混合複素ガウスモデルに基づく深層ベイズ音源分離

坂東 宜昭[1,a)]　佐々木 洋子[1]　吉井 和佳[2]

概要：本稿では，多チャネル混合音のみを用いて，深層ニューラルネットワーク (DNN) に基づく音源分離を教師なし学習する枠組みについて述べる．従来の深層音源分離は，高い性能を得るために大量の教師データを必要とする． 一方，空間情報に基づく多チャネル音源分離は，学習データが不要だがパラメータの初期値依存性や方向の近い音源で性能が劣化する．提案法は，各音源の時間周波数 (TF) マスクと到来方向 (DoA) を潜在変数にもつ混合複素ガウスモデル (cGMM) をコスト関数として，TF マスクを推定する分離 DNN と DoA を推定する定位 DNN を学習する．DoA を同時推論することで，空間モデルのパーミュテーション問題を統一的な枠組みで解決できる．さらに学習済みの分離 DNN は，単チャネル分離として動作するだけでなく，cGMM の多チャネル分離アルゴリズムに良い初期値を与えることができる．シミュレーション混合音を用いた評価により，従来の初期化法より信号対歪比が改善することを確認した．

1. はじめに

音源分離は，個別の音イベントを認識する音環境認識にとって不可欠な機能の一つである [1–3]．音声分離や音楽分離といった特定のタスクにおいて，深層ニューラルネットワーク (DNN) が，圧倒的な性能を達成している [4–7]．例えば，パーミュテーション不変学習 (PIT) [5] は，各音源信号の時間周波数 (TF) マスクを出力する DNN を学習する．このような学習の枠組みは，音源信号と混合音のペアからなる教師データを大量に必要とする．日常生活下の音イベントといった音源種類が無数に存在する条件では，事実上そのような教師データの収集が困難なため，教師データの不要な分離法が望ましい．

多チャネル録音信号に含まれる空間情報を用いることで，教師データを用いずに音源を分離できる [8–11]．例えば，マイクロホン間の位相差とパワー差から，各音源の TF マスクを推定する手法が広く研究されている [10,12,13]．複素混合ガウスモデル (cGMM) [10,14,15] は，そのような空間情報を空間相関行列として表現し，各 TF ビンをクラスタリングすることで TF マスクを推定する．cGMM は，各周波数ごとに独立して定式化されるため，周波数ごとに音源インデックスが異なるパーミュテーション問題が存在する．この問題は，各音源の到来方向 (DoA) を用いて解決することができ，DoA と TF マスクを一挙に推論する手法が提案されている [10,15]．また，方向情報を用いることで，混合音に含まれる音源数も同時推定できる [10]．しかし，一般に多チャネル音源分離は，初期値により性能が大きく左右されるうえ，到来方向の近い音源は性能が劣化する．

多チャネル混合音を用いて深層音源分離を教師なし学習する枠組みが近年注目を集めている [16,17]．例えば，学習データに多チャネル分離法を適用し，分離音を教師データとする枠組みが提案されている [16–18]．この枠組みは，多チャネル分離法の性能劣化が，直接ネットワークの学習にも悪影響を与えうる．この問題を解決するため，Drude ら [19] は，cGMM に基づく空間モデルの尤度関数を用いて深層音源分離を直接学習した．彼らは，従来の多チャネル分離法を学習済み分離 DNN で初期化することで，高い分離性能が得られることを示した．この学習法では，パーミュテーション問題を TF マスクの周波数間の相関を用いて揃える処理で解決するため，混合音に含まれる音源数を事前に与える必要がある．そのため，音源数が一般に未知である日常生活下の音源分離への適用は難しい．

本稿では，多チャネル混合音のみの学習データから，深層音源分離を教師なし学習する．本手法は，TF マスクを推定する分離 DNN と，DoA を推定する定位 DNN を同時学習することで，パーミュテーション問題を解決する．パーミュテーションの解決を単一の確率モデルの推論として扱えるうえ，統一的な音源数の推定も期待できる．TF マスクと DoA を潜在変数にもつ cGMM [10] に基づき多チャネル混合音の生成モデルを定式化し，その変分下限を目的関数として DNN を学習する．学習済みの分離 DNN は，単体でも単チャネル音源分離として動作するうえ，多チャネル分離アルゴリズムによい初期値を与えることができる．

[1] 産業技術総合研究所 人工知能研究センター
[2] 理化学研究所 AIP / 京都大学 情報学研究科
[a)] y.bando@aist.go.jp

2. 深層ベイズ音源分離

多チャネル混合音からなる学習データに対し，cGMM に基づく空間モデルを目的関数として，深層音源分離を教師なし学習する．本稿では，Otsuka ら [10] が提案した各音源の TF マスクと DoA を潜在変数にもつ潜在ディリクレ配分 (LDA) モデルと呼ばれる cGMM に基づき目的関数を導出する．目的関数は対数周辺尤度の下限を用いて導出され，その最大化は DNN の出力とモデルの事後分布間の Kullback-Leibler (KL) 擬距離の最小化に対応する．

2.1 多チャネル混合音の確率的生成モデル

本稿で用いる混合音生成モデルでは，K 個の音源 $s_{tfk} \in \mathbb{C}$ を観測した M チャネル混合音 $\mathbf{x}_{tf} \in \mathbb{C}^M$ を，TF 領域における瞬時混合を仮定し以下のように表現する:

$$\mathbf{x}_{tf} = \sum_k \mathbf{a}_{fk} s_{tfk}. \tag{1}$$

ただし，$\mathbf{a}_{fk} \in \mathbb{C}^M$ は，音源 k のステアリングベクトルである．また，各音源信号 $s_{tfk} \in \mathbb{C}$ は，平均 0 の複素ガウス分布 ($\mathcal{N}_{\mathbb{C}}$) に従うとする:

$$s_{tfk} \mid \lambda_{tfk} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{tfk}\right). \tag{2}$$

ここで $\lambda_{tfk} \in \mathbb{R}_+$ は音源 k のパワースペクトル密度を表す．cGMM ではさらに，各 TF ビンは高々 1 つの音源信号のみ含むとするスパース仮定を導入し，TF マスク $z_{tfk} \in \{0,1\}$ ($\sum_k z_{tfk} = 1$) を用いて式 (1) を以下に置き換える:

$$\mathbf{x}_{tf} = \sum_k z_{tfk} \left(\mathbf{a}_{fk} s_{tfk}\right). \tag{3}$$

TF マスク z_{tfk} の各クラス k の出現頻度は，時間フレームごとに大きく変化すると考えられるので，以下のカテゴリカル分布から生成されるとする:

$$[z_{tf1}, \ldots, z_{tfK}]^{\mathsf{T}} \mid \boldsymbol{\pi}_t \sim \mathrm{Cat}(\pi_{t1}, \ldots, \pi_{tK}) \tag{4}$$

ここで，$\pi_{tk} \in \mathbb{R}_+$ ($\sum_k \pi_{tk} = 1$) は，各時間フレームでの音源 k の出現頻度を表すモデルパラメータである．

式 (3) の空間モデルは，周波数ビンごとに独立しているため，パーミュテーション問題が生じる．そこで LDA モデル [10] では，各音源の DoA と TF マスクを同時推論することでパーミュテーション問題に対処する．この同時推論は，潜在的な DoA $d \in \{1, \ldots, D\}$ を考え，方向 d のステアリングベクトル $\mathbf{a}_{fd} \in \mathbb{C}^M$ を用いて式 (3) を置き換えた以下の空間モデルを用いて行う：

$$\mathbf{x}_{tf} = \sum_{k,d} z_{tfk} w_{kd} \left(\mathbf{a}_{fd} s_{tfk}\right). \tag{5}$$

ここで，$w_{kd} \in \{0,1\}$ ($\sum_d w_{kd} = 1$) は，各音源の方向の割り当て変数で，以下のカテゴリカル分布に従い生成する:

$$[w_{k1}, \ldots, w_{kD}]^{\mathsf{T}} \mid \boldsymbol{\phi} \sim \mathrm{Cat}(\phi_1, \ldots, \phi_D). \tag{6}$$

ただし，$\phi_d \in \mathbb{R}_+$ ($\sum_d \phi_d = 1$) は DoA が d となる頻度を表すパラメータである．本稿では，潜在的な DoA として，水平方向に 5° 間隔で分割した $D = 72$ 方向を考える．

観測混合音 $\mathbf{x}_{tf}$ の尤度関数は，式 (2) および (5) から，多変量混合複素ガウス分布として表現できる:

$$\mathbf{x}_{tf} \,|\, \lambda_{tfk}, \mathbf{H}_{fd}, \mathbf{z}_{tf}, \mathbf{w}_k \sim \prod_{k,d} \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \lambda_{tfk} \mathbf{H}_{fd}\right)^{z_{tfk} w_{kd}} \tag{7}$$

ただし，$\mathbf{H}_{fd} = \mathbb{E}[\mathbf{a}_{fd} \mathbf{a}_{fd}^{\mathsf{H}}]$ は，方向 d の空間相関行列を表す．本稿では，$\mathbf{H}_{fd}$ を方向 d に制約しながら推定するために，以下の複素逆ウィシャート分布を仮定する:

$$\mathbf{H}_{fd} \mid \nu, \mathbf{G}_{fd} \sim \mathcal{IW}_{\mathbb{C}}\left(\nu, (\nu - M)\mathbf{G}_{fd}\right) \tag{8}$$

ここで，$\nu \in \mathbb{R}_+$ と $\mathbf{G}_{fd} = \mathbf{b}_{fd} \mathbf{b}_{fd}^{\mathsf{H}} + \epsilon \mathbf{I}$ は，ハイパーパラメータである．ただし，$\mathbf{b}_{fd} \in \mathbb{C}^M$ は事前に準備する方向 d のステアリングベクトルで，ϵ は $\mathbf{G}_{fd}$ を正定値行列にするために加算する．本稿では，$\epsilon = 1.0 \times 10^{-2}$ とし，平面波を仮定した幾何計算により $\mathbf{b}_{fd}$ を得る．

2.2 変分推論

本稿の教師なし学習と多チャネル音源分離は，TF マスク z_{tfk} と DoA w_{kd} の事後分布 $p(\mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{H}, \boldsymbol{\Theta})$ を推定する枠組みである．ただし，$\boldsymbol{\Theta} = \{\boldsymbol{\lambda}, \boldsymbol{\pi}, \boldsymbol{\phi}\}$ はモデルパラメータを表す． TF マスク z_{tfk} と DoA w_{kd} の事後分布は解析的に計算困難なので，以下の独立性を仮定した変分事後分布を用いて，事後分布 $p(\mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{H}, \boldsymbol{\Theta})$ を近似推論する:

$$p(\mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{H}, \boldsymbol{\Theta}) \approx q(\mathbf{Z}) q(\mathbf{W}). \tag{9}$$

この近似推論は，以下で定義される周辺尤度 $p(\mathbf{X} | \mathbf{H}, \boldsymbol{\Theta})$ の変分下限 $\mathcal{L}_q$ を最大化することで行われる:

$$\begin{aligned}\mathcal{L}_q = {} & \mathbb{E}_q\left[\log p(\mathbf{X} \mid \boldsymbol{\lambda}, \mathbf{H}, \mathbf{Z}, \mathbf{W})\right] \\ & - \mathbb{KL}\left[q(\mathbf{Z}) | p(\mathbf{Z} | \boldsymbol{\pi})\right] - \mathbb{KL}\left[q(\mathbf{W}) | p(\mathbf{W} | \boldsymbol{\phi})\right].\end{aligned} \tag{10}$$

変分下限の最大化は，モデルの真の事後分布と変分事後分布の間の KL 擬距離の最小化に対応する．この下限 $\mathcal{L}_q$ は，具体的には以下で計算できる:

$$\begin{aligned}\mathcal{L}_q = {} & -\sum_{t,f,k,d} \hat{z}_{tfk} \hat{w}_{kd} \left(\frac{1}{\lambda_{tfk}} \mathbf{x}_{tf}^{\mathsf{H}} \mathbf{H}_{fd}^{-1} \mathbf{x}_{tf} + \log \left|\lambda_{tfk} \mathbf{H}_{fd}\right|\right) \\ & + \sum_{t,f,k} \hat{z}_{tfk} \log \frac{\pi_{tfk}}{\hat{z}_{tfk}} + \sum_{k,d} \hat{w}_{kd} \log \frac{\phi_d}{\hat{w}_{kd}} + \mathrm{const.}\end{aligned} \tag{11}$$

ただし，$\hat{z}_{tfk}$ は $q(z_{tfk} = 1)$ で，$\hat{w}_{kd}$ は $q(w_{kd} = 1)$ である．

モデルパラメータ $\boldsymbol{\Theta}$ は最尤推定で求め，空間相関行列 $\mathbf{H}$ は事後確率最大推定で求める．これらも直接計算困難なので，変分下限 (式 (10)) を用いて以下のように更新する:

$$\mathbf{H}_{fd} \leftarrow \frac{\mathbf{G}_{fd} + \sum_{t,k} \hat{z}_{tfk} \hat{w}_{kd} \frac{1}{\lambda_{tfk}} \mathbf{x}_{tf} \mathbf{x}_{tf}^{\mathsf{H}}}{\nu + \sum_{t,k} \hat{z}_{tfk} \hat{w}_{kd} + M}, \tag{12}$$

$$\lambda_{tfk} \leftarrow \frac{1}{M} \sum_d \hat{w}_{kd} \mathbf{x}_{tf}^{\mathsf{H}} \mathbf{H}_{fd}^{-1} \mathbf{x}_{tf}, \tag{13}$$

$$\pi_{tk} \leftarrow \frac{1}{F} \sum_f \hat{z}_{tfk}, \qquad \phi_d \leftarrow \frac{1}{K} \sum_k \hat{w}_{kd}. \tag{14}$$

変分事後分布とパラメータは，収束するまで交互更新する．

2.3 償却変分推論による事前学習

事前学習では，N 個の M チャネル混合音 $\mathbf{x}_{tf}^{(n)}$ を用いて，それぞれ TF マスク z_{tfk} と DoA w_{kd} を推定する 2 つの DNN を教師なし学習する．ただし，以降では個々の $\mathbf{x}_{tf}^{(n)}$ について議論するので，添字 $^{(n)}$ を省略する．TF マスクを推定する分離 DNN (以下，g_{tfk}) は，単チャネルスペクトログラムを入力とし，音源 k が時間 t 周波数 f で選択される確率 $q(z_{tfk}=1)$ を出力するように学習する:

$$q_g(z_{tfk}=1)=\hat{z}_{tfk}=g_{tfk}(\log|\mathbf{X}|) \tag{15}$$

ここで，$\log|\mathbf{X}|\in\mathbb{R}^{T\times F}$ は単チャネルの対数振幅スペクトログラムとし，本稿では $m=1$ 番目のマイクロホンの観測とする．一方，DoA を推定する定位 DNN(以下，h_{kd}) は，音源 k の到来方向が d となる確率 $q_h(w_{kd}=1)$ を予測する:

$$q_h(w_{kd}=1)=\hat{w}_{kd}=h_{kd}(\boldsymbol{\omega}) \tag{16}$$

ここで，$\boldsymbol{\omega}=\{\omega_{kd}\}_{k,d}\in\mathbb{R}^{K\times D}$ は，定位 DNN の入力となる空間情報を表す特徴量である．複素数を扱う DNN の学習は難しいので，以下の混合ガウス尤度を特徴量とする:

$$\omega_{kd}=\sum_{t,f}\hat{z}_{tfk}\log\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\mathbf{G}_{fd}\right). \tag{17}$$

本稿の事前学習では，学習データの各混合音に対する対数周辺尤度 $\log p(\mathbf{X}|\mathbf{H},\boldsymbol{\Theta})$ の変分下限 $\mathcal{L}_q$ (式 (10)) を最大化することで，DNN g_{tfk} と h_{kd} を学習する．ただし，$\boldsymbol{\lambda}$ と $\mathbf{H}_{fd}$ は簡単化のためそれぞれ $\frac{1}{TFM}\sum_{t,f}\mathbf{x}_{tf}^{\mathsf{H}}\mathbf{x}_{tf}$ および $\mathbf{G}_{fd}$ に固定する．分離 DNN g_{tfk} と定位 DNN h_{kd} は，具体的には以下の 3 ステップを繰り返すことで学習する:

1) DNN g_{tfk} と h_{kd} を用いて，ミニバッチの各混合音 $\mathbf{x}_{tf}$ に対する TF マスク $\hat{z}_{tfk}$ と DoA $\hat{w}_{kd}$ を推定する
2) モデルパラメータ π_{tk} および ϕ_d を更新する (式 (14))
3) 変分下限 $\mathcal{L}_q$ を計算し，確率的勾配法 (SGD) を用いて DNNg_{tfk} と h_{kd} のパラメータを更新する

このような学習は，事前の学習データを用いて確率モデルの事後分布を予測する DNN を学習するため，償却変分推論 (AVI) と呼ばれる [20–22].

2.4 EM アルゴリズムよる多チャネル音源分離

事前学習した分離 DNN g_{tfk} は，単体でも単チャネル音源分離ができるが，推定マスクを初期値として多チャネル expectation-maximization (EM) アルゴリズムに与えれば，より高い性能が得られる．本稿で述べた cGMM モデルの EM アルゴリズムを説明する．まず，マスク変数 $\hat{z}_{tfk}$ および DoA $\hat{w}_{kd}$ を変分下限 $\mathcal{L}_q$ を最大化するよう更新する:

$$\hat{z}_{tfk}\leftarrow\frac{\pi_{tk}\prod_d\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\lambda_{tfk}\mathbf{H}_{fd}\right)^{\hat{w}_{kd}}}{\sum_k\pi_{tk}\prod_d\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\lambda_{tfk}\mathbf{H}_{fd}\right)^{\hat{w}_{kd}}} \tag{18}$$

$$\hat{w}_{kd}\leftarrow\frac{\phi_d\prod_{t,f}\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\lambda_{tfk}\mathbf{H}_{fd}\right)^{\hat{z}_{tfk}}}{\sum_d\phi_d\prod_{t,f}\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{tf};\mathbf{0},\lambda_{tfk}\mathbf{H}_{fd}\right)^{\hat{z}_{tfk}}} \tag{19}$$

次に，式 (12)–(14) に従い，モデルパラメータ $\boldsymbol{\Theta}$ と空間相関行列 $\mathbf{H}$ を更新する．本アルゴリズムは，これらの変数を収束するまで交互更新するので，局所解に陥いらないよう慎重な初期化が不可欠である．

TF マスク $\hat{z}_{tfk}$ の初期値は，分離 DNN g_{tfk} を用いて初期化する．一方，DoA $\hat{w}_{kd}$ の初期化には，学習データの空間的偏りに過学習している可能性があるため定位 DNN h_{kd} は使用せず，以下を用いて初期化する:

$$\hat{w}_{kd}\propto\exp\left(-\sum_{t,f}\hat{z}_{tfk}\mathbf{x}_{tf}^{\mathsf{H}}\mathbf{G}_{fd}^{-1}\mathbf{x}_{tf}\right). \tag{20}$$

3. 評価実験

インパルス応答を数値シミュレーションして生成した混合音を用いて，評価実験を行った．

3.1 データセット

WSJ0-2mix データセット [4] に含まれる音源信号にインパルス応答を畳み込んで多チャネル混合音を生成した．WSJ0-2mix データセットは，単チャネル音源分離の評価で広く利用されている [4, 5, 23]．このデータセットの混合音は，WSJ0 コーパスに含まれる $K=2$ 発話をランダムに選び，信号対雑音比 (SNR) を -5 dB から $+5$ dB の範囲でランダムに変動させて生成されている．本稿では，鏡像法 [24] に基づく数値シュミレーション[*1]を用いてインパルス応答を生成した．シミュレーションする部屋の大きさは 5 m × 5 m × 3 m から，10 m × 10 m × 4 m まで各混合音ごとにランダムに変動させた．この部屋の中央に直径 8cm の $M=4$ チャネル円形マイクアレイを配置し，2 つの音源を室内のランダムな位置に配置した．残響時間 (RT_{60}) は，0.2 秒から 0.5 秒の範囲でランダムに変動させた．生成したデータセットは，学習セットとバリデーションセットを持ち，それぞれ 2 万個と 5 千個の混合音からなる．また，評価セットは 3 千個の混合音からなり，本セットは他のセットと話者が独立している．本稿では，計算量削減のため，これらの信号を 8 kHz サンプリングで生成した．

3.2 実験設定

分離 DNN g_{tfk} と定位 DNN h_{kd} は，以下で示すアーキテクチャにより構成した．まず，分離 DNN g_{tfk} は，時間方向に前向きと後向きの双方向を持つ長短期記憶型の再帰ニューラルネットワーク (BiLSTM) を 3 層もつ．各 LSTM は 600 のユニットから成り，BiLSTM 層に続き 1 層の全結合層を経て $\log\hat{z}_{tfk}$ を出力する．一方，定位 DNN h_{kd} は，方向ごとの偏りを学習しないように，方向 d を軸とする 1 元次畳込み層 (1D-conv) で構成した．本稿では，1D-conv 3 層の後に入力と残差接続し $\log\hat{w}_{kd}$ を出力する．

[*1] https://github.com/ty274/rir-generator

表 1 評価セットに対する音源分離性能 (教師あり学習は灰字)

手法	初期化	マイク数 学習	マイク数 評価	反復回数	SDR [dB]
EM-cGMM	g_{tfk}	4	4	100	9.1
EM-cGMM	g_{tfk}	4	4	50	**9.4**
EM-cGMM	g_{tfk}	4	4	10	9.1
EM-cGMM	g_{tfk}	4	4	5	8.0
EM-cGMM	(21)–(22)	–	4	100	8.3
EM-cGMM	(21)–(22)	–	4	50	8.6
EM-cGMM	(21)–(22)	–	4	10	8.4
EM-cGMM	(21)–(22)	–	4	5	7.5
AVI-cGMM	–	4	1	–	4.4
AuxIVA+	–	–	4	200	8.5
AuxIVA	–	–	2	200	4.7
PIT	–	1	1	–	7.4
DPCL	–	1	1	–	6.5

分離 DNN g_{tfk} と定位 DNN h_{kd} は，SGD の 1 つである Adam [25] を用いて学習した．Adam の学習率は 1.0×10^{-3} とし，学習データに対するコストが直前のエポックより大きくなる度に 0.7 倍した．スペクトログラム $\mathbf{x}_{tf}$ は，窓幅 512 サンプルでシフト幅 128 サンプルの短時間フーリエ変換を用いて得た．ハイパーパラメータ ν は M+10.0 とした．

提案法は，独立ベクトル分析法 (AuxIVA) [26] と，教師あり分離法である PIT および deep clustering (DPCL) [4] と比較した．AuxIVA は，音源数 K とマイク数 M が一致する決定条件を仮定するため，$M=2$ チャネル混合音で動作する．全 4 マイクを使用するため，仮想的に $M=K=4$ 個の音源が存在するとして AuxIVA を実行し，その後 $K=2$ となるよう分離音をクラスタリングする手法 (AuxIVA+) [27] を評価した．PIT と DPCL で使用した DNN は提案法の分離 DNN g_{tfk} と同じ設定である．DPCL の潜在変数の次元 D は 20 とした．また，EM-cGMM を学習済みの分離 DNN g_{tfk} で初期化する効果を評価するため，Otsuka ら [10] による初期化法で分離した場合を評価した．この手法は，まず音源数 K を多く準備し，方向 $d=1,\dots,D$ を K 個に分割して，方向ごとに TF マスクを割当て初期化する:

$$\hat{w}_{kd} \propto \begin{cases} 1 & (k-1)\frac{D}{K} \le d < k\frac{D}{K} \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

$$\hat{z}_{tfk} \propto \exp\left(-\sum_d \hat{w}_{kd}\mathbf{x}_{tf}^{\mathsf{H}}\mathbf{G}_{fd}^{-1}\mathbf{x}_{tf}\right). \tag{22}$$

本稿では，$K=6$ とした．分離性能は，信号対歪比 (SDR) [28, 29] を用いて評価した．

3.3 実験結果

分離音の評価結果を表 1 に示す[*2]．まず，事前学習したネットワーク g_{tfk} で単チャネル音源分離した結果 (AVI-cGMM) は，SDR が 4.4dB となった．EM-cGMM を用いた多チャネル分離法は，同じチャネル数で動作する AuxIVA+以上の性能となった．さらに，g_{tfk} を用いて初期化した場合は，従来の初期化法 (式 (21)–(22)) に比べ，約 0.8 dB 程度 SDR が改善した．ただし，反復回数が 50 回を超えるとどちらの場合も SDR が低下しているが，これは残響下ではスパース仮定が成り立たないためと数値誤差による影響と考えられる．

[*2] AuxIVA(+) のシフト幅が 256 で評価していたため再実験した．

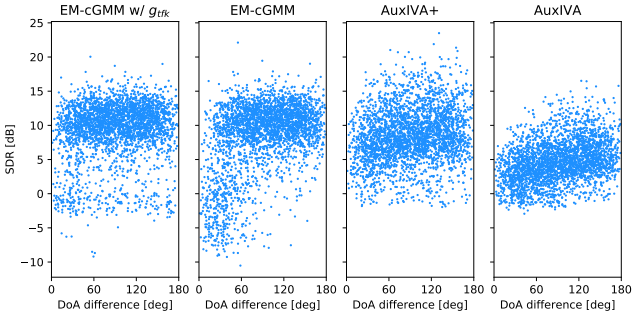

図 1 音源間の DoA 差と SDR の関係を表す散布図

図 1 に，音源間隔である DoA 差と各分離音の SDR の関係を示す[*2]．式 (21)–(22) で初期化した EM-cGMM は，DoA 差が 60° 以下になると急激に SDR が低下している．g_{tfk} を用いて初期化した場合は，そのような方向に依存する性能劣化が改善されている．一方，SDR が −3dB から 0dB 付近となる分離結果が方向によらず増加している．g_{tfk} は単チャネルスペクトログラムを入力とするので，同性の音声の混合音といった，スペクトル特徴から分離が難しい混合音で性能が下がっていると考えられる．教師あり学習を行う PIT や DPCL は，教師なし学習する AVI-cGMM に比べ 3dB 程度性能が高く，ネットワークの表現力の観点では改善の余地がある．償却変分推論時にパワースペクトル密度 λ_{tfk} と空間相関行列 $\mathbf{H}_{fd}$ も同時推論すれば，分離精度の改善が期待できる．

4. おわりに

本稿では，多チャネル混合音のみを用いて，深層音源分離を教師なし学習する枠組みについて述べた．本学習法は，TF マスクと DoA を潜在変数にもつ cGMM の一種である LDA をコスト関数として，TF マスクを推定する分離 DNN と DoA を推定する定位 DNN を学習する．シミュレーション混合音を用いた評価により，従来の初期化法より信号対歪比が改善することを確認した．本稿で述べた学習アルゴリズムによって，パーミュテーション問題を 1 つの確率モデルの中で解決しながら，深層音源分離を教師なし学習できるようになった．今後は，空間相関行列とパワースペクトル密度の同時推定による高精度化と，方向情報に基づく音源数推定を行う．

謝辞: 本研究の一部は科研費スタート支援 No. 18H06490 および基盤研究 (B) No. 19H04137 の支援を受けた．

参考文献

[1] Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N. and Fontecave, J.: The 2016 signal separation evaluation campaign, *International Conference on Latent Variable Analysis and Signal Separation*, pp. 323–332 (2017).

[2] Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B. and Virtanen, T.: DCASE 2017 challenge setup: Tasks, datasets and baseline system, *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, pp. 1–8 (2017).

[3] Barker, J., Marxer, R., Vincent, E. and Watanabe, S.: The third CHiME speech separation and recognition challenge: Dataset, task and baselines, *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 504–511 (2015).

[4] Hershey, J. R., Chen, Z., Le Roux, J. and Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 31–35 (2016).

[5] Yu, D., Kolbæk, M., Tan, Z.-H. and Jensen, J.: Permutation invariant training of deep models for speaker-independent multi-talker speech separation, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 241–245 (2017).

[6] Kolbæk, M., Yu, D., Tan, Z.-H. and Jensen, J.: Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 25, No. 10, pp. 1901–1913 (2017).

[7] Wang, Z.-Q., Le Roux, J. and Hershey, J. R.: Multichannel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation, pp. 1–5 (2018).

[8] Ozerov, A. and Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 3, pp. 550–563 (2010).

[9] Kim, T.: Real-time independent vector analysis for convolutive blind source separation, *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 57, No. 7, pp. 1431–1438 (2010).

[10] Otsuka, T., Ishiguro, K., Sawada, H. and Okuno, H. G.: Bayesian unification of sound source localization and separation with permutation resolution, *AAAI Conference on Artificial Intelligence*, pp. 2038–2045 (2012).

[11] Duong, N. Q., Vincent, E. and Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 7, pp. 1830–1840 (2010).

[12] Ito, N., Araki, S. and Nakatani, T.: Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing, *European Signal Processing Conference (EUSIPCO)*, pp. 1153–1157 (2016).

[13] Itakura, K., Bando, Y., Nakamura, E., Itoyama, K. and Yoshii, K.: A unified Bayesian model of time-frequency clustering and low-rank approximation for multi-channel source separation, *European Signal Processing Conference*, pp. 2280–2284 (2016).

[14] Higuchi, T., Ito, N., Yoshioka, T. and Nakatani, T.: Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5210–5214 (2016).

[15] Azcarreta, J., Ito, N., Araki, S. and Nakatani, T.: Permutation-free cGMM: complex Gaussian mixture model with inverse Wishart mixture model based spatial prior for permutation-free source separation and source counting, *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 51–55 (2018).

[16] Seetharaman, P., Wichern, G., Roux, J. L. and Pardo, B.: Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 356–360 (2019).

[17] Tzinis, E., Venkataramani, S. and Smaragdis, P.: Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information, *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 81–85 (2019).

[18] Drude, L., Hasenklever, D. and Haeb-Umbach, R.: Unsupervised training of a deep clustering model for multichannel blind source separation, *arXiv preprint arXiv:1904.01340 (accepted to ICASSP)* (2019).

[19] Drude, L., Heymann, J. and Haeb-Umbach, R.: Unsupervised training of neural mask-based beamforming, *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 695–699 (2019).

[20] Rezende, D. J. and Mohamed, S.: Variational inference with normalizing flows, *arXiv preprint arXiv:1505.05770* (2015).

[21] Ranganath, R., Gerrish, S. and Blei, D.: Black box variational inference, *Artificial Intelligence and Statistics*, pp. 814–822 (2014).

[22] Ritchie, D., Horsfall, P. and Goodman, N. D.: Deep amortized inference for probabilistic programs, *arXiv preprint arXiv:1610.05735* (2016).

[23] Drude, L., von Neumann, T. and Haeb-Umbach, R.: Deep attractor networks for speaker re-identification and blind source separation, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 11–15 (2018).

[24] Allen, J. B. and Berkley, D. A.: Image method for efficiently simulating small-room acoustics, *The Journal of the Acoustical Society of America*, Vol. 65, No. 4, pp. 943–950 (1979).

[25] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

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

[27] Kitamura, D., Ono, N., Sawada, H., Kameoka, H. and Saruwatari, H.: Relaxation of rank-1 spatial constraint in overdetermined blind source separation, *European Signal Processing Conference*, pp. 1261–1265 (2015).

[28] Vincent, E., Gribonval, R. and Févotte, C.: Performance measurement in blind audio source separation, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp. 1462–1469 (2006).

[29] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., Ellis, D. P. and Raffel, C. C.: mir_eval: A transparent implementation of common MIR metrics, *15th International Society for Music Information Retrieval Conference*, Citeseer (2014).