ChimeraACVAEによる高速多チャンネル変分自己符号化器法*

◎李莉[1,2]，亀岡弘和[1]，牧野昭二[3]
[1] NTT コミュニケーション科学基礎研究所，[2] 名古屋大学，[3] 筑波大学

1 はじめに

本稿では，優決定条件における多チャンネル音源分離問題を扱う．音源の独立性を最大化するように分離フィルタを推定するブラインド音源分離（Blind Source Separation: BSS）手法は優決定音源分離に有効であり，数多く提案されている．例えば，独立低ランク行列分析（Independent Low-Rank Matrix Analysis: ILRMA) [1] 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手法である．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 ILRMA 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，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の同時解決を可能にしている．この手法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．我々が提案した多チャンネル変分自己符号化器法（Multichannel Variational Autoencoder: MVAE）法 [2] 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一方で，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．

そこで以前我々は，MVAE 法の計算コストの削減を目的とし，FastMVAE 法 [3] と呼ぶ高速アルゴリズムを提案した．FastMVAE 法は，クラス識別器つき VAE（Auxiliary Classifier VAE: ACVAE）を用いて音源スペクトログラムの生成モデルであるデコーダと共に，音源クラスの分布と潜在変数の事後分布を近似する識別器分布とエンコーダ分布を学習することで，学習で得られた識別器とエンコーダの順伝播により事後分布が最大となるようなデコーダ入力値を近似する手法である．しかし，FastMVAE 法では音源分離アルゴリズムを高速化できた一方で，未知話者や長い残響の場合など，テスト時において学習時と条件が一致しない場合に分離性能が低下する傾向があった．本稿では，さらにこの問題を解決するため，知識蒸留を用いた新たな ACVAE 音源モデルの学習法とモデル構造を提案する．提案法で学習された音源モデルは従来よりコンパクトな構造かつ高い汎用性を持ち，未知話者でも高速かつ高精度な音源分離を実現できること示す．

2 MVAE 法を用いた音源分離

2.1 多チャンネル音源分離問題の定式化

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

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

$$\mathbf{s}(f,n) = [s_1(f,n), \ldots, s_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I \quad (2)$$

とする．ただし，ここでは $I = J$ の優決定条件を考える．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．$I = J$ の条件においては音源信号ベクトル $\mathbf{s}(f,n)$ と観測信号ベクトル $\mathbf{x}(f,n)$ の間の関係式として瞬時分離系

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

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

を仮定することができる．ここで，$\mathbf{W}^{\mathsf{H}}(f)$ 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

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

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 $\mathbf{s}(f,n)$ は

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

に従う．ここで，$\mathbf{V}(f,n)$ 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 $\mathbf{x}$ は

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

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

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

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．ILRMA や MVAE 法がその例である．

---

*Faster multichannel variational autoencoder method with ChimeraACVAE. by LI, Li (NTT Communication Science Laboratories, Nagoya University), KAMEOKA, Hirokazu (NTT Communication Science Laboratories), MAKINO, Shoji (University of Tsukuba). Shoji Makino is also currently a professor at Waseda University.

2.2 MVAE 法

MVAE 法では，音源クラスラベルを補助入力とした CVAE のデコーダ分布を各音源の複素スペクトログラムの生成モデルとして用いる．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 $\mathbf{c}$ とする．CVAE はエンコーダ分布 $q^*_\phi(\mathbf{z}|\boldsymbol{S},\mathbf{c})$ とデコーダ分布 $p^*_\theta(\boldsymbol{S}|\mathbf{z},\mathbf{c})$ が無矛盾になるように，すなわち，$q^*_\phi(\mathbf{z}|\boldsymbol{S},\mathbf{c})$ と $p^*_\theta(\boldsymbol{S}|\mathbf{z},\mathbf{c})$ から導かれる事後分布 $p^*_\theta(\mathbf{z}|\boldsymbol{S},\mathbf{c}) \propto p^*_\theta(\boldsymbol{S}|\mathbf{z},\mathbf{c})p(\mathbf{z})$ 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

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

$$v(f,n) = g \cdot \sigma^{*2}_\theta(f,n;\mathbf{z},\mathbf{c}) \quad (10)$$

と置く．ただし，分散 $\sigma^{*2}_\theta(f,n;\mathbf{z},\mathbf{c})$ 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 $q^*_\phi(\mathbf{z}|\boldsymbol{S},\mathbf{c})$ は通常の CVAE と同様に，標準正規分布

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

と仮定する．ここで，$\boldsymbol{\mu}^*_\phi(\boldsymbol{S},\mathbf{c})$, $\boldsymbol{\sigma}^{*2}_\phi(\boldsymbol{S},\mathbf{c})$ 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 $\{\boldsymbol{S}_m, \mathbf{c}_m\}_{m=1}^M$ を用いて

$$\mathcal{J} = \mathbb{E}_{\boldsymbol{S},\mathbf{c}}\big[\mathbb{E}_{\mathbf{z}\sim q^*_\phi(\mathbf{z}|\boldsymbol{S},\mathbf{c})}[\log p^*_\theta(\boldsymbol{S}|\mathbf{z},\mathbf{c})] - \mathrm{KL}[q^*_\phi(\mathbf{z}|\boldsymbol{S},\mathbf{c})||p(\mathbf{z})]\big] \quad (12)$$

が最大となるように学習される．$\mathbb{E}_{\boldsymbol{S},\mathbf{c}}[\cdot]$ は学習サンプルによる標本平均を表し，$\mathrm{KL}[\cdot||\cdot]$ は Kullback-Leibler (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 $p^*_\theta(\boldsymbol{S}|\mathbf{z},\mathbf{c},g)$ 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，$\mathbf{c}$ は音源クラスのカテゴリー別の特徴を調整する役割，$\mathbf{z}$ はクラス内の連続的な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，$\mathbf{z}_j$, $\mathbf{c}_j$, 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 $\Psi = \{\mathbf{z}_j, \mathbf{c}_j\}_j$, 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 と同様に反復射影法 (Iterative Projection: IP) [4] を用いることができる．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

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\mathbf{w}_j^{\mathsf{H}}(f)\mathbf{x}(f,n)|^2}{\sigma^2_\theta(f,n;\mathbf{z}_j,\mathbf{c}_j)} \quad (13)$$

により行うことができる．ただし，式 (13) 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. $\mathcal{W}$ を単位行列に初期化し，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．

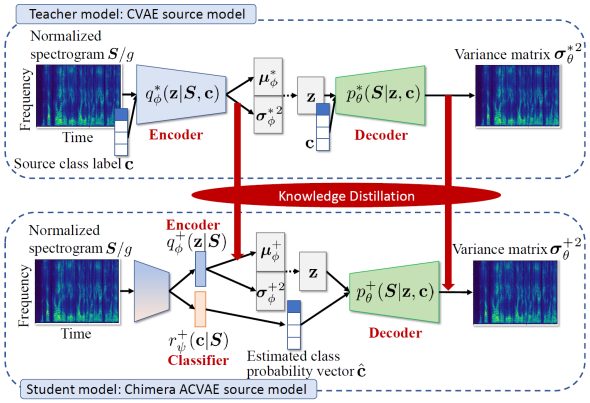

Fig. 1 ChimeraACVAE の学習概念図.

(a) IP 法により $\{\mathbf{w}_j(f)\}_{j,f}$ を更新
(b) 誤差逆伝播法により $\Psi_j = \{\mathbf{z}_j, \mathbf{c}_j\}$ を更新
(c) 式 (13) により g_j を更新

3 提案手法

3.1 基本アイディア

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，$p_\theta(\mathbf{z}_j, \mathbf{c}_j|\boldsymbol{S}_j)$ を最大にするパラメータ $\mathbf{z}_j, \mathbf{c}_j$ を誤差逆伝播法により更新するのに多大な計算コストを要する点に課題があった．FastMVAE 法 [3] では，事後分布 $p_\theta(\mathbf{z},\mathbf{c}|\boldsymbol{S})$ を $p_\theta(\mathbf{z}|\boldsymbol{S},\mathbf{c})p_\theta(\mathbf{c}|\boldsymbol{S})$ のように二つの条件付き分布の積に分解し，各分布を近似するよう分布 $q^*_\phi(\mathbf{z}|\boldsymbol{S},\mathbf{c})$, $r^*_\psi(\mathbf{c}|\boldsymbol{S})$ を NN により表現し，事前学習することで，MVAE 法における誤差逆伝播法によるパラメータ探索をそれぞれの NN の（高速計算可能な）順伝播で代替することが可能となった．しかし，FastMVAE 法におけるエンコーダ $q^*_\phi(\mathbf{z}_j|\boldsymbol{S}_j,\mathbf{c}_j)$ とクラス識別器 $r^*_\psi(\mathbf{c}_j|\boldsymbol{S}_j)$ の出力値は $\mathbf{z}_j, \mathbf{c}_j$ に関する対数尤度の最急上昇方向への更新値を近似したものでしかないため，音源分離精度に関しては FastMVAE 法は MVAE 法に及ばないことが実験的に確認された．

提案する FastMVAE2 法では，FastMVAE 法におけるエンコーダとクラス識別器を単一のマルチタスク NN として統合することでさらなる高速化を実現する．また，当該マルチタスク NN とデコーダを，それぞれの出力分布が，MVAE 法における事前学習で獲得したエンコーダとデコーダのそれぞれの出力分布とできるだけ近くなるように知識蒸留することで，各 NN に MVAE 法における $\mathbf{z}_j, \mathbf{c}_j$ のパラメータ更新に似た振る舞いを模倣させ，MVAE 法に近い分離精度を達成する．そこでまず，潜在変数 $\mathbf{z}$ と音源クラス $\mathbf{c}$ が条件付き独立であることを仮定する．これは，所与のスペクトログラム $\boldsymbol{S}$ が与えられた下で，話者情報 $\mathbf{c}$ と発話内容に関する情報 $\mathbf{z}$ が独立であると仮定することに相当する．つまり，事後確率 $p_\theta(\mathbf{z},\mathbf{c}|\boldsymbol{S})$ を $p_\theta(\mathbf{z}|\boldsymbol{S})p_\theta(\mathbf{c}|\boldsymbol{S})$ と表せると仮定する点が従来と異なる．この二つの条件付き分布の近似分布が得られれば，FastMVAE 法と同様，NN の順伝播によりパラメータ探索を高速に行うことができる．

3.2 ChimeraACVAE 音源モデル

ACVAE は，元々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 $\mathbf{c}$ 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 $\mathbf{c}$ との相互情報量 $I(\mathbf{c},\boldsymbol{S}|\mathbf{z})$ 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．$I(\mathbf{c},\boldsymbol{S}|\mathbf{z})$ を含めた規準を直接最適化することは容易ではないが，CVAE の学習と同様に変分下界

を導入し，その変分下界と $\mathcal{J}(\phi,\theta)$ を合わせた規準を上昇させることで，元となる規準を間接的に大きくすることができる．$I(\mathbf{c},\boldsymbol{S}|\mathbf{z})$ は $\log p(\mathbf{c}|\boldsymbol{S})$ の期待値と定数の和で与えられるが，$p(\mathbf{c}|\boldsymbol{S})$ を適当な補助分布 $r(\mathbf{c}|\boldsymbol{S})$ に置き換えたものが $I(\mathbf{c},\boldsymbol{S}|\mathbf{z})$ の下界となる．この補助分布 $r(\mathbf{c}|\boldsymbol{S})$ をパラメータ ψ の NN でモデル化することで，上記下界を規準として ψ を ϕ や θ とともに学習することができる．パラメータ ψ の NN で表される補助分布を $r_\psi(\mathbf{c}|\boldsymbol{S})$ と表し，クラス識別器と呼ぶ．

これに対し，提案する「[1]ChimeraACVAE」は ACVAE のエンコーダとクラス識別器を一体のマルチタスク NN として表したモデルである．つまり，$\mathbf{z}$ と $\mathbf{c}$ の分布 $q_\phi^+(\mathbf{z}|\boldsymbol{S}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi^+(\boldsymbol{S}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^{+2}(\boldsymbol{S})))$, $r_\psi^+(\mathbf{c}|\boldsymbol{S}) = \mathrm{Mult}(\mathbf{c}|\boldsymbol{\rho}_\psi^+(\boldsymbol{S}))$ をスペクトログラム $\boldsymbol{S}$ から同時推論するモデルとなる[2]．Fig. 1 の下の図に ChimeraACVAE の概念図を示す．ChimeraACVAE は潜在変数 $\mathbf{z}$ を入力スペクトログラムのみから抽出する構造になっているため，音源分離時において $\mathbf{z}$ の推論がクラスラベル $\mathbf{c}$ の推定誤差の影響を受けない利点がある．また，従来の ACVAE モデルに比べてコンパクトなネットワーク構造で記述できるため，より高速な推論が可能となることが期待される．

ChimeraACVAE の学習規準，すなわち NN パラメータ θ,ϕ,ψ に関して最大化すべき目的関数は，CVAE の学習規準

$$\mathcal{J} = \mathbb{E}_{\boldsymbol{S},\mathbf{c}}\big[\mathbb{E}_{\mathbf{z}\sim q_\phi^+(\mathbf{z}|\boldsymbol{S})}[\log p_\theta^+(\boldsymbol{S}|\mathbf{z},\mathbf{c})] - \mathrm{KL}[q_\phi^+(\mathbf{z}|\boldsymbol{S})||p(\mathbf{z})]\big] \quad (14)$$

および，相互情報量

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{S}',\mathbf{z}\sim q_\phi^+(\mathbf{z}|\boldsymbol{S}'),\mathbf{c},\boldsymbol{S}\sim p_\theta^+(\boldsymbol{S}|\mathbf{z},\mathbf{c})}\big[\log r_\psi^+(\mathbf{c}|\boldsymbol{S})\big] \quad (15)$$

の和となる．また，ラベル付き学習サンプル $\{\boldsymbol{S}_m,\mathbf{c}_m\}_m^M$ 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 $\mathbf{c}_m$ の負の交差エントロピー

$$\mathcal{I} = \mathbb{E}_{\boldsymbol{S},\mathbf{c}}[\log r_\psi^+(\boldsymbol{c}|\boldsymbol{S})] \quad (16)$$

も学習規準に含めることができる．ここまではモデル構造を除けば従来の ACVAE と同様である．しかし，以上の学習規準により学習された ACVAE は，テスト条件と学習条件が一致する場合高精度な推論が可能となるが，一致しない場合に推定される潜在変数が仮定した分布から逸脱する傾向があり，モデルの汎化能力は十分ではなかった．そこでモデルの汎化能力を向上させるため，ChimeraACVAE の学習においては上記の学習規準に加え更に以下の規準と知識蒸留を用いる．

式 (14), (15) は各学習サンプルの話者ラベル $\mathbf{c}$ を用いて定義されるが，音源分離時においては $\mathbf{c}$ は未知のため，各分離信号からクラス識別器 $r_\psi^+(\mathbf{c}|\boldsymbol{S})$ で推論されたものを用いてデコーダ出力が算出される流れとなる．音源分離時におけるこのプロセスを学習においても想定するなら，話者ラベルの代わりにクラス識別器 $r_\psi^+(\mathbf{c}|\boldsymbol{S})$ からサンプリングされた $\mathbf{c}$ を用いて定義される規準を考えることもできる．この場合の再構築規準（式 (14) の第一項に相当）とクラス識別規準はそれぞれ

$$\mathcal{J}' = \mathbb{E}_{\boldsymbol{S},\mathbf{z}\sim q_\phi^+(\mathbf{z}|\boldsymbol{S}),\mathbf{c}\sim r_\psi^+(\mathbf{c}|\boldsymbol{S})}\big[\log p_\theta^+(\boldsymbol{S}|\mathbf{z},\mathbf{c})\big] \quad (17)$$

$$\mathcal{L}' = \mathbb{E}_{\boldsymbol{S}',\mathbf{z}\sim q_\phi^+(\mathbf{z}|\boldsymbol{S}'),\mathbf{c}\sim r_\psi^+(\mathbf{c}|\boldsymbol{S}'),\boldsymbol{S}\sim p_\theta^+(\boldsymbol{S}|\mathbf{z},\mathbf{c})}\big[\log r_\psi^+(\mathbf{c}|\boldsymbol{S})\big] \quad (18)$$

[1]Chimera は異種の頭を一体にもつギリシャ神話の怪獣である．
[2]$\mathrm{Mult}(\mathbf{c}|\boldsymbol{\rho}) \propto \prod_i \rho_i{}^{c_i}$ は多項分布である．

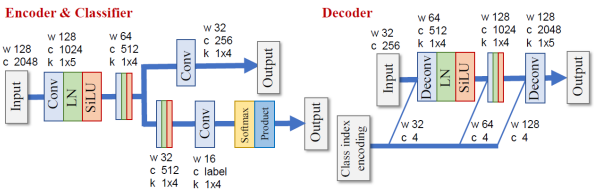

Fig. 2　ChimeraACVAE のネットワーク構造．

となるが，$r_\psi^+(\mathbf{c}|\boldsymbol{S})$ は多項分布のため変数変換トリックを単純適用することができない．そこで，これらの代わりに，$\mathbb{E}_{\mathbf{c}\sim r_\psi^+(\mathbf{c}|\boldsymbol{S})}[\cdot]$ の計算を $\mathbf{c}$ への $\hat{\mathbf{c}} = \mathbb{E}_{\mathbf{c}\sim r_\psi^+(\mathbf{c}|\boldsymbol{S})}[\mathbf{c}] = \boldsymbol{\rho}_\psi^+(\boldsymbol{S})$ の代入操作に置き換えた

$$\mathcal{J}' = \mathbb{E}_{\boldsymbol{S},\mathbf{z}\sim q_\phi^+(\mathbf{z}|\boldsymbol{S})}\big[\log p_\theta^+(\boldsymbol{S}|\mathbf{z},\hat{\mathbf{c}})\big] \quad (19)$$

$$\mathcal{L}' = \mathbb{E}_{\boldsymbol{S}',\mathbf{z}\sim q_\phi^+(\mathbf{z}|\boldsymbol{S}'),\boldsymbol{S}\sim p_\theta^+(\boldsymbol{S}|\mathbf{z},\hat{\mathbf{c}})}\big[\log r_\psi^+(\hat{\mathbf{c}}|\boldsymbol{S})\big] \quad (20)$$

を学習規準に含めることを提案する．また，音源分離時においてもこれに合わせて $\hat{\mathbf{c}}$ を各分離信号からの $\mathbf{c}$ の推論値とすることとした．

知識蒸留（Knowledge Distillation: KD）は事前に大量のデータで学習した大きな NN を教師モデルとし，その知識を軽量または別の NN 構造を持つ生徒モデルに継承させるための方法論であり，汎化能力の高い生徒モデルが得られることが知られている．ここで，未知話者に対しても高い分離精度を実現できる CVAE モデルを教師モデルとし，CVAE で学習した潜在変数の分布 $q_\phi^*(\mathbf{z}|\boldsymbol{S},\mathbf{c})$ とスペクトログラムの生成モデル $p_\theta^*(\boldsymbol{S}|\mathbf{z},\mathbf{c})$ の知識を生徒モデルである ChimeraACVAE に継承させることを考える．具体的には，CVAE で推論した潜在変数の分布 $q_\phi^*(\mathbf{z}|\boldsymbol{S},\mathbf{c})$ と，デコーダで出力した分散 $\boldsymbol{\sigma}_\phi^{*2}$ を用いた正規分布 $\mathcal{N}(\mathbf{0},\mathrm{diag}(\boldsymbol{\sigma}_\theta^{*2}(\mathbf{z},\mathbf{c})))$ をそれぞれ生徒モデルの出力分布 $q_\phi^+(\mathbf{z}|\boldsymbol{S})$ と，デコーダ出力 $\boldsymbol{\sigma}_\phi^{+2}$ を用いた正規分布の事前分布とし，生徒モデルの出力が事前分布に近づくよう学習させる．ただし，教師と生徒モデルの分布の乖離度は KL ダイバージェンスを用いて測る．

$$\mathcal{K}_1 = \mathbb{E}_{\boldsymbol{S},\mathbf{c}}\big[\mathrm{KL}[\mathrm{q}_\phi^*(\mathbf{z}|\boldsymbol{S},\mathbf{c})\|\mathrm{q}_\phi^+(\mathbf{z}|\boldsymbol{S})]\big] \quad (21)$$

$$\mathcal{K}_2 = \mathbb{E}_{\boldsymbol{S},\mathbf{c},\mathbf{z}^*\sim q_\phi^*(\mathbf{z}|\boldsymbol{S},\mathbf{c}),\mathbf{z}^+\sim q_\phi^+(\mathbf{z}|\boldsymbol{S})}\big[\mathrm{KL}[\mathcal{N}(\mathbf{0},\mathrm{diag}(\boldsymbol{\sigma}_\theta^{*2}(\mathbf{z}^*,\mathbf{c})))\| \mathcal{N}(\mathbf{0},\mathrm{diag}(\boldsymbol{\sigma}_\theta^{+2}(\mathbf{z}^+,\mathbf{c})))]\big] \quad (22)$$

$$\mathcal{K}_3 = \mathbb{E}_{\boldsymbol{S},\mathbf{c},\mathbf{z}^*\sim q_\phi^*(\mathbf{z}|\boldsymbol{S},\mathbf{c}),\mathbf{z}^+\sim q_\phi^+(\mathbf{z}|\boldsymbol{S})}\big[\mathrm{KL}[\mathcal{N}(\mathbf{0},\mathrm{diag}(\boldsymbol{\sigma}_\theta^{*2}(\mathbf{z}^*,\mathbf{c})))\| \mathcal{N}(\mathbf{0},\mathrm{diag}(\boldsymbol{\sigma}_\theta^{+2}(\mathbf{z}^+,\hat{\mathbf{c}})))]\big] \quad (23)$$

以上より ChimeraACVAE の最大化すべき学習規準は

$$\mathcal{J} + \lambda_{\mathcal{L}}\mathcal{L} + \lambda_{\mathcal{I}}\mathcal{I} + \lambda_{\mathcal{J}'}\mathcal{J}' + \lambda_{\mathcal{L}'}\mathcal{L}' - \lambda_{\mathcal{K}_1}\mathcal{K}_1 - \lambda_{\mathcal{K}_2}\mathcal{K}_2 - \lambda_{\mathcal{K}_3}\mathcal{K}_3 \quad (24)$$

となる．ここで，$\lambda_\cdot$ は非負値であり，各規準の重み係数である．Fig. 1 に知識蒸留を用いた ChimeraACVAE の学習の概念図を示す．

Fig. 2 に ChimeraACVAE のネットワーク構造を示す．エンコーダと識別器の各層は畳み込み層，Layer Normalization (LN) と Sigmoid Linear Unit (SiLU) により構成され，デコーダの各層は逆畳み込み層，LN と SiLU により構成される．ここで，LN を用いることによって，学習と推論時における正規化の計算方法の不整合を回避できる．SiLU は CVAE 音源モデルに用いられた Gated Linear Unit (GLU) と同様に階層間に受け渡す情報をゲートにより制御するデータ

Table 1　話者依存と任意話者条件での平均 SDR, SIR, SAR [dB]

scenario	method	SDR	SIR	SAR
spk-dep	ILRMA	13.62	19.79	15.83
	MVAE [3]	**17.03**	**23.75**	**18.61**
	FastMVAE [3]	13.95	19.54	16.33
	FastMVAE2	15.44	21.57	17.53
spk-ind	ILRMA	14.43	20.98	17.45
	MVAE [3]	**17.58**	**25.13**	**19.26**
	FastMVAE [3]	14.41	21.21	17.35
	FastMVAE2	16.90	24.66	18.83

駆動の活性化関数であり，GLU のパラメータ数を半減することができる．上記のネットワーク構造を持つ ChimeraACVAE 音源モデルのパラメータ数は従来の ACVAE 音源モデルの約 42%まで削減することができた．

3.3　FastMVAE2 法：高速な推論アルゴリズム

ChimeraACVAE で学習したエンコーダとクラス識別器を用いることで，従来の MVAE 法における $p_\theta(\mathbf{z}_j, \mathbf{c}_j|\boldsymbol{S}_j)$ の最大化ステップを $q_\phi^+(\mathbf{z}_j|\boldsymbol{S}_j)$ と $r_\psi^+(\mathbf{c}_j|\boldsymbol{S}_j)$ の順伝播に置き換えることができる．よって，以下のアルゴリズムが得られる．これを FastMVAE2 法と呼ぶ．

1. 式 (24) を学習規準として θ，ϕ，ψ を学習する．
2. $\mathcal{W}$ を単位行列に初期化する．
3. 各 j について下記ステップを繰り返す．
   (a) IP 法により $\{\mathbf{w}_j(f)\}_{j,f}$ を更新
   (b) $\mathbf{z}_j$ を $\boldsymbol{\mu}_\phi^+(\boldsymbol{S}_j)$ に更新，$\mathbf{c}_j$ を $\boldsymbol{\rho}_\psi^+(\boldsymbol{S}_j)$ に更新
   (c) 式 (13) により g_j を更新

4　評価実験

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 [5] を用いた話者依存の分離実験と WSJ0 音声データベース [6] を用いた任意話者の分離実験を行った．比較対象は ILRMA[1],[3]MVAE 法 [2]，FastMVAE 法 [3] とし，評価規準として source-to-distortion ratio (SDR)，source-to-interference ratio (SIR) と sources-to-artifact ratio (SAR) [7] を用いた．[4]スペースの制限により学習と評価用データの詳細および MVAE 法と FastMVAE 法の実験設定を割愛する．すべての手法においては分離行列 $\mathbf{W}(f)$ を単位行列に初期化し，60 回更新を行った．ILRMA の基底数を 2 とした．Table 1 に実験結果を示す．いずれの条件においても，提案法が ILRMA と FastMVAE 法より高い分離性能を示し，MVAE 法との差を大幅に縮めた．

2 音源より多い音源数における各手法の分離性能および計算時間を評価するため，WSJ0 音声データベースから，18 話者の発話を利用して音源数が {2, 3, 6, 9, 12, 15, 18} の混合信号を作成した．インパルス応答は鏡像法により作成し，壁の反射係数を 0.2 とした．Fig. 3 にマイクと音源の配置を示す．各条件について混合信号を 10 文作成した．すべての処理は Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz と Tesla V100

[3]Code: https://github.com/lili-0805/MVAE
[4][3] を参照されたい．

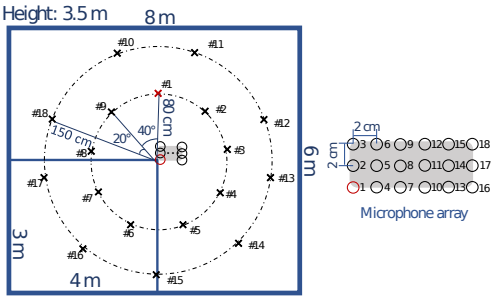

Fig. 3　マイクと音源の配置

Table 2　各音源数における平均 SDR 改善値 [dB]

method	# of sources and channels						
	2	3	6	9	12	15	18
ILRMA	20.79	26.96	15.66	12.07	12.03	9.25	7.94
MVAE	26.54	**29.08**	**19.45**	**19.71**	**19.49**	**18.03**	**16.44**
FastMVAE	15.68	11.50	11.45	14.69	12.83	12.97	11.90
FastMVAE2	**27.28**	25.50	14.56	15.21	14.75	13.82	13.08

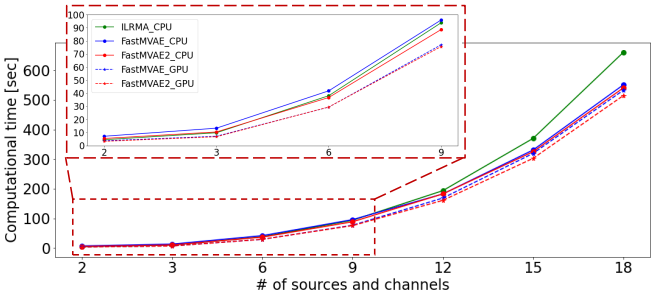

Fig. 4　各反復における計算時間 [sec]

GPU を用いて計算した．Table 2 に各条件における平均 SDR 改善値を示し，Fig. 4 に各手法の計算時間を示す．提案手法において性能改善が確認できる．また，提案手法は 3 音源以下の場合に ILRMA と同等の計算時間，3 音源以上の場合に ILRMA より短い計算時間で分離を実現できることを確認した．

5　おわりに

本稿では，優決定音源分離で高い分離性能が確認されている MVAE 法の高速アルゴリズムである FastMVAE 法の改良を行った．エンコーダとクラス識別器を一体化し，よりコンパクトな構造を持つ ChimeraACVAE 音源モデルを提案し，知識蒸留を用いた学習規準を提案した．様々な音源数条件における実験的評価により，ChimeraACVAE を用いた提案法 FastMVAE2 法は任意話者でも高速かつ高性能の分離ができることを確認した．

謝辞　本研究は JST CREST JPMJCR19A3 の助成を受けて行われた．

参考文献

[1] D. Kitamura, *et al.*, *IEEE/ACM TASLP*, 2016.
[2] H. Kameoka, *et al.*, *Neural Computation*, 2019.
[3] L. Li, *et al.*, *IEEE Access*, 2020.
[4] N. Ono, *WASPAA*, 2011.
[5] J. Lorenzo-Trueba, *et al.*, *arXiv* , 2018.
[6] J. S. Garofolo, *et al.*, 1993.
[7] E. Vincent, *et al.*, *IEEE TASLP*, 2006.