第33回 信号処理シンポジウム
2018年11月6日〜8日（東京電機大学 東京千住キャンパス）

ヘビーテイル生成モデルに基づく独立深層学習行列分析による多チャネル音源分離

Multichannel audio source separation using independent deeply learned matrix analysis based on heavy-tailed generative model

牧島直輝† 最上伸一† 高宗典玄† 北村大地†† 猿渡洋† 高橋祐††† 近藤多伸††† 中嶋広明†††
†東京大学 ††香川高専 †††ヤマハ株式会社

Naoki MAKISHIMA† Shinichi MOGAMI† Norihiro TAKAMUNE† Daichi KITAMURA††
Hiroshi SARUWATARI† Yu TAKAHASHI††† Kazunobu KONDO††† Hiroaki NAKAJIMA†††
†The University of Tokyo ††National Institute of Technology, Kagawa College
†††Yamaha Corporation

アブストラクト 独立深層学習行列分析（IDLMA）は，事前に学習済みの deep neural network による音源の時間周波数構造の推定と音源間の統計的独立性を用いる多チャネル音源分離手法である。従来の IDLMA では生成モデルとして時変複素ガウス分布を仮定するが，本稿ではヘビーテイルで外れ値に頑健な生成モデルに基づく IDLMA を提案する。音楽信号を用いた音源分離実験により，提案手法の有用性を示す。

1 はじめに

音源分離とは複数の音源が混合された観測信号から混合前の各音源信号を推定する技術である。音源分離時に観測信号の情報のみを用いる手法は特にブラインド音源分離（blind source separation: BSS）と呼ばれる。

観測マイク数が音源数以上である優決定条件では，混合系の逆システムである分離系を推定する手法が主流である。分離系の推定手法として，各音源信号の独立性を利用した独立成分分析 [1] に基づく手法及びその拡張手法が発展してきた。特に，独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[2] は，音源の詳細な時間周波数構造を非負値行列因子分解（nonnegative matrix factorization: NMF）[3] による低ランク行列として捉えながら線形分離フィルタを学習するものであり，様々な BSS 手法の中で最も高い分離性能を示している。また，[2] では音源の生成モデルが時変複素ガウス分布に従うことを仮定しているが，生成モデルを時変複素ガウス分布から時変複素 Student's t 分布に一般化することで更に高精度な分離性能が得られることが知られている [4]。

一方，観測マイク数が音源数未満である劣決定条件では，Duong モデル [5] に基づき混合系を推定する手法が主流である。Duong モデルでは，各音源の空間的な情報を表す空間相関行列を各周波数ビン毎に EM アルゴリズムにより推定する。Duong モデルに基づいた音源分離手法として，音源の時間周波数構造を NMF により低ランクモデル化する多チャネル NMF（multichannel NMF: MNMF）[6], [7] が知られている。混合系を推定する手法は時変な分離フィルタを扱うことができる一方で，分離系を推定する手法に比べて数値的に不安定であり，実験的には分離性能が低いことが知られている [2]。

また，BSS 手法とは異なるが，音源モデル推定に学習済みの deep neural network（DNN）を利用する手法も近年盛んに研究されている [8]–[10]。これらの研究では，音源に関してその種類を既知とし学習データを仮定する一方，混合系は BSS と同様に未知とする。これは，空間的な伝達系がマイクの位置や部屋の形状などの膨大な要因に依存して変動し，DNN による学習が困難なためである。[8] では，劣決定条件の下で混合系を Duong モデルと仮定しており，音源の時間周波数構造を学習済みの DNN により推論する。[8] の手法を以降 Duong+DNN 法と呼ぶ。[9], [10] では，優決定条件の下で分離系を推定する独立深層学習行列分析（independent deeply learned matrix analysis: IDLMA）を提案している。IDLMA では音源の生成モデルを時変複素ガウス分布と仮定し，観測信号の尤度を最大化するように分離系の推定と学習済み DNN による音源モデルの推論を交互に行う。ここで学習済み DNN とは，入力の混合信号から各音源信号を強調して出力するネットワークである。IDLMA は従来の BSS 手法に比べて高精度な分離性能を示すが，生成モデルの変更が DNN の学習や推論に与える影響は明らかにされていない。

以上より，本稿では優決定条件及び音源モデル教師あ

り音源分離手法を対象として，モデルの柔軟性を増加させ分離性能を向上させるため IDLMA の生成モデルを時変複素 Student's t 分布に一般化する。複素 Student's t 分布はその特殊形として従来の複素ガウス分布を含み，形状パラメータを変化させることでよりヘビーテイルな形状へと制御することが可能である。本稿では提案法の有効性を実験により示す。

2 従来手法

2.1 定式化

音源信号，観測信号，分離信号の短時間フーリエ変換（short-time Fourier transform: STFT）をそれぞれ

$$\boldsymbol{s}_{ij} = (s_{ij1}, \cdots, s_{ijN})^{\mathrm{T}} \in \mathbb{C}^N \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ijM})^{\mathrm{T}} \in \mathbb{C}^M \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij1}, \cdots, y_{ijN})^{\mathrm{T}} \in \mathbb{C}^N \tag{3}$$

と表す。ここで，$i = 1, \ldots, I$，$j = 1, \ldots, J$ はそれぞれ周波数ビン，時間フレームであり $m = 1, \ldots M$，$n = 1, \ldots, N$ はそれぞれチャネル，音源のインデックスである。また $^{\mathrm{T}}$ はベクトル及び行列の転置を表す。

観測信号が周波数ビン毎の時不変な混合行列 $\boldsymbol{A}_i$ を用いて

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

と表せる時，優決定条件では分離信号は次式で与えられる。

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

ここで $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iN})^{\mathrm{H}}$ は空間モデルを表す分離行列であり，$\boldsymbol{w}_{in}^{\mathrm{H}}$ は第 n 音源の分離フィルタである。また $^{\mathrm{H}}$ はエルミート転置を表す。

2.2 ILRMA 及びその複素 Student's t 分布への拡張

ILRMA は多チャネル BSS 手法の 1 つであり，各音源信号の時間周波数構造を低ランク近似することで，周波数ビン毎の分離信号の混在（パーミュテーション問題）を回避する。[2] では y_{ijn} を要素に持つ音源スペクトログラム $\boldsymbol{Y}_n$ の生成モデルとして，以下の時変複素ガウス分布を仮定する。

$$\begin{aligned} p(\boldsymbol{Y}_n) &= \prod_{i,j} p(y_{ijn}) \\ &= \prod_{i,j} \frac{1}{\pi {r_{ijn}}^2} \exp\left(-\frac{|y_{ijn}|^2}{{r_{ijn}}^2}\right) \end{aligned} \tag{6}$$

$${r_{ijn}}^2 = \sum_k t_{ikn} v_{kjn} \tag{7}$$

ここで $p(y_{ijn})$ は分散 ${r_{ijn}}^2$ の原点対称な複素ガウス分布であり i, j, n に関して独立である。また $k = 1, \ldots, K$ は基底のインデックス，t_{ikn}，v_{kjn} はそれぞれ NMF の基底とアクティベーションを表す。

[4] では生成モデルを以下の時変複素 Student's t 分布に拡張している（本稿ではこれを t-ILRMA と呼ぶ）。

$$\begin{aligned} p(\boldsymbol{Y}_n) &= \prod_{i,j} p(y_{ijn}) \\ &= \prod_{i,j} \frac{1}{\pi {r_{ijn}}^2} \left(1 + \frac{2}{\nu} \frac{|y_{ijn}|^2}{{r_{ijn}}^2}\right)^{-\frac{2+\nu}{2}} \end{aligned} \tag{8}$$

$${r_{ijn}}^p = \sum_k t_{ikn} v_{kjn} \tag{9}$$

ここで ν は自由度パラメータであり p はドメインパラメータである。$\nu \to \infty$，$p = 2$ のとき式 (8) 及び式 (9) はそれぞれ式 (6) 及び式 (7) に一致する。また，$\nu = 1$ とすると式 (8) は時変複素コーシー分布に一致する。空間モデル $\boldsymbol{W}_i$ 及び音源モデル t_{ikn}，v_{kjn} は，式（5）及び式（6）または式（8）に基づき観測信号の尤度最大化により求めることができる。

2.3 IDLMA

IDLMA は多チャネル音源モデル教師あり音源分離手法の 1 つであり，各音源信号の時間周波数構造を DNN により推論する。[9] では ILRMA 同様に音源スペクトログラム $\boldsymbol{Y}_n$ の生成モデルとして時変複素ガウス分布を仮定している（以降この手法を Gauss-IDLMA と呼ぶ）。その場合，式 (5) 及び式 (6) から観測信号 $\boldsymbol{x}_{ij}$ の負対数尤度は以下で与えられる。

$$\mathcal{L} \overset{\mathrm{c}}{=} -2J \sum_i \log|\det \boldsymbol{W}_i| + \sum_{i,j,n} \left(\frac{|y_{ijn}|^2}{{r_{ijn}}^2} + \log {r_{ijn}}^2\right) \tag{10}$$

ILRMA では各音源の時間周波数構造 r_{ijn} を式 (7) のように NMF で表現するが，Gauss-IDLMA では r_{ijn} は DNN により推論される。したがって，DNN の持つ柔軟な表現力により，Gauss-IDLMA は ILRMA ではよく表現できない非低ランクな時間周波数構造を持つ音源信号も適切に表現することが期待される。

Gauss-IDLMA の処理の流れを以下に述べる。Gauss-IDLMA では，まず DNN により推定された音源モデル r_{ijn} を用いて空間モデル $\boldsymbol{W}_i$ を最尤推定する。そして式 (5) により推定された分離信号を DNN に再入力することでより分離された音源モデル r_{ijn} を推定する。これらを繰り返すことで空間モデル $\boldsymbol{W}_i$ 及び音源モデル r_{ijn} を交互に更新し，観測信号の負対数尤度 (10) を最小化しながら分離信号を推定する。

空間モデル $\boldsymbol{W}_i$ は式 (10) を最小化する反復射影法（iter-

ative projection: IP）[11] により以下のように推定される。

$$\boldsymbol{U}_{in} = \frac{1}{J}\sum_{j}\frac{1}{{r_{ijn}}^2}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} \tag{11}$$

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i\boldsymbol{U}_{in})^{-1}\boldsymbol{e}_n \tag{12}$$

$$\boldsymbol{w}_{in} \leftarrow \frac{\boldsymbol{w}_{in}}{\sqrt{\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{U}_{in}\boldsymbol{w}_{in}}} \tag{13}$$

ここで $\boldsymbol{e}_n \in \mathbb{R}^{N\times 1}$ は n 番目の要素で値が 1，ほかの要素で値が 0 のベクトルである。

式 (5) により得た分離信号 y_{ijn} は各周波数毎のスケールが統一されていないため DNN の学習データとは大きく異なるものである。そのため，このままでは DNN による推論が適切に行われない。そこで従来の独立成分分析と同様に，次式のプロジェクションバックにより分離信号の各周波数毎のスケールをリファレンスチャネルのものに統一する。

$$y_{ijn} = [\boldsymbol{W}_i^{-1}(\boldsymbol{e}_n \circ \boldsymbol{y}_{ij})]_{m_{\mathrm{ref}}} \tag{14}$$

ここで $\circ$ はアダマール積であり，$[\cdot]_n$ はベクトルの n 番要素，m_{ref} はリファレンスチャネルのインデックスを表す。

以下では表記簡略化のため、ベクトルや行列に対する絶対値記号とドット付き指数は、要素毎の絶対値と指数乗をとった行列を表すものとする。ある音源 n の音源モデルを推論する DNN を DNN_n とする。分離信号 $|\boldsymbol{Y}_n|^{\cdot 1}$ を入力した DNN_n の出力信号を $\mathrm{DNN}_n(|\boldsymbol{Y}_n|^{\cdot 1}) \in \mathbb{R}_{\geq 0}^{I\times J}$ と表すと，DNN_n を用いた音源モデル r_{ijn} の更新は次式となる。

$$\boldsymbol{R}_n \leftarrow \mathrm{DNN}_n(|\boldsymbol{Y}_n|^{\cdot 1}) \tag{15}$$

$$r_{ijn} \leftarrow \max(r_{ijn}, \epsilon) \tag{16}$$

ここで，$\boldsymbol{R}_n \in \mathbb{R}^{I\times J}$ は各要素が r_{ijn} の行列である。ϵ は IP の安定性を向上させるための微小値である。

DNN_n は分離対象音源の学習信号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 $|\tilde{\boldsymbol{S}}_n|^{\cdot 1}$ を予測するよう学習される。$|\tilde{\boldsymbol{X}}|^{\cdot 1}$ を入力した DNN_n の出力信号を $\boldsymbol{D}_n = \mathrm{DNN}_n(|\tilde{\boldsymbol{X}}|^{\cdot 1})$ として，DNN_n を学習するための損失関数には次式の板倉斎藤擬距離を用いる [8]。

$$L = \frac{1}{IJ}\sum_{i,j}\left[\frac{|\tilde{s}_{ijn}|^2+\delta_1}{{d_{ijn}}^2+\delta_1} - \log\frac{|\tilde{s}_{ijn}|^2+\delta_1}{{d_{ijn}}^2+\delta_1} - 1\right] \tag{17}$$

ここで，$\tilde{s}_{ijn}$ 及び d_{ijn} はそれぞれ $\tilde{\boldsymbol{S}}_n$ 及び $\boldsymbol{D}_n$ の要素である。また δ_1 は零除算を防ぐための微小値である。

3 提案手法

3.1 動機

ILRMA においては，生成モデルを時変複素ガウス分布から時変複素 Student's t 分布へ拡張することで音源分離性能の改善が示されている [4]。本報告は生成モデルをヘビーテイル分布にすることで音源モデルの低ランク性が強調されることを示しているが，IDLMA において生成モデルをヘビーテイル分布にすることが与える影響は未だ示されていない。本稿では IDLMA において生成モデルを時変複素 Student's t 分布へと一般化する（以後、本提案法を t-IDLMA と呼ぶ）。なお，複素 Student's t 分布は再生性のある複素ガウス分布や複素コーシー分布およびその中間に相当する分布を含んでおり，統計的に有用な分布である。従来法の Duong+DNN 法においても時変複素ガウス分布や時変複素コーシー分布の尤度を最大化する DNN を用いているが，空間モデル（空間相関行列）には時変複素ガウス分布のみを仮定している。したがって，従来法では音源モデルと空間モデルとの間に不一致が生じている。一方，提案法では空間モデルを時変複素 Student's t 分布でモデル化し，その最適化手法を最尤推定により導出することで，空間モデルと音源モデル間の不一致を回避し尤度を一貫して最大化することができる。

3.2 空間モデル及び音源モデルの更新則

提案手法の概略を図 1 に示す。提案手法は従来の Gauss-IDLMA 同様に，空間モデル $\boldsymbol{W}_i$ をブラインドに推定し音源信号を DNN により推論するが，音源の生成モデルとして複素ガウス分布よりもヘビーテイルな分布を仮定することができる。

式 (5) 及び式 (8) から時変複素 Student's t 分布に基づく観測信号の負対数尤度は以下の式で与えられる。

$$\mathcal{L}_t \overset{\mathrm{c}}{=} -2J\sum_i \log|\det\boldsymbol{W}_i| + \sum_{i,j,n}\left[\left(1+\frac{\nu}{2}\right)\log\left(1+\frac{2}{\nu}\frac{|y_{ijn}|^2}{{r_{ijn}}^2}\right) + \log {r_{ijn}}^2\right] \tag{18}$$

空間モデル $\boldsymbol{W}_i$ は式 (18) の第一項及び第二項の最小化に基づき音源間の統計的独立性と音源モデル $\boldsymbol{R}_n$ を考慮して推定される。従来の Gauss-IDLMA では，$\boldsymbol{W}_i$ は IP を適用することで更新することが出来る。IP は $|\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2$ 項と $-\log|\det\boldsymbol{W}_i|$ 項の和の最小化に適用可能な高速アルゴリズムである。しかし，t-IDLMA では式 (18) に log 関数中の $|\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{x}_{ij}|^2$ を含むため IP を適用することが出来ない。したがって，本稿では初めに補助関数法を式 (18) に適用し，その補助関数に対して IP を適用することで $\boldsymbol{W}_i$ の更新則を導く。

式 (18) の補助関数を導くため，接線不等式

$$\log z \leq \frac{1}{\alpha}(z-\alpha) + \log\alpha \tag{19}$$

を式 (18) の log 項に適用する。ここで $z \geq 0$ は元の関数の変数，$\alpha \geq 0$ は補助変数である。式 (19) により式 (18) の

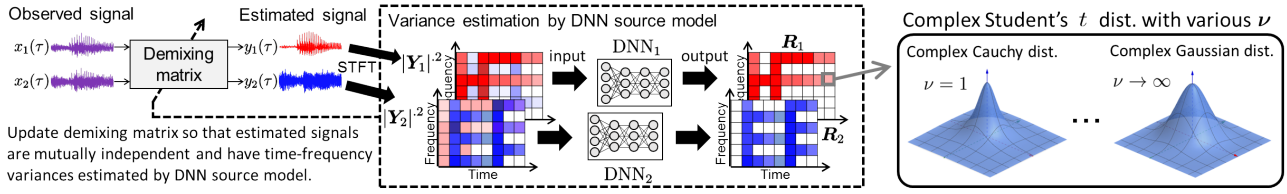

図 1: Principle of source separation based on t-IDLMA in case of $N = M = 2$.

補助関数として以下を得る。

$$\begin{aligned}\mathcal{L}_t &\leq \sum_{i,j,n}\left[\left(1+\frac{\nu}{2}\right)\frac{1}{\alpha_{ijn}}\left(1+\frac{2}{\nu}\frac{|y_{ijn}|^2}{{r_{ijn}}^2}-\alpha_{ijn}\right)\right.\\ &\qquad\left.+\left(1+\frac{\nu}{2}\right)\log\alpha_{ijn}+\log {r_{ijn}}^2\right]\\ &\quad -2J\sum_i \log|\det \boldsymbol{W}_i|\\ &=: \mathcal{L}_t^+ \end{aligned} \tag{20}$$

ここで α_{ijn} は補助変数であり，$\mathcal{L}_t$ と $\mathcal{L}_t^+$ は

$$\alpha = 1+\frac{2}{\nu}\frac{|y_{ijn}|^2}{{r_{ijn}}^2} \tag{21}$$

の時のみ一致する。補助関数 (20) は以下のように変形される。

$$\mathcal{L}_t^+ = J\sum_{i,n}\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{U}_{in}\boldsymbol{w}_{in} - 2J\sum_i \log|\det \boldsymbol{W}_i| + \text{const.} \tag{22}$$

$$\boldsymbol{U}_{in} = \frac{1}{J}\left(1+\frac{2}{\nu}\right)\sum_j \frac{1}{\alpha_{ijn}{r_{ijn}}^2}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} \tag{23}$$

式 (22) に IP を適用し式 (21) を代入することで分離フィルタ $\boldsymbol{w}_{in}$ に関して以下の更新則を得る。

$$c_{ijn} = \frac{\nu}{\nu+2}{r_{ijn}}^2 + \frac{2}{\nu+2}|y_{ijn}|^2 \tag{24}$$

$$\boldsymbol{U}_{in} = \frac{1}{J}\sum_j \frac{1}{c_{ijn}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}} \tag{25}$$

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i\boldsymbol{U}_{in})^{-1}\boldsymbol{e}_n \tag{26}$$

$$\boldsymbol{w}_{in} \leftarrow \frac{\boldsymbol{w}_{in}}{\sqrt{\boldsymbol{w}_{in}^{\mathrm{H}}\boldsymbol{U}_{in}\boldsymbol{w}_{in}}} \tag{27}$$

DNN 音源モデルによる r_{ijn} の更新は従来の Gauss-IDLMA 同様に式 (15) 及び式 (16) により行うことができる。ただし，DNN の学習時には 3.3 節で述べるように式 (18) に基づいた損失関数を用いることで，式 (15) が時変複素 Student's t 分布に基づいた生成モデルの分散を出力するよう DNN を学習させる必要がある。

3.3 DNN 学習時の損失関数

t-IDLMA において DNN_n の損失関数には以下の式を用いる。

$$L_t = \sum_{i,j}\left[\left(1+\frac{\nu}{2}\right)\log\left(1+\frac{2}{\nu}\frac{|\tilde{s}_{ijn}|^2+\delta_1}{{d_{ijn}}^2+\delta_1}\right)\right.$$
$$\left.+\log({d_{ijn}}^2+\delta_1)\right] \tag{28}$$

式 (28) の最小化は式 (18) における分散 r_{ijn} の最尤推定と等価である。したがって，式 (28) を用いて学習を行った DNN_n は，時変複素 Student's t 分布に基づいて分離信号 y_{ijn} の生成モデルパラメータを最尤推定することが期待される。

3.4 自由度パラメータと空間モデル推定の安定性

従来の Gauss-IDLMA では，式 (11) により定義される $\boldsymbol{U}_{in}$ は ${r_{ijn}}^{-2}$ で重みづけされた空間相関行列 $\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}}$ と解釈できる。r_{ijn} はスペクトログラムの振幅に対応しているため，DNN_n による推定では ReLU など負値を零値に変換する活性化関数を出力層に組み込む必要がある。したがって，r_{ijn} は時間周波数領域で非常にスパースなものとなりやすく，その二乗逆数である ${r_{ijn}}^{-2}$ を用いる IP の計算は数値的に不安定なものとなりやすい。

ヘビーテイル分布に基づく IDLMA ではこの問題を緩和することができる。t-IDLMA において式 (25) 中の c_{ijn} は ${r_{ijn}}^2$ と $|y_{ijn}|^2$ を $\nu:2$ で重みづけした算術平均である。y_{ijn} は線形空間フィルタによって得られる分離信号であるため ${r_{ijn}}^2$ に比べスパース性は小さい。したがって，従来の複素ガウス分布（$\nu\to\infty$）から ν の値を小さくすることで，${r_{ijn}}^2$ を平滑化し数値的に安定化させる役割が期待される。

一方，ν の値を小さくすることは DNN による強力な推論結果を平滑化により弱めることに対応するため，収束が遅くなってしまう。収束が遅すぎると，空間モデルは分離途中の分離信号に対する局所最適解に陥ってしまう。以上より ν の大きさにはトレードオフがある。最適な ν の選択は次章で議論する。

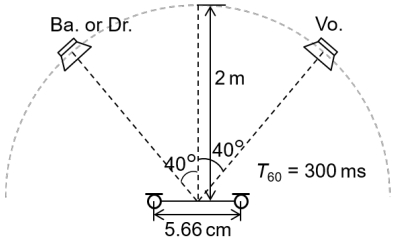

図 2: Recording environment of impulse response.

4 評価実験

4.1 実験条件

音楽信号の 2 音源分離を通して提案する t-IDLMA の性能評価を行った。なお最適な ν の値を調べるため $\nu = 1, 10, 100, 1000$ の 4 通りに関して実験を行った。比較対照法として，BSS 手法の t-ILRMA（$K = 20$），教師あり手法の Duong+DNN 法，DNN+WF 法（単一チャネル DNN に Wiener フィルタを接続したもの）及び従来の Gauss-IDLMA [9] に関して実験を行った。MNMF に関しては，その分離性能が ILRMA に劣ることが [10] で示されているため，比較対照法には加えなかった。Duong+DNN 法，Gauss-IDLMA，t-IDLMA では，空間モデルを IP で 10 回更新するごとに DNN により音源モデルを 1 回更新した。

実験には SiSEC2016 [12] 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.）をドライソース及び DNN の学習データとして用いた。dev データ 50 曲を DNN の学習データとして用い，test データのアルファベット順の上位 25 曲を性能評価に用いた。test データは 30 秒から 60 秒の区間を音源分離の対象とした。但し，該当区間の混合信号のパワー平均値が一定以下となる楽曲については，60 秒から 90 秒の区間を対象とした。観測信号に残響を与えるため，図 2 に示す RWCP データベース [13] から得られる E2A インパルス応答を Ba. と Vo.（Ba./Vo.）及び Dr. と Vo.（Dr./Vo.）に畳み込んで分離対象となる 2 チャネル信号を作成した。全ての信号は 8 kHz にダウンサンプリングを行った。Ba./Vo. の分離ではシフト長 256 ms，窓長 512 ms のハミング窓による STFT を行い，Dr./Vo. の分離ではシフト長 128 ms，窓長 256 ms のハミング窓による STFT を行った。各手法において 1 反復は IP の更新 1 回に相当し，全手法に対して 150 回の反復を行った。音源分離性能の客観評価尺度には source-to-distortion ratio（SDR）[14] を用いた。

4.2 DNN 音源モデルの構造と訓練手法

音源モデル推定 DNN は 4 つの隠れ層をもつ全結合型ニューラルネットワークとした。ユニット数はすべての隠れ層で 1024 とし，各ユニットの活性化関数は ReLU を用いた。

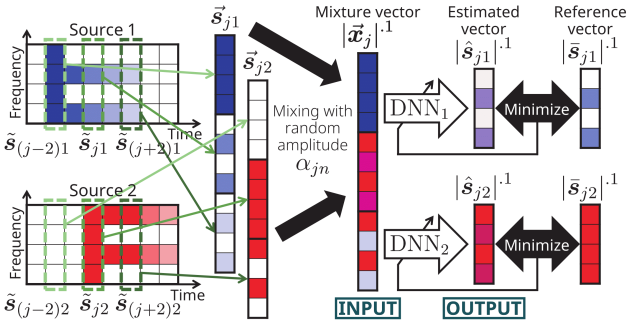

図 3: Training of DNN using feature vectors.

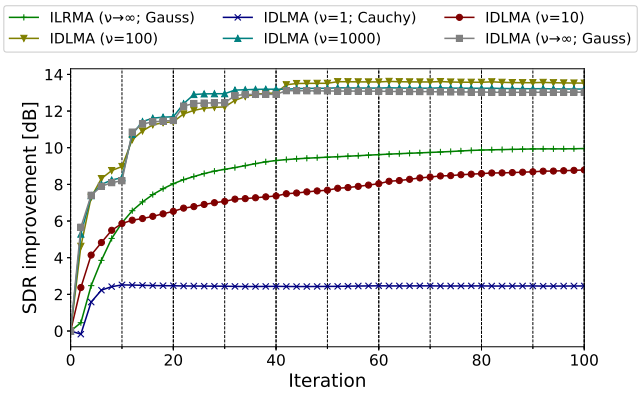

図 4: SDR improvements for each method for Ba./Vo.

訓練手法の概略を図 3 に示す。DNN の入出力には以下に示す特徴量を用いた。

$$\vec{\boldsymbol{x}}_j = \frac{\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}}{\|\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}\|_2 + \delta_2} \in \mathbb{C}^{I(2c+1)\times 1} \tag{29}$$

$$\bar{\boldsymbol{s}}_{jn} = \frac{\alpha_{jn}\tilde{\boldsymbol{s}}_{jn}}{\|\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}\|_2 + \delta_2} \in \mathbb{C}^{I\times 1} \tag{30}$$

$$\vec{\boldsymbol{s}}_{jn} = [\tilde{\boldsymbol{s}}^{\mathrm{T}}_{(j-2c)n}, \tilde{\boldsymbol{s}}^{\mathrm{T}}_{(j-2c+2)n}, \cdots, \tilde{\boldsymbol{s}}^{\mathrm{T}}_{(j+2c)n}]^{\mathrm{T}} \in \mathbb{C}^{I(2c+1)\times 1} \tag{31}$$

ここで α_{jn} は時間フレーム j における音源 n 毎の混合比である [15]。DNN の入力と出力はそれぞれ $|\vec{\boldsymbol{x}}_j|^{\cdot 1}$ 及び $|\bar{\boldsymbol{s}}_{jn}|^{\cdot 1}$ である。入力 $\vec{\boldsymbol{x}}_j$ は時間フレーム j を中心にして 1 フレームおきに前後 c フレームの周波数ベクトルを結合した混合音ベクトルであり，出力 $\bar{\boldsymbol{s}}_{jn}$ は時間フレーム j における教師音の周波数ベクトルである。また，δ_2 は零除算を防ぐための微小値である。DNN の学習時はバッチサイズを 128，エポック数を 2000 として ADADELTA [16] による最適化を行った。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。その他のパラメータは，$\delta_1 = \delta_2 = 10^{-5}$，$\epsilon = 10^{-1} \times (IJ)^{-1} \sum_{i,j} r_{ijn}$，$c = 3$，及び $\lambda = 10^{-5}$ とした。

4.3 評価結果

図 4 に ILRMA，Gauss-IDLMA 及び t-IDLMA の収束の様子を示す。3.4 節に記したように，より大きな ν の値では 2 反復目のような分離開始時の更新幅は大きいが収束先の分離性能は最大となっておらず，収束先の分離性能

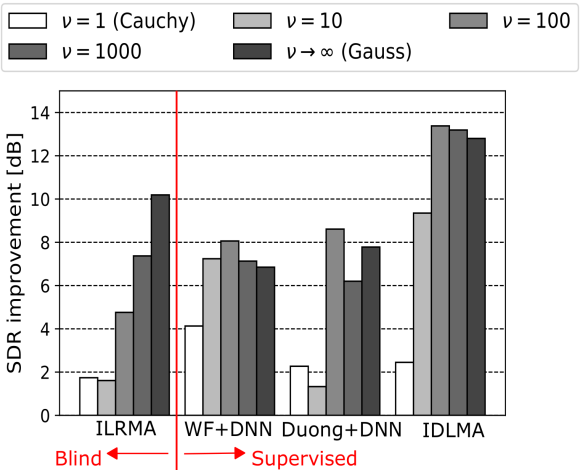

図 5: Average SDR improvement of Ba./Vo. 25 songs.

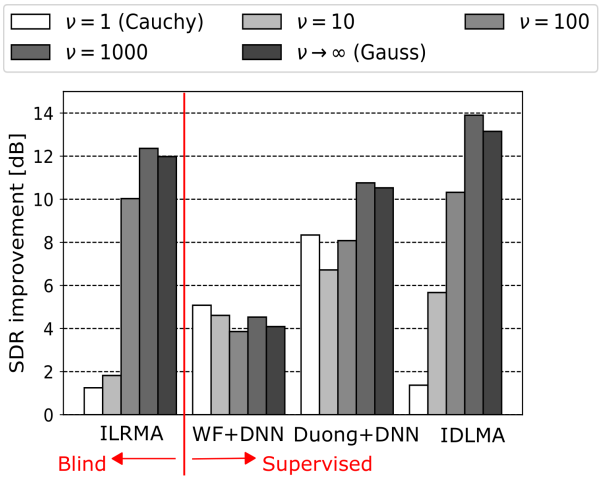

図 6: Average SDR improvement of Dr./Vo. 25 songs.

は適切な大きさの ν (= 100) で最大となることが確認できる。図 5 及び図 6 にそれぞれ Ba./Vo. 及び Dr./Vo に関する 25 曲平均の SDR 改善量を示す。いずれの音源の組み合わせに関しても，提案手法である t-IDLMA が最大の分離性能を示している。そして，Ba./Vo. の組み合わせでは ν = 100 で最大値をとるのに対して Dr./Vo. の組み合わせでは ν = 1000 で最大値をとることが確認できる。これは 3.4 節に記した ν の大きさによる更新スピードと平滑化のトレードオフや音源毎の最適な生成モデルの違いが影響したためであると考えられる。

5 結論

本稿では時変複素 Student's t 分布を音源生成モデルとした IDLMA として t-IDLMA を提案した。また評価実験により従来の Gauss-IDLMA に比べて分離性能が改善されることを示した。

謝辞 本研究の一部はセコム科学技術振興財団，JSPS 科研費 17H06101，及び 17H06572 の助成を受けたものである。

参考文献

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

[2] 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. 14, no. 9, pp. 1626 – 1641, 2016.

[3] 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.

[4] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 28, pp. 1–25, 2018.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 角野隼斗, 北村大地, 高宗典玄, 高道慎之介, and 猿渡洋, "独立深層学習行列分析に基づく多チャネル音源分離," in 日本音響学会 *2018* 年春季研究発表会講演論文集, no. 1-4-16, 2018.

[10] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for multichannel audio source separation," in *Proc. EUSIPCO*, 2018, pp. 1571–1575.

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

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

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

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

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

[16] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *CoRR*, vol. abs/1212.5701, 2012.