音源クラス識別器つき多チャンネル変分自己符号化器を用いた高速セミブラインド音源分離 *

◎李 莉 [1], 亀岡 弘和 [2], 牧野 昭二 [1]
[1] 筑波大学 [2] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

1 はじめに

ブラインド音源分離（Blind Source Separation: BSS）とは，音源信号や音源からマイクまでの伝達特性が未知の場合に，複数の音源信号が混合された観測信号から音源信号を推定する問題である．周波数領域で定式化される BSS のアプローチは，周波数ごとの音源分離の問題と周波数ごとに得られる分離信号がそれぞれどの音源のものであるかを対応づけるパーミュテーション整合と呼ぶ問題を併せて解く必要があるが，音源の混合過程を畳み込み演算を含まない瞬時混合系で表せるため比較的効率の良いアルゴリズムを実現できる利点がある．また，音源に関する時間周波数領域で成り立つ様々な仮定やマイクロホンアレーの周波数応答に関する仮定が有効活用できるようになることも大きな利点である．例えば，多チャンネル非負値行列因子分解（Multichannel Non-negative Matrix Factorization: MNMF）[1–4] は，各音源のパワースペクトログラムを非負値行列とみなし，二つの非負値行列の積で表現するアプローチである．これは，各時間フレームでパワースペクトルを時間変化する振幅でスケーリングされた基底スペクトルの線形和によって近似することに相当する．これにより MNMF は音源のスペクトル構造を手がかりにしながら周波数ごとの音源分離とパーミュテーション整合の同時解決を可能にしている．[3] では優決定条件に特化した MNMF の枠組が初めて導入され，その枠組は後年独立低ランク行列分析（Independent Low-Rank Matrix Analysis: ILRMA）と呼ばれている [4]．MNMF や ILRMA は，低ランク構造を持つ特定の音源に対して有効である一方で，限られた基底の線形和で正しく表現できない音源に対しては分離性能が制限される．

この問題を解決するため，ニューラルネットワーク（Neural Network: NN）が持つ豊かな関数表現能力を活かし，行列積に代わるパワースペクトログラムモデルとして NN を用いた手法が提案されている [5, 6]．独立深層学習行列分析（Independent Deeply Learned Matrix Analysis: IDLMA）[5] は，単一フレームのクリーンパワースペクトルを出力する NN を各音源ごとに事前学習し，音源分離アルゴリズムにおいて，学習した NN のフィードフォワード計算により各音源のパワースペクトルを更新する手法である．IDLMA は高い音源分離精度が得られることが実験的に示されているが，このアルゴリズムでは，音源のパワースペクトログラムを更新する際に尤度関数を増大させる保証がないため，分離行列の局所解への収束性が保証されない点に課題があった．一方，多チャンネル変分自己符号化器（Multichannel Variational Autoencoder: MVAE)[6] 法は，条件付き VAE（Conditional VAE: CVAE）により表現される音源スペクトログラムの生成モデルを事前学習し，分離時において CVAE のデコーダ入力を分離行列と共に推定する手法である．この手法では，各反復計算で尤度関数が上昇するようにパラメータが更新されるため，尤度関数の停留点への収束が保証される．しかし，この手法ではデコーダ入力値の更新に誤差逆伝播法（Backpropagation）が用いられるため，高い計算コストを要する点に課題があった．本稿では，MVAE の推論プロセスを代替する音源クラス識別器を組み込んだ多チャンネル変分自己符号化器を提案し，これを用いることで高速な音源分離アルゴリズムを実現できることを示す．

2 MVAE を用いた音源分離

2.1 問題の定式化

I 個のマイクロホンで J 個の音源から到来する信号を観測する場合を考える．マイク i の観測信号，音源 j の信号の複素スペクトログラムをそれぞれ $x_i(f,n)$, $s_j(f,n)$ とする．また，これらを要素としたベクトルを

$$\boldsymbol{x}(f,n) = [x_1(f,n), \ldots, x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I, \quad (1)$$

$$\boldsymbol{s}(f,n) = [s_1(f,n), \ldots, s_J(f,n)]^{\mathsf{T}} \in \mathbb{C}^J \quad (2)$$

とする．ただし，優決定条件下においては $I = J$ とおく．ここで $(\cdot)^{\mathsf{T}}$ は転置を表し，f と n はそれぞれ周波数と時間のインデックスである．音源信号ベクトル $\boldsymbol{s}(f,n)$ と観測信号ベクトル $\boldsymbol{x}(f,n)$ の間の関係式として瞬時分離系

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\mathsf{H}}(f)\boldsymbol{x}(f,n), \quad (3)$$

$$\boldsymbol{W}(f) = [\boldsymbol{w}_1(f), \ldots, \boldsymbol{w}_I(f)] \in \mathbb{C}^{I\times I} \quad (4)$$

を仮定する．ここで，$\boldsymbol{W}^{\mathsf{H}}(f)$ は分離行列を表し，$(\cdot)^{\mathsf{H}}$ はエルミート転置である．以上の瞬時混合系の仮定の下で，更に音源信号 j の複素スペクトログラム $s_j(f,n)$ が平均 0，分散 $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$ の複素正規分布

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)) \quad (5)$$

に従う確率変数とすると，各音源信号 $s_j(f,n)$ と $s_{j'}(f,n)$, $j \neq j'$ が統計的に独立のときには，音源信号 $\boldsymbol{s}(f,n)$ は

$$\boldsymbol{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}(f,n)|\boldsymbol{0}, \boldsymbol{V}(f,n)) \quad (6)$$

に従う．ここで，$\boldsymbol{V}(f,n)$ は $v_1(f,n), \ldots, v_I(f,n)$ を要素に持つ対角行列である．式 (3), (6) より，観測信号 $\boldsymbol{x}$ は

$$\boldsymbol{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}(f,n)|\boldsymbol{0}, (\boldsymbol{W}^{\mathsf{H}}(f))^{-1}\boldsymbol{V}(f,n)\boldsymbol{W}(f)^{-1}) \quad (7)$$

に従う．従って，分離行列 $\mathcal{W} = \{\boldsymbol{W}(f)\}_f$ と各音源のパワースペクトログラム $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$ が与えられた下での観測信号 $\mathcal{X} = \{\boldsymbol{x}(f,n)\}_{f,n}$ の対数条件

---

*Fast algorithm for semi-blind source separation using multichannel variational autoencoder with auxiliary source label classifier, Li Li (University of Tsukuba), Hirokazu Kameoka (NTT Communication Science Laboratories), Shoji Makino (University of Tsukuba)

付き分布は

$$\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) \stackrel{c}{=} 2N\sum_f \log|\det \boldsymbol{W}^{\mathsf{H}}(f)| - \sum_{f,n,j}\Big(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{v_j(f,n)}\Big) \quad (8)$$

となる．ここで，$\stackrel{c}{=}$ はパラメータに依存する項のみに関する等号を表す．音源パワースペクトログラム $v_j(f,n)$ に制約がない場合，式 (8) は周波数 f ごとの項に分解されるため，式 (8) に基づいて求める $\mathcal{W}$ で得られた分離信号のインデックスにはパーミュテーションの任意性が生じる．$v_j(f,n)$ が周波数方向に構造的制約を持つ場合，その制約を活かすことでパーミュテーション整合と音源分離を同時解決するアプローチを導くことができる．独立ベクトル分析 (Independent Vector Analysis: IVA) や ILRMA がその例である．

2.2 MVAE

MVAE は，各音源の複素スペクトログラムの生成モデルとして，音源クラスラベルを補助入力とした CVAE のデコーダ分布を用いた混合信号モデルである．ある音源信号の複素スペクトログラムを $\boldsymbol{S} = \{s(f,n)\}_{f,n}$ とし，対応する音源クラスラベルを one-hot ベクトル c とする．CVAE はエンコーダ分布 $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ とデコーダ分布 $p_\theta(\boldsymbol{S}|\boldsymbol{z},c)$ が無矛盾になるように，かつ $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ と $p_\theta(\boldsymbol{S}|\boldsymbol{z},c)$ から導かれる事後分布 $p_\theta(\boldsymbol{z}|\boldsymbol{S},c) \propto p_\theta(\boldsymbol{S}|\boldsymbol{z},c)p(\boldsymbol{z})$ ができるだけ一致するようにエンコーダとデコーダの NN パラメータ ϕ，θ を学習する．ここで，CVAE のデコーダ分布を式 (5) の局所ガウス音源モデルと同形の確率モデル

$$p_\theta(\boldsymbol{S}|\boldsymbol{z},c,g) = \prod_{f,n}\mathcal{N}_{\mathbb{C}}(s(f,n)|0,v(f,n)), \quad (9)$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\boldsymbol{z},c) \quad (10)$$

とする．ただし，分散 $\sigma_\theta^2(f,n;\boldsymbol{z},c)$ はデコーダネットワークの出力であり，g はパワースペクトログラムのスケールを表す変数である．一方，エンコーダ分布 $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ は通常の CVAE と同様に，標準正規分布

$$q_\phi(\boldsymbol{z}|\boldsymbol{S},c) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_\phi(\boldsymbol{S},c),\mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\boldsymbol{S},c))) \quad (11)$$

と仮定する．ここで，$\boldsymbol{\mu}_\phi(\boldsymbol{S},c)$, $\boldsymbol{\sigma}_\phi^2(\boldsymbol{S},c)$ はエンコーダの出力である．CVAE のパラメータ θ, ϕ は，各種クラスの音源信号の複素スペクトログラムの学習サンプル $\{\boldsymbol{S}_m,c_m\}_{m=1}^M$ を用いて

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{S},c)}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},c)] - KL[q_\phi(\boldsymbol{z}|\boldsymbol{S},c)||p(\boldsymbol{z})]] \quad (12)$$

が最大となるように学習される．$\mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\cdot]$ は学習サンプルによる標本平均を表し，$KL[\cdot||\cdot]$ は KL ダイバージェンスである．以上により学習したデコーダ分布 $p_\theta(\boldsymbol{S}|\boldsymbol{z},c,g)$ を CVAE 音源モデルと呼ぶ．CVAE 音源モデルは，学習サンプルに含まれる様々なクラスの音源の複素スペクトログラムを表現可能なユニバーサルな生成モデルとなっており，c は音源クラスのカテゴリカルな特徴を調整する役割，$\boldsymbol{z}$ はクラス内の変動を調整する役割を担った変数と見なせる．

音源 j の複素スペクトログラム $\boldsymbol{S}_j = \{s_j(f,n)\}_{f,n}$ の生成モデルを，$\boldsymbol{z}_j$，c_j，g_j を入力としたデコーダ分布により表現することで，音源モデルのパラメータの尤度関数は式 (8) と同形の尤度関数に帰着させることができる．従って，式 (8) が大きくなるように分離行列 $\mathcal{W}$，CVAE 音源モデルパラメータ $\Psi = \{\boldsymbol{z}_j,c_j\}_j$，スケールパラメータ $\mathcal{G} = \{g_j\}_j$ を反復更新することで，式 (8) の停留点を探索することができる．式 (8) を上昇させる $\mathcal{W}$ の更新には ILRMA，IDLMA と同様に反復射影法 (Iterative Projection: IP)

$$\boldsymbol{w}_j(f) \leftarrow (\mathbf{W}^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f))^{-1}\boldsymbol{e}_j, \quad (13)$$

$$\boldsymbol{w}_j(f) \leftarrow \frac{\boldsymbol{w}_j(f)}{\sqrt{\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f)\boldsymbol{w}_j(f)}} \quad (14)$$

を用いることができる．ただし，$\boldsymbol{\Sigma}_j(f) = \frac{1}{N}\sum_n \frac{\boldsymbol{x}(f,n)\boldsymbol{x}^{\mathsf{H}}(f,n)}{v_j(f,n)}$ であり，$\boldsymbol{e}_j$ は $I\times I$ の単位行列 $\boldsymbol{I}$ の第 j 列ベクトルである．また式 (8) を上昇させる Ψ の更新は誤差逆伝播法，$\mathcal{G}$ の更新は

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z}_j,c_j)} \quad (15)$$

により行うことができる．ただし，式 (15) は $\mathcal{W}$ と Ψ が固定された下で式 (8) を最大にする更新式である．以上より MVAE の推論プロセスは以下のようにまとめられる．

1. 式 (12) を学習規準として θ，ϕ を学習する．
2. $\mathcal{W}$，Ψ を初期化する．
3. 各 j について下記ステップを繰り返す．
   (a) 式 (13), (14) により $\boldsymbol{w}_j(0),\ldots,\boldsymbol{w}_j(F)$ を更新する．
   (b) 誤差逆伝播法により $\Psi_j = \{\boldsymbol{z}_j,c_j\}$ を更新する．
   (c) 式 (15) により g_j を更新する．

音源クラスベクトル c_j はテスト時において推定されるパラメータになるため，MVAE は音源分離と音源クラス識別を同時に行うことができる．以上が MVAE 法である．

3 提案手法：音源クラス識別器つき MVAE

MVAE 法では，各反復計算で対数尤度が上昇するようにパラメータの更新が行われるため，対数尤度の停留点への収束が保証される利点がある一方で，誤差逆伝播法を用いた $\Psi_j = \{\boldsymbol{z}_j,c_j\}$ の更新に多大な計算コストを要する点に課題があった．CVAE の学習ではエンコーダ分布 $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ が $p_\theta(\boldsymbol{z}|\boldsymbol{S},c)$ を近似するものとして得られるが，$p_\theta(c|\boldsymbol{S})$ を近似するネットワーク $r_\psi(c|\boldsymbol{S})$ を同様に得ることができれば，誤差逆伝播法による $p_\theta(\boldsymbol{z},c|\boldsymbol{S}) = p_\theta(\boldsymbol{z}|\boldsymbol{S},c)p_\theta(c|\boldsymbol{S})$ の最大値探索を $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$，$r_\psi(c|\boldsymbol{S})$ のフォワード計算に（近似的に）置き換えることができるため，大幅な高速化が可能となる．これを実現するアイディアとして，以下，クラス識別器つき VAE を導入した MVAE 法の高速化版を提案する．本稿では提案法を「FastMVAE 法」と呼ぶ．

3.1 クラス識別器つき VAE

クラス識別器つき VAE（Auxiliary Classifier VAE: ACVAE）[7] は，音声変換に応用する目的で提案された CVAE の拡張版で，クラスラベル入力 c のデコーダ出力への影響力を強調するためにデコーダ出力とクラスラベル c との相互情報量を正則化項としてエンコー

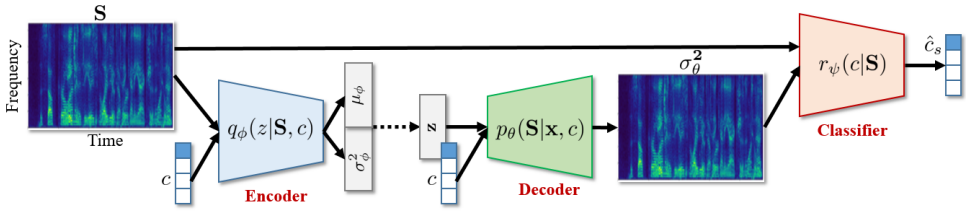

Fig. 1　ACVAE の概念図.

ダとデコーダを学習する方式である．潜在変数 $\boldsymbol{z}$ が与えられたときの $\boldsymbol{S}$ と c の相互情報量は

$$I(c,\boldsymbol{S}|\boldsymbol{z}) \tag{16}$$
$$=\mathbb{E}_{c\sim p(c),\boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c),c'\sim p(c|\boldsymbol{S})}[\log p(c'|\boldsymbol{S})]+H(c)$$

と書ける．ここで，$H(c)$ はクラス c のエントロピーを表し，定数項と見なせる．式 (16) のとおり，相互情報量 $I(c,\boldsymbol{S}|\boldsymbol{z})$ を求めるには事後分布 $p(c|\boldsymbol{S})$ を計算する必要があるが，$p_\theta(\boldsymbol{S}|\boldsymbol{z},c)$ から $p(c|\boldsymbol{S})$ を解析的に求めることは難しく，$I(c,\boldsymbol{S}|\boldsymbol{z})$ を規準に含めて最適化することは困難である．そこで ACVAE では $I(c,\boldsymbol{S}|\boldsymbol{z})$ の代わりに，式 (16) の右辺第一項の変分下界

$$\begin{aligned}
&\mathbb{E}_{c\sim p(c),\boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c),c'\sim p(c|\boldsymbol{S})}[\log p(c'|\boldsymbol{S})]\\
=&\mathbb{E}_{c\sim p(c),\boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c),c'\sim p(c|\boldsymbol{S})}[\log \frac{r(c'|\boldsymbol{S})p(c'|\boldsymbol{S})}{r(c'|\boldsymbol{S})}]\\
\geq&\mathbb{E}_{c\sim p(c),\boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c),c'\sim p(c|\boldsymbol{S})}[\log r(c'|\boldsymbol{S})]\\
=&\mathbb{E}_{c\sim p(c),\boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c)}[\log r(c|\boldsymbol{S})]
\end{aligned} \tag{17}$$

を規準として学習を行う．この不等式は $r(c|\boldsymbol{S})=p(c|\boldsymbol{S})$ のときに等号が成立する．よって，この変分下界を大きくすることは分布 $r(c|\boldsymbol{S})$ を用いて事後分布 $p(c|\boldsymbol{S})$ を近似することに相当し，この変分下界を $r(c|\boldsymbol{S})$ と θ に関して上昇させることで，$I(c,\boldsymbol{S}|\boldsymbol{z})$ を間接的に大きくすることができる．ここで，$p_\theta(\boldsymbol{z}|\boldsymbol{S},c)$ の近似分布のモデル化と同様に，データ $\boldsymbol{S}$ が与えられた下でのクラスラベル c の条件付分布 $r_\psi(c|\boldsymbol{S})$ のパラメータを出力する NN を用いて近似分布 $r(c|\boldsymbol{S})$ をモデル化し，NN のパラメータ ψ を

$$\mathcal{L}(\phi,\theta,\psi) \tag{18}$$
$$=\mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c),q_\phi(\boldsymbol{z}|\boldsymbol{S},c)}[\mathbb{E}_{c\sim p(c),\boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c)}[\log r_\psi(c|\boldsymbol{S})]].$$

が大きくなるようエンコーダ及びデコーダのパラメータとともに学習する．また，ラベル付き学習サンプル $\{\boldsymbol{S}_m,c_m\}_{m=1}^M$ も学習に用いることができるため，学習データ $\boldsymbol{S}_m$ と対応するクラスラベル c_m の交差エントロピー

$$\mathcal{I}(\psi)=\mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\log r_\psi(c|\boldsymbol{S})] \tag{19}$$

も学習規準に含めることができる．従って，ACVAE の NN パラメータ学習規準は

$$\mathcal{J}(\phi,\theta)+\lambda_{\mathcal{L}}\mathcal{L}(\phi,\theta,\psi)+\lambda_{\mathcal{I}}\mathcal{I}(\psi) \tag{20}$$

となる．ここで，$\lambda_{\mathcal{L}}\geq 0$ と $\lambda_{\mathcal{I}}\geq 0$ は各規準の重み係数である．Fig. 1 に ACVAE の概念図を示す．

ACVAE で学習されるクラス識別器 $r_\psi(c|\boldsymbol{S})$ は $p(c|\boldsymbol{S})$ を近似した分布となるため，学習した $r_\psi(c|\boldsymbol{S})$ とエンコーダ分布 $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ の積は $p(\boldsymbol{z},c|\boldsymbol{S})$ を近似した分布となる．MVAE 法では音源 j ごとに $p(\boldsymbol{z}_j,c_j|\boldsymbol{S}_j)$ が最大となる $\Psi_j=\{\boldsymbol{z}_j,c_j\}$ を探索するため誤差逆伝播法が用いられたが，ACVAE を用いることによりこのパラメータ更新部を $r_\psi(c|\boldsymbol{S})$ と $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ のフィードフォワード計算に置き換えることができる．

3.2　高速アルゴリズム

ACVAE で学習したエンコーダとクラス識別器を q_ϕ，r_ψ とすると，$\boldsymbol{S}_j$ のエンコーダ出力およびクラス確率は $q_\phi(\boldsymbol{z}_j|\boldsymbol{S}_j,c_j)$，$r_\psi(c_j|\boldsymbol{S}_j)$ で与えられる．従来の MVAE 法における $p_\theta(\boldsymbol{z}_j,c_j|\boldsymbol{S}_j)$ の最大化ステップを，$q_\phi(\boldsymbol{z}_j|\boldsymbol{S}_j,c_j)$ および $r_\psi(c_j|\boldsymbol{S}_j)$ のフォワード計算に置き換えることにより，以下のアルゴリズムを得る．これを fMVAE アルゴリズムと呼ぶ．

1. 式 (20) を学習規準として θ，ϕ，ψ を学習する．
2. $\mathcal{W}$ を初期化する．
3. 各 j について下記ステップを繰り返す．
   (a) 式 (13), (14) により $\boldsymbol{w}_j(0),\ldots,\boldsymbol{w}_j(F)$ を更新する．
   (b) 音源クラス識別器の出力 $r_\psi(c_j|\boldsymbol{S}_j)$（連続値ベクトル）に c_j を更新する．
   (c) エンコーダ $q_\theta(\boldsymbol{z}_j|\boldsymbol{S}_j,c_j)$ の平均値に $\boldsymbol{z}_j$ を更新する．
   (d) 式 (15) により g_j を更新する．

このアルゴリズムは NN のフィードフォワード計算で音源モデルのパラメータ $\Psi_j=\{\boldsymbol{z}_j,c_j\}$ を更新できるため，MVAE に比べ計算時間を大幅に削減できる．

4　評価実験

提案手法による音声分離及び識別性能を検証するため，Voice Conversion Challenge (VCC) 2018 音声データベース [9] を用いて既存手法 ILRMA[4], IDLMA[5], MVAE[6] と提案手法の比較実験を行った．

音声データは男性話者 2 名（SM1, SM2）と女性話者 2 名（SF1, SF2）の発話データを用いた．各話者の発話データ 116 文のうち 81 文を CVAE と ACVAE の学習データとし，残りの 35 文を用いて評価用データを作成した．多チャンネル観測信号は鏡像法を用いて残響時間 (RT_{60}) をそれぞれ 78 ms と 351 ms として生成したインパルス応答と，RWCP データベース [10] に収録された ANE(RT_{60}=173 ms) と E2A(RT_{60}=225 ms) のインパルス応答，それぞれと発話データを畳み込んで作成したデータの 2 種類を用いた．各残響環境において 4 パターンの話者の組み合わせ（SF1+SM1, SM1+SM2, SM2+SF2, SF1+SF2 の）について計 40 文の混合信号を作成した．すべての音声信号のサンプリング周波数を 16 kHz とし，フレーム長 256 ms，フレームシフト 128 ms で短時間 Fourier 変換を行い，観

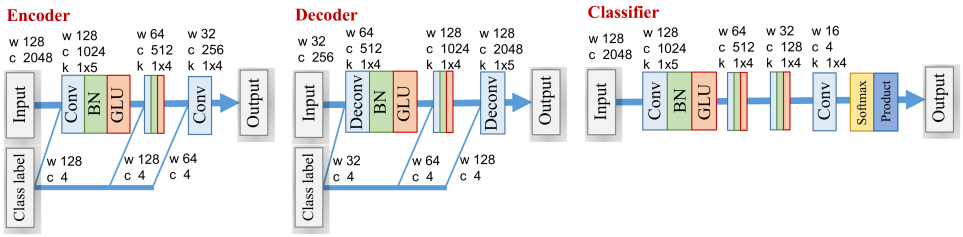

Fig. 2　ACVAE のネットワーク構造.

Table 1　4 種類の残響環境下における各手法で得られた SDR，SIR，SAR の平均値.

method	SDR [dB]	SIR [dB]	SAR [dB]
ILRMA	9.0089	15.5027	11.9961
IDLMA	6.3979	11.1837	9.7275
MVAE	13.2996	21.2205	15.4373
fMVAE	**13.9827**	**22.1369**	**15.6460**

Table 2　各手法の計算時間.

method	iteration[sec]	total [sec]
MVAE (GPU)	6.079274	202.1153
fMVAE (CPU)	0.360914	20.57482
fMVAE (GPU)	**0.069550**	**10.902709**
ILRMA (CPU)	0.117794	12.675797

Table 3　音源クラス識別率.

method	all iterations	final estimation
MVAE	30.63%	40.63%
fMVAE	**78.35%**	**78.75%**

測信号 $\boldsymbol{x}(f,n)$ を算出した．Fig. 2 に ACVAE のネットワーク構造を示す．通常の CVAE で音源モデルを学習したときのエンコーダとデコーダネットワーク構造も Fig. 2 に示した構造を用いた．ILRMA と IDLMA は 100 回反復更新を行った．また，MVAE と fMVAE は ILRMA を 30 回反復した $\mathcal{W}$ を初期値として 30 回反復更新を行った．すべての実験は Intel (R) Core i7-6800K CPU@3.40 GHz と GeForce GTX 1080Ti GPU を用いて行った．

Table 1 には，音源分離性能を評価する signal-to-distortion ratios (SDRs)，signal-to-interference ratios (SIRs) と signal-to-artifacts ratios (SARs)[11] の結果を示す．MVAE と fMVAE は ILRMA と IDLMA より高い分離性能を示した．更に，fMVAE は MVAE より一層の性能向上が得られた．IDLMA のネットワーク構造は [5] に記載されている構造に 1 層を増やしたものを用いた．この構造は学習データに対して最適ではない可能性があり，IDLMA が ILRMA より低い分離性能を導いた原因だと考えられる．Table 2 に各手法について 1 回パラメータ更新を行う計算時間とアルゴリズム全体の計算時間を 160 回の試行の平均値を示す．提案手法 fMVAE は MVAE より約 20 倍速くなったことが確認できた．更に GPU を利用した場合には ILRMA より速く分離を行うことができた．音源クラス識別率を Table 3 として示す．音源クラス識別器を用いた fMVAE は誤差逆伝播法で音源クラスベクトル c_j を推定する MVAE に比べ識別率が約 40% 向上した．

5　おわりに

MVAE は，各音源のスペクトログラムの生成過程を CVAE を用いてモデル化した混合信号のモデルであり，これを用いることで収束性が保証された反復アルゴリズムによりパーミュテーションフリーかつ高精度なセミブラインド音源分離を実現することができる．従来の MVAE では音源モデルパラメータの更新に誤差逆伝播法が用いられたため，高い計算コストを要していた．本稿では，この推論プロセスを代替する音源クラス識別器を組み込んだ MVAE を提案し，これを用いることで高速な音源分離及び高精度な音源クラス識別アルゴリズムが実現できることを実験により確認した．

謝辞　本研究は JSPS 科研費 17H01763 と 18J20059 及び SECOM 科学技術振興財団の助成を受けて行われた．また，本研究に関して多くの議論を頂いた筑波大学の井上翔太氏に謝意を表する．

参考文献

[1] A. Ozerov *et al.*, IEEE Trans. ASLP, vol. 18, no. 3, pp. 550–563, 2010.
[2] H. Sawada *et al.*,, IEEE Trans. ASLP, vol. 21, no. 5, pp. 971–982, 2013.
[3] H. Kameoka *et. al.*, in Proc. LVA/ICA, pp. 245–253, 2010.
[4] D. Kitamura *et. al.*, IEEE/ACM Trans. ASLP, vol. 24, no. 9, pp. 1622–1637, 2016.
[5] S. Mogami *et. al.*, in Proc. EUSIPCO, pp. 1571–1575, 2018
[6] H. Kameoka*et. al.*, epint arXiv: 1808.00892, 2018.
[7] H. Kameoka *et. al.*, eprint arXiv: 1808.05092, 2018.
[8] X. Chen *et. al.*, in Proc. NIPS, pp. 2172–2180, 2016.
[9] J. Lorenzo-Trueba *et. al.*, eprint arXiv: 1804.04262, 2018.
[10] S. Nakamura *et. al.*, in Proc. LREC, pp 965–968, 2000.
[11] E. Vincent *et. al.*, IEEE Trans. ASLP, vol. 14, no. 4, pp. 1462–1469, 2006.