識別的変分自己符号化器学習による特定話者モノラル音声分離*

☆村島允也[1]，亀岡弘和[2]，李莉[1]，関翔悟[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

本稿では特定話者のモノラル音声分離を扱う．複数話者の音声信号が混在する混合信号から各話者の音声信号を分離する音声分離技術は，音声認識の高精度化や音声通信の高品質化に直結する重要技術である．近年 Deep Clustering（DC）[1] をはじめとする Deep Neural Network (DNN) を用いた識別的アプローチによるモノラル音声分離手法が，その有効性により注目を集めている [2–4]．識別的アプローチの手法としては，混合信号から各話者に対応する時間周波数マスクを直接予測する DNN，または時間周波数マスクを得る手がかりとなる特徴量を出力する DNN を学習する方法などが検討されている．また，最近は各話者の分離波形を直接予測する DNN を学習する方法も検討されている．これらの手法は特定条件において高い精度の分離を達成できることが示されているが，残響などにより学習条件とテスト条件の間にミスマッチがある場合に分離精度が低下しうる点に課題がある．

一方，非負行列分解 (Non-negative Matrix Factorization: NMF) [5] の教師あり学習に基づく音声分離法である教師あり NMF (Supervised NMF: SNMF) 法 [6] をはじめとする生成的アプローチは，観測信号のモデル化に基づくため，学習条件とテスト条件のミスマッチに柔軟に対応できる点において有利になりうる．これは例えば，ミスマッチの原因となる過程を模したモデルを観測信号の生成モデルに明示的に組み込み，テスト時に観測信号に応じてすべての未知パラメータを最適推定する方法により実現できる．SNMF 法は，各時刻の観測信号のスペクトルが有限個の基底スペクトルの非負結合で近似できるという仮定に基づいており，この近似は観測信号のスペクトログラムを 2 つの非負行列（基底行列と係数行列）の積で近似することに相当する．SNMF 法 [6] は，各話者の音声サンプルを用いて基底スペクトルを事前学習し，それらを観測された混合信号にフィッティングすることで各話者に対応するスペクトログラムを推定し，Wiener フィルタにより分離を行う方法である．しかしこの方法では，基底スペクトルの学習規準が分離時の目的関数と一致しておらず，テスト時の分離信号が最適となるような規準にはなっていない．識別的 NMF (Discriminative NMF: DNMF) [7] 法はこの不整合を解決することを目的として提案された手法である．DNMF 法は，テスト時と同じ手順の音声分離プロセスを考え，分離信号（Wiener フィルタの出力）が直接最適になるような規準により基底スペクトルを学習する方式である．しかし，DNMF 法を含む NMF に基づく手法の分離性能は，音声のように低ランク行列で正確に近似することが難しいスペクトログラムを持つ音源に対しては限定的であった．

近年，音源のスペクトログラムをより精緻にモデル化する試みとして，DNN を導入した生成的アプローチに基づく手法が提案されている [8–16]．多チャンネル信号を対象とした手法としては，条件付き変分自己符号化器 (Conditional Variational Autoencoder: CVAE) [17] を音声スペクトログラムのモデル化に利用した多チャンネル VAE (Multichannel VAE: MVAE) 法 [10–12] が提案され，NMF 型の音源スペクトログラムモデルに基づく独立低ランク行列分析法 [18, 19] の性能を大きく凌駕することに成功している．このことは，CVAE が音源モデルとして高い表現能力および当該音と干渉音を正しく弁別するための弁別能力を有することを示している．同様の動機で，VAE と NMF をそれぞれ音声と雑音のスペクトログラムのモデル化に利用したモノラル音声強調法 (VAE-NMF 法) [13,14] およびその多チャンネル拡張 [15,16] も提案されている．

本稿では，CVAE を音源モデルとして利用したモノラル音声分離 (VAE-based Speech Separation: VASS) 法 とともに，DNMF 法の枠組みを拡張し CVAE 音源モデルを識別的規準により学習する識別的 VASS (Discriminative VASS: DVASS) 法を提案する．2 話者混合信号を用いた音声分離実験により SNMF 法，VASS 法，DVASS 法の比較を行い，提案法の有効性を示す．

2 モノラル音声分離の従来法

2.1 音声分離問題の定式化

観測信号中に J 話者の音声信号が混在する場合を考え，観測信号と話者 j の音声信号の複素スペクトログラムをそれぞれ $\mathbf{Y} = \{y(f,n)\}_{f,n} \in \mathbb{C}^{F\times N}$ と $\mathbf{S}_j = \{s_j(f,n)\}_{f,n} \in \mathbb{C}^{F\times N}$ とする．ただし，f, n はそれぞれ周波数と時刻のインデックスである．ここで，$\mathbf{S}_j$ の各要素 $s_j(f,n)$ を

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n) \mid 0, v_j(f,n)) \tag{1}$$

のように平均が 0，分散が $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$ の複素正規分布に従う独立な確率変数と仮定する．式 (1) に従う $s_j(f,n)$ を局所 Gauss 音源モデル (Local Gaussian Model: LGM) と呼び，多くの音声分離手法において広く用いられている．ここで，$\mathbf{S}_1, \ldots, \mathbf{S}_J$ が互いに独立と仮定すると，$\mathbf{Y} = \sum_j \mathbf{S}_j$ という関係より，$\mathbf{Y}$ の各要素 $y(f,n)$ は

$$y(f,n) \sim \mathcal{N}_{\mathbb{C}}(y(f,n) \mid 0, v(f,n)) \tag{2}$$

に従う．ただし，$v(f,n) = \sum_j v_j(f,n)$ である．$\mathbf{V}_j = \{v_j(f,n)\}_{f,n}$ とすると，$\mathbf{Y}$ が与えられた下での $\mathbf{V} = \{\mathbf{V}_1, \ldots, \mathbf{V}_J\}$ の負の対数尤度 $-\log p(\mathbf{Y} \mid \mathbf{V})$ は，定数項を除いて $\tilde{y}(f,n) = |y(f,n)|^2$ と $v(f,n)$ の板倉斎藤擬距離

$$\mathcal{D}_{\mathrm{IS}}(\hat{\mathbf{Y}}|\mathbf{V}) = \sum_{f,n} \left(\frac{\tilde{y}(f,n)}{v(f,n)} - \log \frac{\tilde{y}(f,n)}{v(f,n)} - 1 \right) \tag{3}$$

と等しい．また，$\mathbf{Y}$ と $\mathbf{S}_1, \ldots, \mathbf{S}_J$ の結合 Gauss 性より，$\mathbf{Y}$ と $\mathbf{V}_1, \ldots, \mathbf{V}_J$ が与えられた下での $\mathbf{S}_j$ の最小

*Single-channel Multi-speaker Separation via Discriminative Training of Variational Autoencoder by Naoya Murashima (University of Tsukuba), Hirokazu Kameoka (NTT), Li Li (University of Tsukuba), Shogo Seki (NTT), Shoji Makino (University of Tsukuba).

平均二乗誤差推定量 $\mathbb{E}[\mathbf{S}_j|\mathbf{Y}]$ は

$$\mathbb{E}[\mathbf{S}_j|\mathbf{Y}] = \frac{\mathbf{V}_j}{\sum_{j'}\mathbf{V}_{j'}} \odot \mathbf{Y} \tag{4}$$

で与えられる．ただし，式 (4) の係数は Wiener フィルタを表し，$\div$ と $\odot$ は行列要素ごとの除算，乗算を表す．よって，$\mathbf{V}_1, \ldots, \mathbf{V}_J$ を推定できれば，式 (4) より各音声信号を推定できることを意味する．以上より，モノラル音声分離の問題は，$\mathbf{V}_1, \ldots, \mathbf{V}_J$ に関して何らかの制約や仮定を設けた上で，式 (3) を規準として $\mathbf{V}_1, \ldots, \mathbf{V}_J$ を最適推定する問題に帰着する．

2.2 SNMF 法 [6]

SNMF 法では $\mathbf{V}_j$ を非負行列積 $\mathbf{W}_j\mathbf{H}_j$，すなわち $v_j(f,n) = \sum_k w_{j,k}(f)h_{j,k}(n)$ で表し，事前に各話者の音声サンプルを用いて $\mathbf{W}_j$ を学習する．スペクトログラムを非負行列積（低ランク行列）で表すことは各時刻のスペクトルを有限個の基底スペクトルの非負結合で表すことに相当しており，$\mathbf{W}_j$ の事前学習により各話者固有の基底スペクトルが得られることとなる．テスト時は全話者の基底スペクトルを連結したもの $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_J]$ を固定した上で，観測混合信号のスペクトログラム $\mathbf{Y}$ に $\mathbf{WH}$ が当てはまるよう $\mathbf{H} = [\mathbf{H}_1^\intercal, \ldots, \mathbf{H}_J^\intercal]^\intercal$ を推定し，$\mathbf{V}_1, \ldots, \mathbf{V}_J$ を得ることで Wiener フィルタにより分離を行える．

$\mathbf{W}_j$ の事前学習は，話者ごとの全学習サンプルを時間方向に連結したパワースペクトログラム $\tilde{\mathbf{S}}_j'$ と $\mathbf{V}_j = \mathbf{W}_j\mathbf{H}_j$ の誤差を規準とした最適化問題

$$\{\hat{\mathbf{W}}_j, \hat{\mathbf{H}}_j\} = \underset{\mathbf{W}_j, \mathbf{H}_j}{\operatorname{argmin}} \mathcal{D}(\tilde{\mathbf{S}}_j'|\mathbf{W}_j\mathbf{H}_j) \tag{5}$$

として定式化できる．ただし，$\mathcal{D}$ は行列間の誤差を測る関数を表し，板倉斎藤擬距離などが用いられる．テスト時は観測混合信号のパワースペクトログラム $\tilde{\mathbf{Y}}$ が与えられた下で，事前学習した基底行列 $\hat{\mathbf{W}} = [\hat{\mathbf{W}}_1, \ldots, \hat{\mathbf{W}}_J]$ を固定し，

$$\hat{\mathbf{H}} = \underset{\mathbf{H}}{\operatorname{argmin}} \mathcal{D}(\tilde{\mathbf{Y}}|\hat{\mathbf{W}}\mathbf{H}) \tag{6}$$

の解を探索することで，観測信号に含まれる各話者のパワースペクトログラムの成分 $\hat{\mathbf{W}}_j\hat{\mathbf{H}}_j$ を推定する．各話者に対応する分離信号の複素スペクトログラムは以下により得られる．

$$\hat{\mathbf{S}}_j = \frac{\hat{\mathbf{W}}_j\hat{\mathbf{H}}_j}{\sum_{j'}\hat{\mathbf{W}}_{j'}\hat{\mathbf{H}}_{j'}} \odot \mathbf{Y} \tag{7}$$

2.3 DNMF 法 [7]

テスト時において式 (7) により最終的な分離信号を得る場合，SNMF 法における基底スペクトルの学習規準は分離信号を直接的に最適にするような規準になっていない．DNMF 法は，学習においてもテスト時と同じ手順の音声分離プロセスを考え，Wiener フィルタの出力が最適となるように基底スペクトル学習を行うよう改良された手法である．

SNMF 法ではテスト時において式 (6) で $\tilde{\mathbf{Y}}$ から係数行列 $\hat{\mathbf{H}}$ を得る目的，式 (7) で Wiener フィルタを構成する目的でそれぞれ基底スペクトルが使用される．実はこれらのステップで使用される基底スペクトルは必ずしも同一である必要はなく，異なる変数として扱った上で目的に合わせた規準でそれぞれを別々に学習する方がテスト時において有利なはずである．そこで，各ステップにおける基底行列を $\mathbf{W}$, $\mathbf{B}$ と表し，それぞれを異なる変数として学習することを考える．各話者の音声サンプルを適当に混合した混合信号のパワースペクトログラム $\tilde{\mathbf{Y}}' = \{|y'(f,n)|^2\}_{f,n}$ を用いることで，テスト時と同じ音声分離プロセスをシミュレートしながら混合前の各音声信号を学習の目標信号とすることができる．すなわち，式 (5) で学習した基底行列 $\hat{\mathbf{W}}$ を用いて式 (6) により $\hat{\mathbf{H}}$ が求まれば，式 (7) の出力値が目標信号にできるだけ一致するように $\mathbf{B}$ を学習することができる．よって，学習に用いる混合信号と目標信号の振幅スペクトログラムをそれぞれ $|\mathbf{Y}'|, |\mathbf{S}_1'|, \ldots, |\mathbf{S}_J'|$（ただし $|\cdot|$ は行列要素ごとの絶対値を表す）とすると，

$$\hat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}} \sum_j \mathcal{D}\left(|\mathbf{S}_j'| \middle| \frac{\mathbf{B}_j\hat{\mathbf{H}}_j}{\mathbf{B}\hat{\mathbf{H}}} \odot |\mathbf{Y}'|\right) \tag{8}$$

が $\mathbf{B}$ の学習目標となる．テスト時は，以上により学習した $\hat{\mathbf{W}}$, $\hat{\mathbf{B}}$ を用いて式 (6) と式 (7) と同じプロセスにより分離信号を得ることができる．

SNMF 法も DNMF 法も，各音声信号のスペクトログラムが低ランク行列で表せることを想定した手法になっているが，この仮定は必ずしも正確ではなく，これがいずれの手法においても分離性能を限定的なものにしている．

3 提案法

3.1 VAE によるスペクトログラムモデル

行列積表現 $\mathbf{Wh}$ は，$\mathbf{h}$ を入力とした 1 層の線形全結合型 NN と見なせるので，NMF 型のスペクトログラムモデルの代わりに，その自然な拡張として，より高い表現能力をもつことが期待される DNN を用いることが考えられる．そこで LGM（式 (1)）において，分散 $v_j(f,n)$ を DNN の出力として表現するモデルを考えることができるが，これは後述のように VAE により記述できる．

前述の MVAE 法では，多チャンネル音源分離問題を対象とし，話者ラベルで条件付けした CVAE を用いた音源スペクトログラムモデルが導入されているが，本稿ではこの音源モデルをベースにしたモノラル音声分離法を提案する．

3.2 CVAE 音源モデル

CVAE はエンコーダとデコーダからなる自己符号化器型の NN モデルの一種で，エンコーダもデコーダも確率分布モデルで表現されている点，いずれも補助変数で条件付けされる点が特徴である．ある話者の音声の複素スペクトログラムを $\mathbf{S}$ とし，対応する話者ラベルの one-hot 表現を $\mathbf{c}$ とすると，CVAE のデコーダ分布を式 (1) の LGM と同形かつ $\mathbf{c}$ の条件付き分布の形

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

$$v(f,n) = \eta \cdot \sigma_\theta^2(f,n;\mathbf{z},\mathbf{c}) \tag{10}$$

にしたものを CVAE 音源モデルと呼ぶ．ただし，分散 $\sigma_\theta^2(f,n;\mathbf{z},\mathbf{c})$ は $\mathbf{z},\mathbf{c}$ を入力としたデコーダネットワークの出力 $\boldsymbol{\sigma}_\theta^2(\mathbf{z},\mathbf{c})$ の第 (f,n) 要素であり，$\mathbf{z}$ はエンコーダ分布から生成された潜在変数，η はパワースペクトログラムの総エネルギーを表すパラメータで

ある．一方，エンコーダ分布 $q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})$ は正規分布形

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

のものを考え，$\mathbf{z}$ の事前分布 $p(\mathbf{z})$ は標準正規分布とする．ただし，平均 $\boldsymbol{\mu}_\phi(\mathbf{S},\mathbf{c})$ および分散 $\boldsymbol{\sigma}^2_\phi(\mathbf{S},\mathbf{c})$ は $\mathbf{S},\mathbf{c}$ を入力としたエンコーダネットワークの出力である．ここで，CVAE の NN パラメータ θ, ϕ を，各学習サンプルの複素スペクトログラムと話者ラベルのペア $\{\mathbf{S}_m, \mathbf{c}_m\}_{m=1}^M$ を用いて，エンコーダ分布 $q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})$ とデコーダ分布 $p_\theta(\mathbf{S}|\mathbf{z},\mathbf{c})$ が矛盾しないように，すなわち，$q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})$ と $p_\theta(\mathbf{z}|\mathbf{S},\mathbf{c}) \propto p_\theta(\mathbf{S}|\mathbf{z},\mathbf{c})p(\mathbf{z})$ ができるだけ一致するように学習することで，デコーダ分布を話者ごとのスペクトログラムの分布に近づけることができる．$q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})$ と $p_\theta(\mathbf{z}|\mathbf{S},\mathbf{c})$ の Kullback-Leibler (KL) ダイバージェンスの期待値は，定数項を除き

$$\begin{aligned}\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\mathbf{S},\mathbf{c})\sim p_D(\mathbf{S},\mathbf{c})}\big[&\mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})||p(\mathbf{z})] \\ &- \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})}[\log p_\theta(\mathbf{S}|\mathbf{z},\mathbf{c})]\big]\end{aligned} \quad (12)$$

と等しくなるので，これを最小にすることが θ, ϕ の学習目標となる．ただし，$\mathbb{E}_{(\mathbf{S},\mathbf{c})\sim p_D(\mathbf{S},\mathbf{c})}[\cdot]$ は学習サンプルによる標本平均を表し，$\mathrm{KL}[\cdot||\cdot]$ は KL ダイバージェンスを表す．ところで実は式 (12) の第二項は，式 (10) のデコーダ分布が LGM と同形となっていることから，定数項を除いて $\hat{\mathbf{S}} = \{|s(f,n)|^2\}_{f,n}$ と $v(f,n)$ の間の板倉斎藤擬距離の期待値に等しい．

なお，CVAE 音源モデルにおいて，潜在変数 $\mathbf{z}$ が発話内容に相当するコンテキスト情報を表す量，デコーダの NN パラメータ θ がコンテキスト情報からスペクトログラムへの変換則を司る量と解釈できるため，それぞれ NMF 型の音源モデルにおける係数行列 $\mathbf{H}$ および基底行列 $\mathbf{W}$ に対応していると見なせる．

3.3 提案法①：VASS 法

VASS 法は，SNMF 法において NMF 型の音源モデルを CVAE 音源モデルに置き換えたものに相当し，SNMF 法と同様，音源モデルの事前学習ステップと，観測混合信号のスペクトログラムに対する音源モデルフィッティング，Wiener フィルタによる分離信号の推定，のテストステップからなる．CVAE 音源モデルの特長は同一パラメータで全話者のスペクトログラムを表現できる点にあり，式 (12) を規準に事前学習を行うことができる．事前学習で得られたパラメータを $\hat{\theta}$ とすると，テスト時の第 1 ステップとなる観測混合信号に対する音源モデルフィッティングは，観測信号の複素スペクトログラム $y(f,n) = \sum_j s_j(f,n)$ が

$$y(f,n) \sim \mathcal{N}_{\mathbb{C}}(y(f,n)|0, v(f,n)) \quad (13)$$

$$v(f,n) = \sum_j \underbrace{v_j(f,n)}_{\eta_j \sigma^2_{\hat{\theta}}(f,n;\mathbf{z}_j,\mathbf{c}_j)} \quad (14)$$

に従うことから，$\mathbf{Y} = \{y(f,n)\}_{f,n}$, $\mathbf{Z} = \{z_j\}_j$, $\mathbf{C} = \{\mathbf{c}_j\}_j$, $\boldsymbol{\eta} = \{\eta_j\}_j$ と置くと，$\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ を $\mathbf{Z}$, $\mathbf{C}$, $\boldsymbol{\eta}$ に関して最大化する最尤推定問題

$$\{\hat{\mathbf{Z}}, \hat{\mathbf{C}}, \hat{\boldsymbol{\eta}}\} = \underset{\mathbf{Z},\mathbf{C},\boldsymbol{\eta}}{\mathrm{argmax}} \log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta}) \quad (15)$$

として定式化できる．2.1 節で述べたように $-\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ は $|y(f,n)|^2$ と $v(f,n)$ の板倉斎藤擬距離と定数項を除いて等しいため，この問題は，観測混合信号のパワースペクトログラム $\tilde{\mathbf{Y}} = \{|y(f,n)|^2\}_{f,n}$ に対し，$\hat{\theta}$ を固定した上で $\mathcal{D}_{\mathrm{IS}}(\tilde{\mathbf{Y}}|\mathbf{V})$ （ただし，$\mathbf{V} = \{v(f,n)\}_{f,n}$）が最小となる $\mathbf{Z}$, $\mathbf{C}$, $\boldsymbol{\eta}$ を求める問題と等価である．$\hat{\mathbf{Z}}$, $\hat{\mathbf{C}}$, $\hat{\boldsymbol{\eta}}$ が求まれば観測信号に含まれる各話者のパワースペクトログラムの成分を推定することができるので，Wiener フィルタ

$$\mathbf{S}_j = \frac{\eta_j \boldsymbol{\sigma}^2_{\hat{\theta}}(\hat{\mathbf{z}}_j, \hat{\mathbf{c}}_j)}{\sum_{j'} \eta_{j'} \boldsymbol{\sigma}^2_{\hat{\theta}}(\hat{\mathbf{z}}_{j'}, \hat{\mathbf{c}}_{j'})} \odot \mathbf{Y} \quad (16)$$

により分離信号を得ることができる．

式 (15) の解法にはいくつかの方法が考えられる．1 つ目は，$\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ または $\mathcal{D}_{\mathrm{IS}}(\tilde{\mathbf{Y}}|\mathbf{V})$ を規準として，$\mathbf{Z}$, $\mathbf{C}$, $\boldsymbol{\eta}$ を単純に勾配法（$\mathbf{Z}$ と $\mathbf{C}$ については誤差逆伝播法）で最適化する方法である．2 つ目は，各話者に対応する複素スペクトログラム $s_j(f,n)$ を潜在変数とした期待値最大化 (Expectation Maximization: EM) 法である．EM 法では，$\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ を直接最大化する代わりに，補助関数 $\mathbb{E}_{\mathbf{S}\sim p(\mathbf{S}|\mathbf{Y},\mathbf{Z}',\mathbf{C}',\boldsymbol{\eta}')}[\log p(\mathbf{S}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})]$ を用いて E ステップと M ステップと呼ぶ更新を反復的に行うことで，$\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ を間接的に大きくすることができる．M ステップは，補助関数が増加するように $\mathbf{Z}$, $\mathbf{C}$, $\boldsymbol{\eta}$ を更新する処理となり，$\mathbf{Z}$ と $\mathbf{C}$ の更新については誤差逆伝播法により行うことができる．$\boldsymbol{\eta}$ については，$\mathbf{Z}$, $\mathbf{C}$ が固定のとき補助関数を最大にする更新則が解析的に得られる．一方 E ステップは，更新した $\mathbf{Z}$, $\mathbf{C}$, $\boldsymbol{\eta}$ を $\mathbf{Z}'$, $\mathbf{C}'$, $\boldsymbol{\eta}'$ に代入し，補助関数（の期待値計算）を更新する処理となる．$\log p(\mathbf{S}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ は $\sum_j \sum_{f,n} \log p(s_j(f,n)|0, v_j(f,n))$ のように j ごとの項に分解された形になるため，M ステップにおいて $(\mathbf{z}_1,\mathbf{c}_1,\boldsymbol{\eta}_1),\ldots,(\mathbf{z}_J,\mathbf{c}_J,\boldsymbol{\eta}_J)$ の更新を並列に行うことができる．3 つ目は，EM 法を包含する概念である補助関数法の原理に基づき，EM 法とは異なる補助関数を用いて $\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ を反復的に大きくする方法である．紙面の都合により，EM 法と補助関数法に基づく具体的なアルゴリズムの導出は省略するが，後述の実験では EM 法のアルゴリズムを用いた．

3.4 提案法②：DVASS 法

前述の VASS 法では SNMF 法と同様，CVAE 音源モデルのパラメータ θ の学習規準が分離信号（Wiener フィルタ出力）が最適となるような規準になっていない．そこで，DNMF 法のアイディアをヒントにし，この不整合を解消するよう VASS 法を改良する．

DNMF 法は，係数行列を得る目的の基底行列と Wiener フィルタを構成する目的の基底行列を別変数として扱う点がポイントだったが，同様の考え方で，式 (15) を求める目的の CVAE 音源モデルと Wiener フィルタを構成する目的の CVAE 音源モデルのそれぞれのパラメータを異なる変数 θ, ϑ として学習することを考える．DNMF 法と同様，テスト時と同じ音声分離プロセスをシミュレートすることで混合前の各音声信号を目標信号とした学習を行うことができる．すなわち，式 (12) の規準で学習した $\hat{\theta}$ を用いて式 (15) により $\hat{\mathbf{Z}}$, $\hat{\mathbf{C}}$, $\hat{\boldsymbol{\eta}}$ が求まれば，式 (16) の出力値と目標信号の誤差ができるだけ小さくなるように ϑ を学習すること

$$\hat{\vartheta} = \underset{\vartheta}{\mathrm{argmin}} \sum_j \mathcal{D}\left(|\mathbf{S}'_j| \,\middle|\, \frac{\eta_j \boldsymbol{\sigma}^2_{\vartheta}(\hat{\mathbf{z}}_j, \hat{\mathbf{c}}_j)}{\sum_{j'} \eta_{j'} \boldsymbol{\sigma}^2_{\vartheta}(\hat{\mathbf{z}}_{j'}, \hat{\mathbf{c}}_{j'})} \odot |\mathbf{Y}'|\right) \quad (17)$$

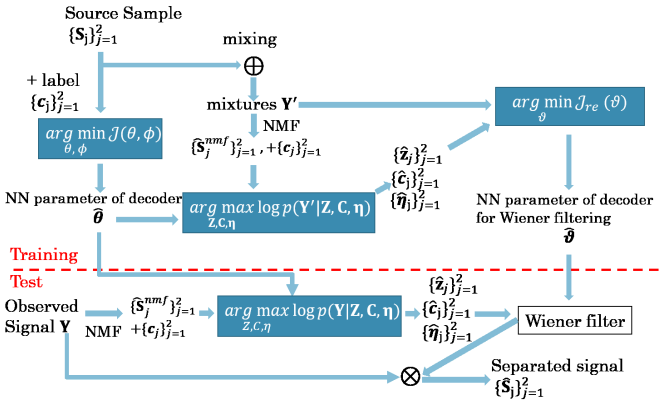

Fig. 1　Schematic overview of DVASS.

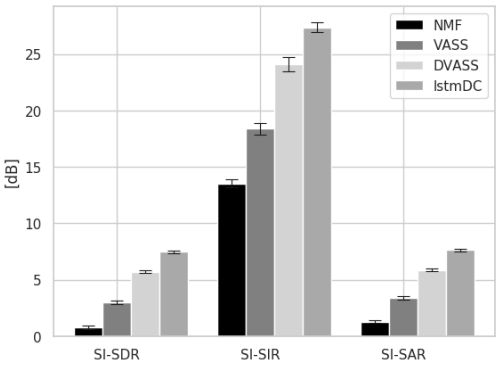

Fig. 2　Averaged separation performances [dB].

が目標となる．テスト時は，以上により学習した $\hat{\theta}$, $\hat{\vartheta}$ を用いて式 (15) と式 (16) と同じプロセスにより分離信号を得ることができる．式 (17) の規準を $\mathcal{J}_{re}(\vartheta)$ とし，以上の手法（DVASS 法）の全体像を示した図が Fig. 1 である．

4　評価実験

提案法の有効性を検証するため，2 話者の音声分離による実験的評価を行った．SNMF 法と DC [1] を比較対象とし，提案法である VASS 法と DVASS 法と比較した．音声データとして，CMU ARCTIC データベース [20] にある男性話者 2 名（bdl，rms）と女性話者 2 名（clb，slt）の音声発話を用いた．各話者ごとに 1000 発話を学習に利用し，132 発話をテストに用いた．テストの混合信号は音声信号のパワー比が 1 となるように 3 パターンの話者の組み合わせ（bdl+clb, bdl+rms，clb+slt）について計 244 文を作成した．また，DVASS の学習に用いる混合信号 $\mathbf{Y}'$ も同様に各組み合わせに 560 文を作成した．すべての音声信号のサンプリング周波数を 8 kHz とし，フレーム長 512 ms，フレームシフト 256 ms で短時間 Fourier 変換を行った．VASS で用いる CVAE 音源モデルの NN 構造は [10] で用いられる NN と同様に各 3 層のゲート付き CNN を使用した．DVASS で Wiener フィルタを構成するパラメータを推論するデコーダは CVAE 音源モデルのデコーダと同じ構造とした．NN 学習とモデルパラメータ $\mathbf{z}$ と $\mathbf{c}$ の更新には Adam [21] が用いた．VASS と DVASS は SNMF を 100 回反復して得られた $\hat{\mathbf{H}}$ を用いて構成された Wiener フィルタにより初期分離信号を求め，CVAE のエンコーダにより $\mathbf{z}$ の初期化を行った．話者が既知のため，$\mathbf{c}$ を正解ラベルの one-hot ベクトルに固定した．VASS と DVASS の更新回数を 2 とした．SNMF 法では，基底数を各音源に 10 とし，KL ダイバージェンス規準を用いた．評価指標として，scale-invariant signal-to-distortion ratio (SI-SDR)，scale-invariant signal-to-interference ratio (SI-SIR) と scale-invariant signals-to-artifacts ratio (SI-SAR) [22] を用いた．

実験結果を Fig.2 に示す．ベースラインの SNMF 法と比較すると，提案法の VASS 法による高い分離性能が確認できた．SNMF 法と VASS 法の差異は，音源モデルの違いによるものであることから，VAE に基づく表現能力の高い音源モデルが分離性能の向上に寄与していることを示している．また，すべての指標において DVASS 法は VASS 法よりも高い分離性能を示し，識別的学習の有効性が確認できた．しかし，DVASS 法は DC の高い分離性能までまだ改善の余地があった．

5　おわりに

本稿では，CVAE 音源モデルを利用したモノラル音声分離手法として VASS 法を提案するとともに，識別的規準により CVAE 音源モデルを学習する DVASS 法を提案した．特定話者 2 話者音声分離実験により，提案法の有効性を調査した．実験的評価により，提案法はベースライン手法より高い分離性能が確認された．

謝辞　本研究の一部は JST CREST JPMJCR19A3 の支援を受けたものである．

参考文献

[1] J. R. Hershey, *et al.*, *ICASSP*, 31–35, 2016.
[2] Y. Liu, *et al.*, *IEEE/ACM Trans. ASLP*, 27(12), 2092–2102, 2019.
[3] J. Le Roux, *et al.*, *IEEE JSTSP*, 13(2), 370–382, 2019.
[4] D. Wang, *et al.*, *IEEE/ACM Trans. ASLP*, 26(10), 1702–1726, 2018.
[5] D. D. Lee, *et al.*, *NIPS*, 556–562, 2001.
[6] P. Smaragdis, *et al.*, *ICA*, 414–421, 2007.
[7] F. Weninger, *et al.*, *Interspeech*, 865–869, 2014.
[8] A. A. Nugraha, *et al.*, *IEEE/ACM Trans. ASLP*, 24(9), 1652–1664, 2016.
[9] N. Makishima, *et al.*, *IEEE/ACM Trans. ASLP*, 27(10), 1601–1615, 2019.
[10] H. Kameoka, *et al.*, *Neural Computation*, 31(9), 1891–1914, 2019.
[11] L. Li, *et al.*, *IEEE Access*, 8(1), 228740–228753, 2020.
[12] S. Seki, *et al.*, *IEEE Access*, 7(1), 168104–168115, 2019.
[13] Y. Bando, *et al.*, *ICASSP*, 716–720, 2018.
[14] S. Leglaive, *et al.*, *MLSP*, 2018.
[15] K. Sekiguchi, *et al.*, *APSIPA*, 1233–1239, 2018.
[16] S. Leglaive, *et al.*, *ICASSP*, 101–105, 2019.
[17] D. P. Kingma, *et al.*, *NIPS*, 2014.
[18] H. Kameoka, *et al.*, *LVA/ICA*, 245–253, 2010.
[19] D. Kitamura, *et al.*, *IEEE/ACM Trans. ASLP*, 24(9), 1626–1641, 2016.
[20] J. Kominek., *et al.*, *WSS*, 2004.
[21] D. P. Kingma, *et al.*, *ICLR*, 2015.
[22] J. Le Roux, *et al.*, *ICASSP*, 626–630, 2019.