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
EA2019-77(2019-12)

多チャンネル変分自己符号化器法による任意話者の音源分離

李　莉†　　亀岡　弘和††　井上　翔太†　　牧野　昭二†

† 筑波大学大学院 システム情報工学研究科　〒 305-0877 茨城県つくば市天王台 1-1-1
†† 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
〒 243–0198 神奈川県厚木市森の里若宮 3–1
E-mail: †lili@mmlab.cs.tsukuba.ac.jp, s1920622@s.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp,
††hirokazu.kameoka.uh@hco.ntt.co.jp

あらまし　多チャンネル変分自己符号化器 (MVAE) は，各音源のスペクトログラムの生成過程を条件付変分自己符号化器 (CVAE) を用いてモデル化した混合信号のモデルであり，これを用いたパーミュテーションフリーかつ高精度な音源分離手法である MVAE 法，およびその計算コストを大幅に削減した FastMVAE 法が提案されている．MVAE 法と FastMVAE 法は教師あり音源分離法に位置づけられるが，本稿では，十分なデータでネットワークを学習させることによりいずれの手法も任意話者に対する音源分離を既知話者の場合と同等の性能で行えることを示す．また，Product-of-Experts に基づいて潜在空間変数の事前確率を考慮した推論アルゴリズムを提案する．話者依存及び任意話者の音源分離実験において提案法の高い分離性能を確認した．

キーワード　多チャンネル音源分離、話者分離、多チャネル変分自己符号化器（MVAE）法、FastMVAE 法、条件付き変分自己符号化器 (CVAE)

Speaker-independent source separation with multichannel variational autoencoder

Li LI†, Hirokazu KAMEOKA††, Shota INOUE†, and Shoji MAKINO†

† Graduate School of Systems and Information Engineering, University of Tsukuba
Tennodai 1-1-1, Tsukuba-shi, Ibaraki, 305-0877 Japan
†† NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
Morinosatowakamiya 3–1, Atsugi-shi, Kanagawa, 243–0198 Japan
E-mail: †lili@mmlab.cs.tsukuba.ac.jp, s1920622@s.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp,
††hirokazu.kameoka.uh@hco.ntt.co.jp

Abstract　The multichannel variational autoencoder method (MVAE) is a recently proposed determined source separation method, which uses a conditional variational autoencoder (CVAE) to learn the spectrograms of source signals given a source-class ID as an auxiliary input. The trained decoder distribution can be used as a universal generative model capable of generating spectrograms of all the sources involved in the training samples. The decoder distribution can then be exploited to estimate the spectrograms of sources in a mixture. The MVAE methods, including the original MVAE method and its fast version called FastMVAE, were shown to significantly outperform conventional methods under speaker-dependent conditions, where the target speakers are seen in the training dataset. In this paper, we investigate the performances of the two MVAE methods under speaker-independent conditions. To further enhance the ability of FastMVAE to estimate the latent space variables for unknown speakers, we propose a prior-aware inference algorithm based on the concept of product-of-experts. Experimental results revealed that the MVAE methods could perform well even under speaker-independent conditions.

Key words　Multichannel source separation, determined source separation, multichannel variational autoencoder (MVAE), FastMVAE, conditional variational autoencoder (CVAE)

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，周波数帯域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点も特長の 1 つである．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，各時刻のパワースペクトルを限られた個数の基底スペクトルの線形和で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帯域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，こ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，低ランク行列で近似可能なスペクトログラムを持つ特定の音源に対して有効である一方で，音声など，それ以外の音源に対しては分離性能が限定的となる．

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]〜[8]．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の各反復計算において，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報告されているが，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加させる保証がないため，反復アルゴリズム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], [7] 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減少しないようにデコーダ入力値が誤差逆伝搬法（Backpropagation）により更新されるため，尤度関数の停留点への収束が保証される．その一方，デコーダ入力の推定に高い計算コストを要する点に課題があった．そこで，[8] で我々は当該処理に要する計算コストを削減する目的で，音源クラス識別器を利用して CVAE を学習する方式を考え，誤差逆伝播法の代わりにエンコーダと音源クラス識別器を用いてデコーダ入力の最適な更新値を近似的に求められる FastMVAE（または fMVAE）法を提案している．

MVAE 法と FastMVAE 法は，教師あり音源分離法に位置づけられ，NN の学習データに含まれる話者に対して優れた分離性能が得られることが実験的に検証されている．しかしながら，利用場面によっては話者が未知の場合もあるため，提案法が不特定話者に対しても遜色なく動作するかどうかを検証することも重要である．そこで，本稿では，多数の話者を含む学習データを用いることにより CVAE が不特定話者の音源モデルとしても機能するかどうかを検証するため，学習データに含まれない任意話者に対する MVAE 法及び FastMVAE 法の音源分離性能を評価する．更に，FastMVAE 法が未知話者の場合でも安定かつ高精度で推論を行わせるため，エンコーダの確率モデルと潜在空間変数の事前確率モデルを Product-of-Experts (PoE) [9] の枠組により統合し，新たな推論アルゴリズムを提案する．

2. MVAE による音源分離

2.1 問題の定式化

I 個のマイクで 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

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

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

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 $\boldsymbol{x}(f,n)$ の間の関係として瞬時分離系

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

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

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

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

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

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

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

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

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

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

となる．ここで，$=^c$ 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，(8) は周波数 f ごとの項に分解されるため，各周波数帯域の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．しかし，$v_j(f,n)$ 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帯域ごとの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) [10], [11] や ILRMA がその例である．

2.2 MVAE 法

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 の学習はエンコーダ分布 $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ とデコーダ分布 $p_\theta(\boldsymbol{S}|\boldsymbol{z},c)$ 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 を探索することにより実現される．ここで，CVAE のデコーダ分布を (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 $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ と潜在空間変数 $\boldsymbol{z}$ の事前確率分布は通常の CVAE と同様に，多次元正規分布

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

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I}) \quad (12)$$

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

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{S},c)}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},c)] - KL[q_\phi(\boldsymbol{z}|\boldsymbol{S},c)||p(\boldsymbol{z})]] \quad (13)$$

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 (8) と同形の尤度関数に帰着させることができる．従って，(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，(8) の停留点を探索することができる．(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) [12]:

$$\boldsymbol{w}_j(f) \leftarrow (\mathbf{W}^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f))^{-1}\boldsymbol{e}_j, \quad (14)$$

$$\boldsymbol{w}_j(f) \leftarrow \frac{\boldsymbol{w}_j(f)}{\sqrt{\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f)\boldsymbol{w}_j(f)}} \quad (15)$$

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 (8) を上昇させる Ψ の更新は誤差逆伝播法，$\mathcal{G}$ の更新は

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n} \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z}_j,c_j)} \quad (16)$$

により行う．ただし，(16) は $\mathcal{W}$ と Ψ が固定された下で (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） (13) 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．
  (3a) 式 (14), (15) により $\boldsymbol{w}_j(0), \ldots, \boldsymbol{w}_j(F)$ を更新する．
  (3b) 誤差逆伝播法により $\Psi_j = \{\boldsymbol{z}_j, c_j\}$ を更新する．
  (3c) (16) 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．

2.3 FastMVAE 法

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．これを実現するためにクラス識別器つき VAE（Auxiliary Classifier VAE: ACVAE）[13] を導入したのが FastMVAE 法である．

ACVAE 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

$$I(c, \boldsymbol{S}|\boldsymbol{z}) = \mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c), c'\sim p(c|\boldsymbol{S})}[\log p(c'|\boldsymbol{S})] + H(c) \tag{17}$$

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．(17) 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，(17) の右辺第一項の変分下界

$$\begin{aligned}&\mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c), c'\sim p(c|\boldsymbol{S})}[\log p(c'|\boldsymbol{S})] \\ &\qquad \geqq \mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c)}[\log r(c|\boldsymbol{S})]\end{aligned} \tag{18}$$

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

$$\begin{aligned}&\mathcal{L}(\phi, \theta, \psi) \\ &= \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c), q_\phi(\boldsymbol{z}|\boldsymbol{S},c)}[\mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c)}[\log r_\psi(c|\boldsymbol{S})]].\end{aligned} \tag{19}$$

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

$$\mathcal{I}(\psi) = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\log r_\psi(c|\boldsymbol{S})] \tag{20}$$

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

$$\mathcal{J}(\phi, \theta) + \lambda_{\mathcal{L}}\mathcal{L}(\phi, \theta, \psi) + \lambda_{\mathcal{I}}\mathcal{I}(\psi) \tag{21}$$

となる．ここで，$\lambda_{\mathcal{L}} \geqq 0$ と $\lambda_{\mathcal{I}} \geqq 0$ は各規準の重み係数である．

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 $\Psi_j = \{\boldsymbol{z}_j, c_j\}$ を探索するために誤差逆伝播法が用いられたが，FastMVAE 法では 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．ただし，クラスラベル c はクラス識別器が出力した連続値

$$c \leftarrow r_\psi(c|\boldsymbol{S}) \tag{22}$$

を用いた fMVAE_c と，学習データと同様の one-hot ベクトル形式の離散表現

$$c \leftarrow \underset{c\in\{1,2,\ldots,C\}}{\operatorname{argmax}}\ r_\psi(c|\boldsymbol{S}) \tag{23}$$

に整形した fMVAE_o の 2 バージョンを考える．従って，FastMVAE 法の推論プロセスは以下のようにまとめられる．

（1） (21) を学習規準として θ，ϕ，ψ 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．
- (3a) 式 (14), (15) により $\boldsymbol{w}_j(0), \ldots, \boldsymbol{w}_j(F)$ を更新する．
- (3b) (22) や (23) により c_j を更新する．
- (3c) エンコーダのフォワード計算により $\boldsymbol{z}_j$ を更新する．
- (3d) (16) により g_j を更新する．

3. 提案手法：潜在空間変数の事前確率を考慮した推論アルゴリズム

[8] で我々は，上記の近似計算により FastMVAE 法は MVAE 法と同等な分離性能を実現しつつ，計算時間を 1/20 に削減できることを確認した．しかし，FastMVAE 法では，潜在空間変数 $\boldsymbol{z}$ の更新においてエンコーダ $q_\phi(\boldsymbol{z}|\boldsymbol{S}, c)$ のフィードフォワード計算が用いられるが，CVAE 音源モデルの学習時に仮定した潜在空間変数 $\boldsymbol{z}$ の事前確率分布 $p(\boldsymbol{z})$ を考慮した最適化規準と同一になっていなかった．そのため，未知話者を対象とした場合，推定される潜在空間変数が仮定した分布から逸脱し，分離性能の劣化を招いていた（後述）．この点に着目し，本章では，学習と推論時の最適化規準のミスマッチを解消し，未知話者の場合においても安定かつ高精度な推論を行うために，エンコーダの確率モデルと潜在空間変数の事前確率モデルを Product-of-Experts (PoE) の枠組により統合した確率モデルに基づいて $\boldsymbol{z}$ の更新式を導出する．

PoE は観測データの確率分布を Experts と呼ばれる個々の確率分布の積で表現する手法であり，そのモデルは複数の Experts の論理積を取ったような確率モデルとなる．従って，エンコーダ分布と $\boldsymbol{z}$ の事前確率分布を Experts として統合した確率モデルの最尤推定量は

$$\begin{aligned}\hat{\boldsymbol{z}} &= \underset{\boldsymbol{z}}{\operatorname{argmax}}\, p(\boldsymbol{z}|\boldsymbol{S}, c)p(\boldsymbol{z})^\alpha \\ &\approx \underset{\boldsymbol{z}}{\operatorname{argmax}}\, q_\phi(\boldsymbol{z}|\boldsymbol{S}, c)p(\boldsymbol{z})^\alpha \\ &= \underset{\boldsymbol{z}}{\operatorname{argmax}} \log q_\phi(\boldsymbol{z}|\boldsymbol{S}, c) + \alpha \log p(\boldsymbol{z})\end{aligned} \tag{24}$$

と定義できる．ここで，α は事前確率分布の重みを表すパラメータである．また，$q_\phi(\boldsymbol{z}|\boldsymbol{S}, c)$ と $p(\boldsymbol{z})$ は多次元正規分布であるため，(24) の確率モデルは

$$\begin{aligned}&\log q_\phi(\boldsymbol{z}|\boldsymbol{S}, c) + \alpha \log p(\boldsymbol{z}) \\ &\overset{c}{=} -\frac{1}{2}\big(\boldsymbol{z} - \boldsymbol{\mu}_\phi(|\boldsymbol{S}, c)\big)^\mathsf{T}\boldsymbol{\Sigma}_\phi^{-1}\big(\boldsymbol{z} - \boldsymbol{\mu}_\phi(\boldsymbol{S}, c)\big) - \frac{\alpha}{2}\boldsymbol{z}^\mathsf{T}\boldsymbol{z} \\ &\overset{c}{=} -\frac{\boldsymbol{\Sigma}_\phi^{-1} + \alpha\boldsymbol{I}}{2}(\boldsymbol{z} - \boldsymbol{\mu})^\mathsf{T}(\boldsymbol{z} - \boldsymbol{\mu})\end{aligned} \tag{25}$$

のように変形できる．ただし，$\boldsymbol{\Sigma}_\phi = \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\boldsymbol{S}, c))$ であり，

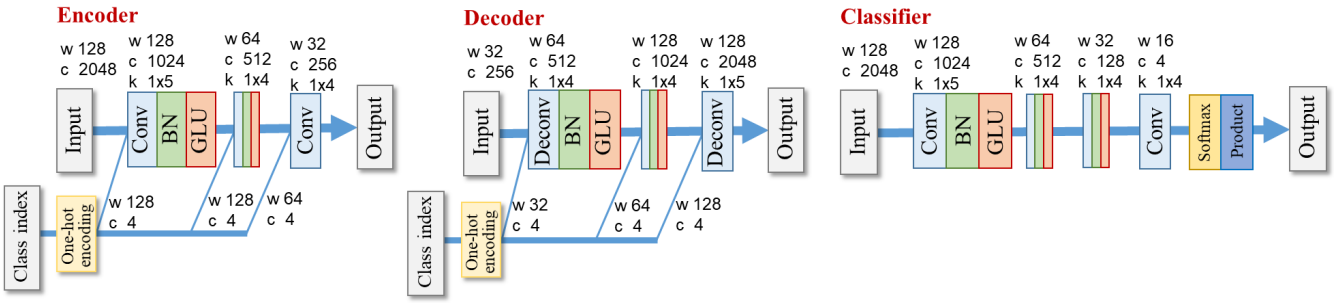

図 1: ACVAE のネットワーク構造. 入力と出力スペクトログラムの周波数軸をチャネル軸と見なし，1 次元畳み込み NN を用いる．“w”，“c” と “k” はそれぞれデータの系列長，チャネル数とフィルタサイズを表し，“Conv”，“Deconv”，“BN” と “GLU” はそれぞれ畳み込み，逆畳み込み，バッチ正規化と gated linear unit と呼ばれるデータ駆動の活性化関数を表す．

$\boldsymbol{\mu} = \frac{\boldsymbol{\Sigma}_\phi^{-1}}{\boldsymbol{\Sigma}_\phi^{-1}+\alpha \boldsymbol{I}}\boldsymbol{\mu}_\phi(\boldsymbol{S}, c)$ である．よって，二次関数である (25) の最大値で与えられる $\boldsymbol{z}$ の更新式は

$$\boldsymbol{z} \leftarrow \frac{\boldsymbol{\Sigma}_\phi^{-1}}{\boldsymbol{\Sigma}_\phi^{-1}+\alpha \boldsymbol{I}}\boldsymbol{\mu}_\phi(\boldsymbol{S}, c) \tag{26}$$

となる．$\alpha = 0$ の時には，(26) は従来の FastMVAE 法で用いられた $\boldsymbol{z}$ の更新式と等しい．

4. 評 価 実 験

4.1 実 験 条 件

提案手法による音声分離性能を検証するため，Voice Conversion Challenge (VCC) 2018 音声データベース [14] を用いた話者依存の分離実験と WSJ0 音声データベース [15] を用いた任意話者の分離実験を行った．比較対象は ILRMA [4], IDLMA [5], MVAE [7], PoE に基づく fMVAE_o と fMVAE_c とし，評価規準として source-to-distortion ratio (SDR), source-to-interferences ratio (SIR) と sources-to-artifact ratio (SAR) [16] を用いた．

話者依存の実験では男性話者 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によりシミュレートした 2 種類のインパルス応答（残響時間 (RT_{60}) はそれぞれ 78 ms 及び 351 ms），及び，RWCP データベース [17] に収録された 2 種類の実測インパルス応答（ANE(RT_{60}=173 ms) と E2A（RT_{60}=225 ms））を用いて作成した．それぞれの混合条件ごとに 4 パターンの話者の組み合わせ（SF1+SM1, SM1+SM2, SM2+SF2, SF1+SF2）の混合信号を計 40 文作成した．一方，任意話者の実験ではフォルダ `si_tr_s` 内の 101 話者のデータを学習データとし，フォルダ `si_dt_05` と `si_et_05` 内の 18 話者のデータを評価用データの作成に用いた．残響時間がそれぞれ 78 ms と 351 ms のインパルス応答を用いて各混合条件について混合信号を 100 文作成した．

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 の下で短時間 Fourier 変換を行い，観測信号 $\boldsymbol{x}(f, n)$ を算出した．図 1 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時のエンコーダとデコーダネットワーク構造も図 1 に示した構造を用いた．ILRMA の基底数は 2 とした．ILRMA と IDLMA1 においては 100 回反復更新を行った．また，MVAE, fMVAE_c と fMVAE_o 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．

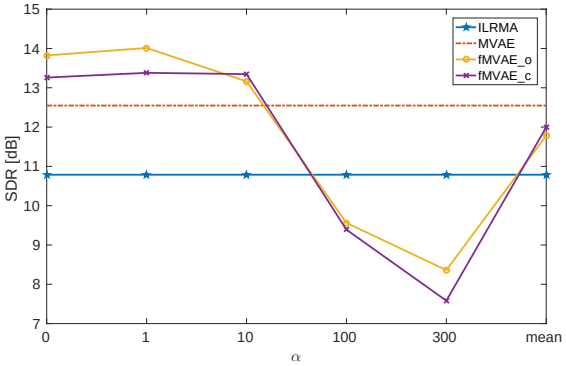

図 2: 話者依存条件での α の設定値による SDR 平均値 [dB] の違い．

表 1: 話者依存条件での $\alpha = 1$ のときの SDR，SIR と SAR の平均値 [dB].

method	RT_{60} = 78 [ms] (simu)			RT_{60} = 351 [ms] (simu)		
	SDR	SIR	SAR	SDR	SIR	SAR
ILRMA	18.33	25.42	20.47	5.96	13.78	7.91
IDLMA	9.31	13.31	12.72	4.53	10.46	7.45
MVAE	20.60	26.88	23.25	**6.98**	**15.48**	**8.63**
fMVAE_o	22.76	30.36	24.84	6.76	15.45	7.89
fMVAE_c	**23.35**	**31.03**	**25.25**	6.21	14.63	7.63
method	RT_{60} = 173 [ms] (ANE)			RT_{60} = 225 [ms] (E2A)		
	SDR	SIR	SAR	SDR	SIR	SAR
ILRMA	14.16	20.68	16.85	4.70	13.16	6.42
IDLMA	7.68	10.97	12.25	4.07	9.99	6.49
MVAE	16.44	22.20	19.08	**6.17**	15.97	**7.62**
fMVAE_o	**20.53**	**27.64**	**22.43**	6.00	**15.98**	7.25
fMVAE_c	19.26	26.37	21.33	4.70	13.91	6.49

4.2 実 験 結 果

図 2 と図 3 は話者依存と任意話者の分離実験において異なる α で得られた SDR 平均値を示す．話者依存の実験では $\alpha = \{0, 1, 10, 100, 300, \text{mean}\}$，任意話者の実験では更に $\alpha = \{500, 700, 1000\}$ の設定で実験を行った．ここで，“mean”

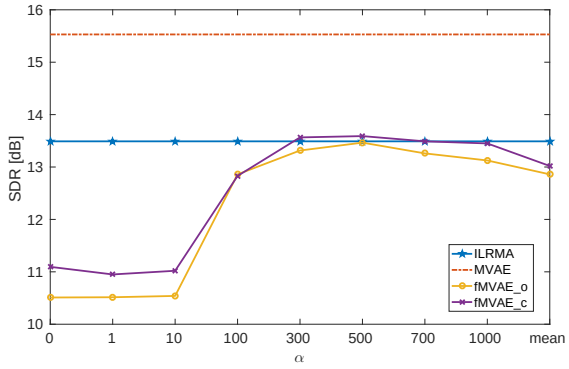

図 3: 任意話者条件での α の設定値による SDR 平均値 [dB] の違い．

表 2: 任意話者条件での $\alpha = 500$ のときの SDR，SIR と SAR の平均値 [dB]．

method	RT_{60} = 78 ms (simu)			RT_{60} = 351 ms (simu)		
	SDR	SIR	SAR	SDR	SIR	SAR
ILRMA	19.66	27.12	22.96	7.32	17.98	8.51
MVAE	**23.86**	**31.46**	**27.97**	**7.57**	18.31	**8.99**
fMVAE_o	19.44	26.55	23.29	7.49	**18.51**	8.70
fMVAE_c	19.74	27.14	23.53	7.44	18.48	8.62

は $\alpha = \frac{1}{K}\sum_k^K \sigma_\phi^2(k; \boldsymbol{S}, c)$ のようなデータ依存のパラメータ設定を表す．話者依存の実験において $\alpha = 1$ の時に最高分離性能が得られたことから，話者が学習データに含まれる条件においては学習されたエンコーダ分布の出力を直接用いたアルゴリズムでも $\boldsymbol{z}$ を正確に推定できることを確認した．一方，話者が未知の条件では，エンコーダ分布の出力を直接用いた場合の分離性能には劣化が見られた．しかし，$\boldsymbol{z}$ の事前確率を考慮した場合において性能に改善が見られたことから，提案法の有効性を確認できた．また，話者依存の実験においては fMVAE_c と fMVAE_o の分離性能に優劣が見られなかったが，任意話者の実験においては fMVAE_c の曲線が fMVAE_o の曲線を全体的に上回ったことから，クラス識別器の出力の連続値を c の推論結果とした方が未知話者のデータに有効であることが分かった．表 1 と表 2 に比較対象手法の話者依存と任意話者の分離実験における結果を示す．MVAE 法は両方の実験において従来の ILRMA と IDLMA を凌駕した分離性能を示した．一方，FastMVAE 法は話者依存の実験では MVAE 法より一層の性能向上が得られたが，任意話者の実験では分離性能の劣化が見られ，ILRMA と同等な分離性能であった．

5. おわりに

本稿では，多数の話者が含まれるデータでネットワークを学習させることにより話者に汎用性のある CVAE 音源モデルの構築が可能であることと，それを用いた MVAE 法と FastMVAE 法は任意話者に対しても高い分離性能を実現できることを実験により確認した．PoE の枠組により潜在空間変数の確率モデルを統合し，CVAE 音源モデルの学習時に用いられた規準と同一になるような規準に基づく潜在空間変数の更新式を提案し，その有効性を任意話者の分離実験により確認した．

謝辞 本研究は JSPS 科研費 17H01763 と 18J20059，JST CREST JPMJCR19A3 及び SECOM 科学技術振興財団の助成を受けて行われた．

文　献

[1] 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.

[2] 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.

[3] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le. Roux and K. Kashino, “Statistical model of speech signals based on composite autoregressive system with application to blind source separation,” in Proc. LVA/ICA, pp. 245–253, 2010.

[4] 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. 1622–1637, 2016.

[5] 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, pp. 1571–1575, 2018

[6] H. Kameoka, L. Li, S. Inoue and S. Makino, “Semi-blind source separation with multichannel variational autoencoder,” arXiv: 1808.00892, 2018.

[7] H. Kameoka, L. Li, S. Inoue and S. Makino, “Supervised determined source separation with multichannel variational autoencoder,” Neural computation, vol. 31, no. 9, pp. 1891–1914, 2019.

[8] L. Li, H. Kameoka and S. Makino, “Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier,” in Proc. ICASSP, pp. 546–550, 2019.

[9] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” Neural computation, vol. 14, no. 8, pp. 1771–1800, 2002.

[10] T. Kim, T. Eltoft and T.-W. Lee, “Independent vector analysis: An extension of ICA to multivariate components,” in Proc. ICA, pp. 165–172, 2006.

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

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

[13] H. Kameoka, T. Kaneko, K. Tanaka and N. Hojo, “ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder,” arXiv: 1808.05092, 2018.

[14] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F, Villavicencio, T. Kinnunen and Z. Ling, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods,” arXiv: 1804.04262, Apr. 2018.

[15] J. S. Garofolo, et al. CSR-I (WSJ0) Complete LDC93S6A. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[16] 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.

[17] S. Nakamura, K. Hiyane, F. Asano and T. Endo, “Sound scene data collection in real acoustical environments,” J. Acoust. Soc. Jpn. (E), vol. 20, no. 3, pp. 225–231, 1999.