は混合信号から分離対象音源の分散行列を推定するネットワークと解釈できる．

学習済みの DNN 音源モデルを用いた分散 $r_{ij,n}$ の更新は次式となる．

$$|\hat{\boldsymbol{R}}_n|^{\cdot\frac{1}{2}} = \mathrm{DNN}_n(|\hat{\boldsymbol{Y}}_n|^{\cdot 1}) \tag{14}$$

$$r_{ij,n} \leftarrow \max(\hat{r}_{ij,n}, \varepsilon) \tag{15}$$

ここで，$\hat{r}_{ij,n}$ は DNN_n の推定分散行列 $\hat{\boldsymbol{R}}_n \in \mathbb{R}_{\geq 0}^{I \times J}$ の要素，ε は IP の安定性を向上させるための微小値である．本稿で用いる DNN 音源モデルの入出力特徴量とアーキテクチャの詳細については 4.2 節で示す．

4. 評価実験

4.1 実験条件

IDLMA の性能評価のため，ILRMA，学習済みの DNN_n の全音源（$n=1,\cdots,N$）分の出力を用いて Wiener フィルタを構成し観測信号に適用する手法（DNN+WF）[24]，Duong+DNN 法，及び IDLMA の 4 手法について，音楽信号を対象とした 2 音源の分離性能を比較した．実験には，SiSEC2016 [33] の音楽信号データセット DSD100 中のボーカル（Vo.），ベース（Ba.），ドラム（Dr.）の 3 音源を用いた．DSD100 の dev データ 50 曲を DNN 音源モデルの学習データとし，test データ 50 曲の内アルファベット順の上位 25 曲のそれぞれについて，30 秒から 60 秒の区間を音源分離の対象（評価データ）とした．但し，当該区間がすべて無音となる楽曲については，60 秒から 90 秒の区間とした．多チャネル観測信号は，RWCP データベース [34] に収録されている E2A インパルス応答と音源信号を畳み込み混合して生成した．実験に用いた 2 種のインパルス応答（以後，IR1 及び IR2 と呼ぶ）を Fig. 3 に示す．音源の混合は Ba./Vo. と Dr./Vo. の 2 種とし，サンプリング周波数は 8 kHz に落として実験した．ILRMA の基底数は $K=20$ とし，ILRMA，Duong+DNN 法，及び IDLMA の最適化は空間モデルの反復更新を 100 回行った時点で終了とした．また，Duong+DNN 法及び IDLMA は，空間モデルの最適化を 10 回反復する毎に DNN 音源モデルを 1 回適用した．音源分離性能の客観評価尺度には signal-to-distortion ratio (SDR) [35] を用いた．

4.2 DNN の構造と学習方法

DNN は全結合型ニューラルネットワークを用いた．隠れ層は 4 層，各隠れ層のユニット数は STFT の窓長によらず 1024 とし，各隠れ層及び出力層に対し，活性化関数として rectified linear unit [36] を用いた．また，DNN の学習時の入出力特徴量には，Fig. 4 に示す次のベクトルを用いた．

$$\vec{\boldsymbol{x}}_j = \frac{\sum_n \alpha_{j,n} \vec{\boldsymbol{s}}_{j,n}}{\|\sum_n \alpha_{j,n} \vec{\boldsymbol{s}}_{j,n}\|_2 + \delta_2} \in \mathbb{C}^{I(2c+1)\times 1} \tag{16}$$

$$\bar{\boldsymbol{s}}_{j,n} = \frac{\alpha_{j,n} \tilde{\boldsymbol{s}}_{j,n}}{\|\sum_n \alpha_{j,n} \vec{\boldsymbol{s}}_{j,n}\|_2 + \delta_2} \in \mathbb{C}^{I\times 1} \tag{17}$$

$$\vec{\boldsymbol{s}}_{j,n} = [\tilde{\boldsymbol{s}}_{j-2c,n}^{\mathrm{T}}, \tilde{\boldsymbol{s}}_{j-2c+2,n}^{\mathrm{T}}, \cdots, \tilde{\boldsymbol{s}}_{j+2c,n}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{I(2c+1)\times 1} \tag{18}$$

ここで，$\vec{\boldsymbol{x}}_j$ 及び $\bar{\boldsymbol{s}}_{j,n}$ はそれぞれ時間フレーム j において同じ値で正規化された混合信号ベクトル及び分離対象音源ベクトルであり，DNN_n の入力ベクトル及び出力ベクトルはそれぞれ $|\vec{\boldsymbol{x}}_j|^{\cdot 1}$ 及び $|\bar{\boldsymbol{s}}_{j,n}|^{\cdot 1}$ である．また，$\alpha_{j,n}$ は $[0.05, 1]$ の一様乱数に従う確率変数であり，DNN による音源分離が様々な混合比に対する汎化性を獲得することを期待している [24]．$\tilde{\boldsymbol{s}}_{j,n} \in \mathbb{C}^I$ は音源信号 $\tilde{\boldsymbol{S}}_n$ の時間フレーム j における周波数ベクトルであり，$\vec{\boldsymbol{s}}_{j,n}$ は時間フレーム j を中心として 1 フレームおきに前後のフレームを c 個ずつ集めて結合したベクトルである．さらに，$\|\cdot\|_2$ は L_2 ノルム，δ_2 は零除算を防ぐための微小値である．なお，IDLMA の反復更新において式 (14) を適用する場合も，$\hat{\boldsymbol{Y}}_n$ に対して式 (18) と同様の操作で前後の時間フレームを結合したベクトル $\vec{\hat{\boldsymbol{y}}}_{j,n} \in \mathbb{C}^{I(2c+1)\times 1}$ を作成し，$|\vec{\hat{\boldsymbol{y}}}_{j,n}|^{\cdot 1}$ を DNN_n に入力する．

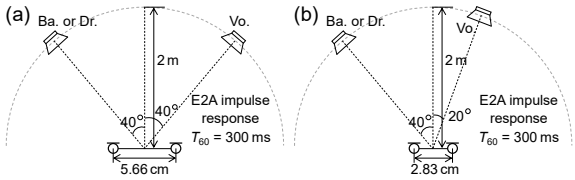

Fig. 3 Impulse responses: (a) IR1 and (b) IR2.

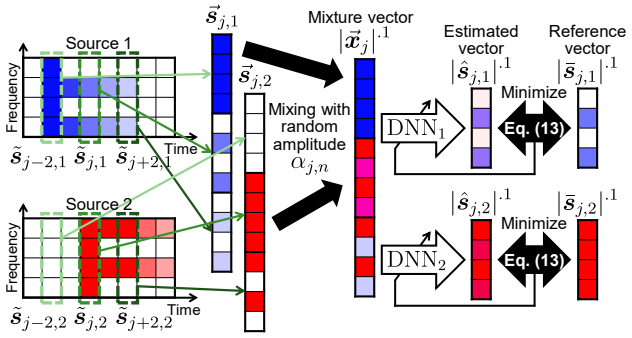

Fig. 4 Learning process of DNN source model and its input and output vectors, where $I=4$, $J=8$, $N=2$, and $c=1$.

DNN の最適化には ADADELTA [37] を用い，ミニバッチサイズを 128，エポック数を 1000 として誤差逆伝播学習を行った．また，過学習を避けるための正則化項として $(\lambda/2)\sum_q g_q^2$ を損失関数 (13) に加えた．ここで，g_q は DNN の重み係数を表す．その他のパラメタは，$\delta_1=\delta_2=10^{-5}$，$\varepsilon=10^{-1}\times(IJ)^{-1}\sum_{i,j}\hat{r}_{ij,n}$，$c=3$，及び $\lambda=10^{-5}$ とし，ADADELTA のハイパーパラメタは $\rho=0.95$ 及び $\epsilon=10^{-6}$ とした．

4.3 実験結果

4.3.1 STFT の窓長と各手法の性能の関係

ILRMA と IDLMA は，いずれも式 (4) のランク 1 空間モデルに基づく分離であり，この仮定は STFT の窓長が長いほど（あるいは収録環境の残響長が短いほど）妥当なものとなる [11]．一方で，フルランクの空間相関行列を仮定する Duong+DNN 法は，式 (4) で表現できない混合系も取り扱えることから，その推定の困難さはさておき，原理的にはランク 1 空間モデルよりも高精度に分離できる．従って両者の違いは，窓長に対する空間モデル由来の制約の有無である．

本項では，窓長に対する各手法の性能を実験的に確認するために，128 ms，256 ms，512 ms，及び 1024 ms の 4 種の窓長を用いて分離を行った．但し，シフト長は常に窓長の半分とした．DNN+WF 法，Duong+DNN 法，及び IDLMA に関しては，各窓長に対して DNN 音源モデルの学習を行い，共通の音源モデルで分離性能を評価した．さらに，ランク 1 空間モデル (4) を仮定した場合の理想的な分離フィルタ（ideal filter）の性能

Table 1　Average SDR improvements (dB) for Ba./Vo. separation with various window lengths

Impulse response	Method	Window length in STFT			
		128 ms	256 ms	512 ms	1024 ms
IR1	Ideal filter	18.08	19.78	21.18	**22.64**
	ILRMA	3.40	4.65	**6.53**	4.91
	DNN+WF	7.07	7.34	**7.76**	6.66
	Duong+DNN	9.25	10.60	**11.14**	8.82
	IDLMA	10.88	12.90	**13.48**	11.55
IR2	Ideal filter	14.32	15.63	16.60	**17.43**
	ILRMA	2.36	3.60	**5.82**	2.86
	DNN+WF	6.84	7.35	**7.66**	6.85
	Duong+DNN	8.35	9.89	**10.67**	8.15
	IDLMA	11.08	11.60	**12.51**	9.62

Table 2　Average SDR improvements (dB) for Dr./Vo. separation with various window lengths

Impulse response	Method	Window length in STFT			
		128 ms	256 ms	512 ms	1024 ms
IR1	Ideal filter	16.35	17.37	18.21	**19.25**
	ILRMA	6.88	**8.13**	7.46	6.20
	DNN+WF	2.67	**3.35**	3.31	2.00
	Duong+DNN	3.75	**9.58**	9.05	6.17
	IDLMA	6.82	**10.91**	8.66	6.52
IR2	Ideal filter	12.21	12.96	13.43	**13.96**
	ILRMA	4.01	4.65	**4.90**	3.44
	DNN+WF	3.20	**4.21**	3.88	2.57
	Duong+DNN	4.02	**9.62**	9.04	5.94
	IDLMA	6.93	**9.24**	6.93	5.53

も同時に示した．この理想的な分離フィルタは，音源信号 $\boldsymbol{S}_n$ と分離信号 $\boldsymbol{Y}_n$ 間の歪み最小基準より次式で与えられる．

$$\boldsymbol{W}_i = \boldsymbol{S}_i \boldsymbol{X}_i^{\mathrm{H}} \left(\boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{H}} \right)^{-1} \tag{19}$$

ここで，$\boldsymbol{S}_i \in \mathbb{C}^{N \times J}$ 及び $\boldsymbol{X}_i \in \mathbb{C}^{M \times J}$ はそれぞれ $s_{ij,n}$ 及び $x_{ij,m}$ を要素にもつ行列である．Duong+DNN 法は，原理的にはこの理想的な分離フィルタの性能を上回ることができる点に注意する．

Tables 1 及び 2 は，各窓長での音源分離結果をテストデータ 25 曲に関して平均した SDR 改善量である．前述の通り，理想的な分離フィルタは全てのデータに対して窓長が長いほど性能が向上している．ILRMA は，ランク 1 空間モデルの妥当性と統計バイアスの影響のトレードオフ [11] により，Ba./Vo. に対しては 512 ms，Dr./Vo. に対しては 256 ms か 512 ms の窓長が最適となっている．DNN+WF は単一チャネルの音源分離手法であるが，窓長の増加に伴って DNN 音源モデルの入出力特徴量の次元が増加する．今回の実験では，隠れ層の層数やユニット数は窓長によらず一定としたため，長い窓長では学習が困難となり，性能が劣化している．この DNN 音源モデルの性能劣化は Duong+DNN 法及び IDLMA の分離性能にも影響する．従って IDLMA は，空間モデルの妥当性，統計バイアス，及び DNN 音源モデルの学習困難性の 3 つの観点から，窓長に関する性能のトレードオフが存在する．Duong+DNN 法と IDLMA を比較すると，原理的にはランク 1 空間モデルを用いない Duong+DNN 法が有利であるが，フルランクの空間相関行列の推定が困難であることに起因して，Dr./Vo. の IR2 以外のデータで IDLMA の性能が上回っていることが確認できる．

4.3.2　反復更新毎の性能比較

各手法の反復更新に対する全 25 曲の平均 SDR 改善量を Fig. 5 に示す．但し，Ba./Vo. は 512 ms，Dr./Vo. は 256 ms の窓長を用いた場合の結果である．DNN+WF は反復手法ではないため，SDR 改善量の値を水平線で示している．結果より，IDLMA は DNN 音源モデルを通す度（反復 10 回毎）に推定分散の改善がみられ，それに追従して分離行列もより良い解に誘導されていく様子が確認できる．Duong+DNN 法も同様の傾向が確認できるが，反復回数 50 回以降では DNN 音源モデルが SDR を下げてしまう現象が確認でき，多くの場合で IDLMA よりも悪い性能に収束している．

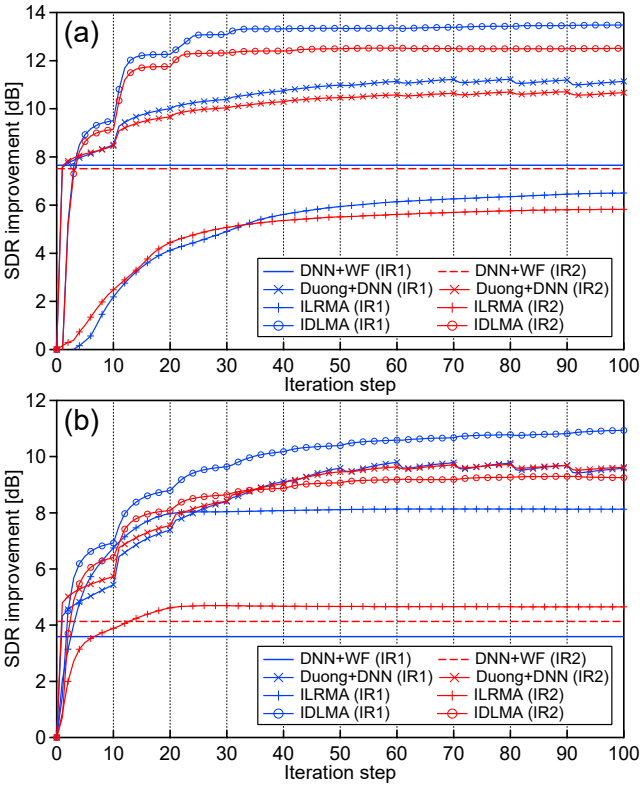

Fig. 5　Average SDR improvements in each iteration step for (a) Ba./Vo. separation with 512-ms-long window and (b) Dr./Vo. separation with 256-ms-long window.

4.3.3　計算量比較

ILRMA や IDLMA における空間モデル $\boldsymbol{W}_i$ の更新には式 (9)–(11) に示す IP を用いている．この反復更新にはサイズ N の行列 $\boldsymbol{W}_i \boldsymbol{U}_{i,n}$ の逆行列演算が含まれているため，全音源全周波数の IP による更新はおよそ $O(IN^4)$ の計算量が必要となる．一方で，Duong 法の空間相関行列の expectation-maximization（EM）アルゴリズムに基づく反復更新 [18] には，E ステップで各時間周波数に対してサイズ M の逆行列演算が必要であり，さらに M ステップでサイズ M の逆行列演算が音源と周波数毎に必要になるため，それぞれのステップでおよそ $O(IJM^3)$ と $O(INM^3)$ の計算量が必要となる．従って，$N = M$ の場合には IP に基づく手法の方が高速である．Table 3 は，30 秒の信号に対して空間モデルを 100 回更新した際の各手法の実際の計

Table 3　Examples of computational time for 100 spatial updates in each method

Method	Computational time [s]
ILRMA	23.31
Duong+DNN	287.06
IDLMA	26.56

算時間例を手法毎に示している．但し，計算には Python 3.5.2 と Chainer 2.1.0 の環境で，Intel Core i7-6850K（3.60 GHz，6 Cores）の CPU を用いている．また，DNN 音源モデルによる分散推定には GeForce GTX 1080 Ti の GPU を用いている．この結果より，IDLMA の計算時間は ILRMA とほぼ同程度，Duong+DNN 法よりも 10 倍以上高速であることが分かる．

5. IP における分離フィルタの更新順の影響と適切な更新順の自動選択

5.1 分離フィルタの更新順の影響

前節の実験では，Ba. あるいは Dr. を $n=1$，Vo. を $n=2$ とし，式 (9)–(11) の IP による分離行列 $\boldsymbol{W}_i$ の更新を $n=1,2$ という順番（インデクスの昇順）で行った場合の結果を示した．この IP における分離フィルタ $\boldsymbol{w}_{i,n}$ の更新の順番には任意性があり，コスト関数 (8) が $\boldsymbol{w}_{i,n}$ に対して非凸であることから，最終的に得られる解は更新する順番に依存する．特に DNN 音源モデルを用いる IDLMA では，得られる分離結果がこの更新の順番に強く依存してしまう．例えば，IR1 を用いた Ba/Vo. の混合信号では，$n=1,2$ の昇順で IP を適用した場合の SDR 改善量が 13.48 dB であるのに対し，$n=2,1$ の降順で IP 適用した場合は 11.05 dB となった．

この現象の原因として，DNN 音源モデルによる分散行列 $\boldsymbol{R}_n$ の推定精度が関連していると考えられる．即ち，1 回の分離行列 $\boldsymbol{W}_i$ の更新の中で逐次的に $\boldsymbol{w}_{i,n}$ を更新する IP では，分散行列 $\boldsymbol{R}_n$ がより高精度に推定された音源の分離フィルタを先に更新することが望ましい．今回の実験では，複雑な周波数変動を含む Vo. 信号よりも，簡素な時間周波数構造を持つ Ba. や Dr. の方が高精度に分散行列 $\boldsymbol{R}_n$ を推定できる．事実として，学習データに対する損失関数の収束値は Vo. よりも Ba. や Dr. の方が小さな値となっている．従って，IDLMA の IP の計算では，Ba. や Dr. の音源（$n=1$）の分離フィルタ $\boldsymbol{w}_{i,n}$ を先に更新する方がより良い解へと収束する可能性が高い．

5.2 適切な更新順の選択基準

より良い音源分離性能をもたらす分離フィルタの更新順を自動的に選択するために，何らかの妥当な選択基準を設けることが望ましい．本稿では，IDLMA の一時的な分離信号 $\hat{\boldsymbol{Y}}_n$ を音源モデル $\mathrm{DNN}_{n'}$ に入力したときに得られる分散行列を $\check{\boldsymbol{R}}_{nn'}=|\mathrm{DNN}_{n'}(|\hat{\boldsymbol{Y}}_n|^{\cdot 1})|^{\cdot 2}$ としたとき，更新順の選択基準として次式で定義される ζ 及び ξ の 2 種類を提案する．

$$\zeta = \frac{1}{N}\sum_{n}\frac{\sum_{i,j}\check{r}_{ij,nn'}}{\sum_{n'}\sum_{i,j}\check{r}_{ij,nn'}} \tag{20}$$

$$\xi = \frac{1}{N}\sum_{i,j,n}\frac{\check{r}_{ij,nn'}}{\sum_{n'}\check{r}_{ij,nn'}} \tag{21}$$

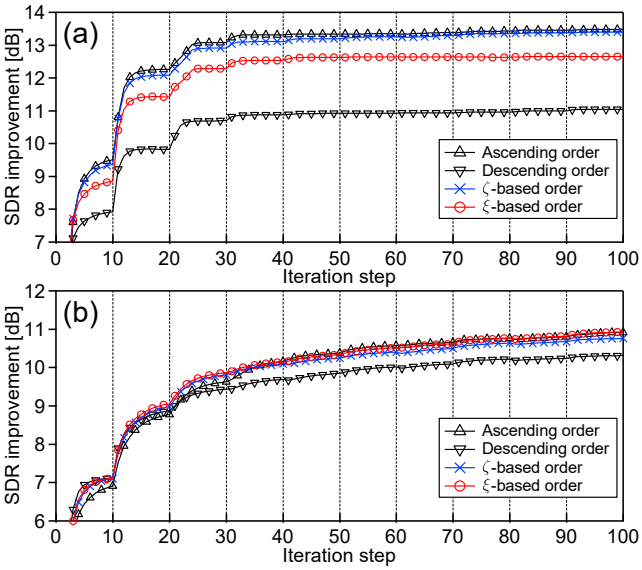

Fig. 6　Average SDR improvements in each iteration step based on several update orders: (a) Ba./Vo. separation with 512-ms-long window and (b) Dr./Vo. separation with 256-ms-long window.

ここで，$\check{r}_{ij,nn'}$ は $\check{\boldsymbol{R}}_{nn'}$ の要素である．式 (20) は分離信号 $\hat{\boldsymbol{Y}}_n$ に含まれる音源 n 成分の割合の推定値を全音源に関して平均した指標であり，式 (21) は各時間周波数要素の Wiener フィルタ値を全音源に関して平均した指標である．IDLMA の反復更新の中で，全通りの分離フィルタの更新順（$N!$ 通り）を試行し，そのそれぞれの更新順で得られる分離信号 $\hat{\boldsymbol{Y}}_n$ から ζ あるいは ξ を計算して値の大きくなる結果を採用することで，より高精度な分離ができる解へと導くことができる．

5.3 更新順の自動選択の実験

4 章と同一の条件で，ζ 及び ξ を用いた更新順の自動選択を行った場合の IDLMA の分離結果を Fig. 6 に示す．但し，IR1 のインパルス応答を用いた結果のみ示している．本実験の結果では，5.1 節で述べた理由から，常に昇順での分離フィルタ更新がテストデータ 25 曲のほぼ全てで最良の結果となったが，Ba./Vo. のデータに対しては ζ に基づく更新順の選択，Dr./Vo. のデータに対しては ξ に基づく更新順の選択が最良と同程度の分離性能となっており，提案する選択基準の有効性が確認できる．

6. ま と め

本稿では，音源間の統計的独立性と DNN に基づく音源モデルを用いた教師あり音源分離手法の IDLMA を提案した．IDLMA は，対象の音源のみを強調する DNN を教師あり音源モデルとして活用しているが，空間モデルは依然としてブラインドに推定できるため，汎用性の高いアルゴリズムである．また，分離フィルタの更新順に対する適切な選択基準を提案した．DNN 音源モデルに基づく従来手法と実験的に比較し，分離性能と計算効率の観点から IDLMA の有効性を実証した．

謝辞　本研究の一部は SECOM 科学技術支援財団，JSPS 科研費 16H01735，17H06101，及び 17H06572 の助成を受けたものである．

文　　献

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

[2] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," *Proc. ICA*, pp. 601–608, 2006.

[3] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: an extension of ICA to multivariate components," *Proc. ICA*, pp. 165–172, 2006.

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

[5] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," *Proc. WASPAA*, pp. 189–192, 2011.

[6] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," *Proc. APSIPA*, 2012.

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

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization" *Proc. NIPS*, pp. 556–562, 2000.

[9] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[10] D. Kitamura, "Algorithms for independent low-rank matrix analysis," [Online]. Available: `http://d-kitamura.net/pdf/misc/AlgorithmsForIndependentLowRankMatrixAnalysis.pdf`

[11] D. Kitamura, N. Ono, and H. Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," *Proc. EUSIPCO*, pp. 1210–1214, 2017.

[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," *Audio Source Separation*, Shoji Makino, Ed. ch. 6, 31 pages, Springer, March 2018 (in press).

[13] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[14] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. SAP*, vol. 11, no. 2, pp. 109–116, 2003.

[15] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," *Proc. ICASSP*, 2004, pp. III-885–III-888.

[16] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[17] S. Kurita, H. Saruwatari, S. Kajita K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. ICASSP*, vol. 5, pp. 3140–3143, 2000.

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

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

[20] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[21] A. R. López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, "Designing multichannel source separation based on single-channel source separation," *Proc. ICASSP*, pp. 469–473, 2015.

[22] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," *Proc. ICASSP*, pp. 3734–3738, 2014.

[23] P.-S. Huang, M. Kim, M. H.-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. ASLP*, vol. 23, no. 12, pp. 2136–2147, 2015.

[24] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," *Proc. ICASSP*, pp. 2135–2139, 2015.

[25] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," *Proc. ICASSP*, pp. 116–120, 2015.

[26] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," *Proc. ICASSP*, pp. 286–290, 2017.

[27] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks ," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[28] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[29] Y. Mitsui, D. Kitamura, N. Takamune, H. Saruwatari, Y. Takahashi, and K. Kondo, "Independent low-rank matrix analysis based on parametric majorization-equalization algorithm," *Proc. CAMSAP*, pp. 98–102, 2017.

[30] K. Yatabe and D. Kitamura, "Determined blind source separation via proximal splitting algorithm," *Proc. ICASSP*, 2018 (in press).

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

[32] M. Mimura, Y. Bando, K. Shimada, S. Sakai, K. Yoshii, and T. Kawahara, "Combined multi-channel NMF-based robust beamforming for noisy speech recognition," *Proc. Interspeech*, pp. 2451–2455, 2017.

[33] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," *Proc. LVA/ICA*, pp. 323–332, 2012.

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

[35] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[36] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," *Proc. AISTATS*, pp. 315–323, 2011.

[37] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv*, 2012. [Online]. Available: `http://arxiv.org/abs/1212.5701`