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
EA2018-154,SIP2018-160,SP2018-116(2019-03)

多チャンネル変分自己符号化器に基づく劣決定音源分離の評価

関　翔悟† 亀岡　弘和†† 李　莉††† 戸田　智基† 武田　一哉†

† 名古屋大学　〒464–8601 愛知県名古屋市千種区不老町
†† 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
††† 筑波大学　〒305–8573 茨城県つくば市天王台 1–1–1
E-mail: †seki.shogo@g.sp.m.is.nagoya-u.ac.jp, ††kameoka.hirokazu@lab.ntt.co.jp

あらまし　本稿では劣決定音源分離を扱う．多チャンネル非負値行列因子分解（MNMF）は劣決定音源分離に有効な手法であり，NMF を音源のパワースペクトログラムモデリングへと適用する．これは優決定条件下での MNMF でもある独立低ランク行列分析（ILRMA）にも取り入れられている．これらの手法は特定の音源に対して有効である一方で，NMF によってモデル化が困難な音源に対してはその分離性能が制限される．この問題に対して，NMF の代わりに条件付き変分自己符号化器（CVAE）を音源モデルとして利用する多チャンネル変分自己符号化器（MVAE）が最近提案され，ネットワークの柔軟な表現能力による高い分離性能が確認されている．本稿では優決定音源分離で定式化されている MVAE を劣決定音源分離へと適用し，停留点への収束が保証された最適化アルゴリズムを導く．2 チャンネルのマイクロフォンアレイを用いた 3 話者の分離による実験的評価を行い提案法の有効性を示す．

キーワード　劣決定音源分離，多チャンネル変分自己符号化器，多チャンネル非負値行列因子分解

An Evaluation of Underdetermined Source Separation Based on Multichannel Variational Autoencoder

Shogo SEKI†, Hirokazu KAMEOKA††, Li LI†††, Tomoki TODA†, and Kazuya TAKEDA†

† Nagoya University　Chikusa-ku Furo-cho, Nagoya, 464–8601 Japan
†† NTT Communication Science Laboratories　3–1 Morinosato Wakamiya, Atsugi, 243–0198 Japan
††† University of Tsukuba　1–1–1 Tennodai, Tsukuba, 305–8573 Japan
E-mail: †seki.shogo@g.sp.m.is.nagoya-u.ac.jp, ††kameoka.hirokazu@lab.ntt.co.jp

Abstract　This paper deals with a multichannel audio source separation problem under underdetermined conditions. Multichannel Non-negative Matrix Factorization (MNMF) is one of powerful approaches, which adopts the NMF concept for source power spectrogram modeling. This concept is also employed in Independent Low-Rank Matrix Analysis (ILRMA), a special class of the MNMF framework formulated under determined conditions. These methods work reasonably for particular types of sound sources, however, one limitation is that they can fail to work for sources with spectrograms that do not comply with the NMF model. To address this limitation, an extension of ILRMA called the Multichannel Variational Autoencoder (MVAE) method was recently proposed, where a Conditional VAE (CVAE) is used instead of the NMF model for source power spectrogram modeling. This approach has shown to perform impressively in determined source separation tasks thanks to the representation power of DNNs. This paper generalizes MVAE originally formulated under determined mixing conditions so that it can also deal with underdetermined cases. The proposed method was evaluated on a underdetermined source separation task of separating out three sources from two microphone inputs. Experimental results revealed that the generalized MVAE method achieved better performance than the MNMF method.

Key words　Underdetermined source separation, Multichannel variational autoencoder, Multichannel non-negative matrix factorization

1. はじめに

ブラインド音源分離（Blind Source Separation: BSS）は，音源とマイクロフォン間の伝達関数が未知であるという条件のもと，マイクロフォンアレイに入力される観測信号から個々の音源信号を分離する問題である．周波数領域 BSS では音源やアレイ応答の時間周波数表現について多様なモデルを設計することが可能である．例として，独立ベクトル分析（Independent Vector Analysis: IVA) [1], [2] では，同一音源に由来する周波数成分が時間とともにコヒーレントに変化するという仮定のもと，周波数成分ごとの音源分離とパーミュテーション整合を同時に解くことが可能である．

周波数領域 BSS における他のアプローチとして，楽曲採譜やモノラル音源分離に適用されていた非負値行列因子分解（Non-negative Matrix Factorization: NMF）[3], [4] の多チャンネル拡張を扱う手法が提案されている [5]〜[10]．NMF に基づく手法では，混合信号のパワースペクトログラムを 2 つの非負値行列の行列積として近似する．このとき，各時間フレームにおける混合信号のパワースペクトルは，少数の基底スペクトルと時変のゲイン成分の線形和として近似される．多チャンネル非負値行列因子分解（Multichannel Non-negative Matrix Factorization: MNMF）は，単チャンネル信号を前提とした NMF に対して多チャンネル信号を入力とする拡張手法であり，チャンネル間の位相差といった空間情報を音源分離における新たな手がかりとして利用する．

MNMF は，観測信号に含まれる音源数がマイクロフォンアレイのチャンネル数より大きいという劣決定条件において提案され [5]，のちに音源数がチャンネル数以下という優決定条件へと限定することで高速化を実現する分離アルゴリズムが提案された [6]．優決定条件下での MNMF の枠組みは，特に独立低ランク行列分析（Independent Low-Rank Matrix Analysis: ILRMA）とよばれる [11]．ILRMA を含む MNMF の枠組みにおいては収束が保証された最適化アルゴリズムを得ることが可能である一方，NMF によるスペクトログラムのモデル化が困難な音源に対しては，分離性能が限定される．

これに対して，現在では多チャンネル変分自己符号化器（Multichannel Variational Autoencoder: MVAE）とよばれる手法が提案されている [12]．MVAE は，ILRMA における各音源のスペクトログラムのモデル化において，NMF の代わりに条件付け変分自己符号化器（Conditional Variational Autoencoder: CVAE）[13], [14] を利用する手法である．MVAE では，音源の種類を表すラベル情報とスペクトログラムを組み合わせた学習データを用いて，事前に CVAE を学習させる．分離時には，学習された CVAE を各音源のスペクトログラムを生成する生成モデルとして，ILRMA における NMF の代わりに利用する．ニューラルネットワークの高い表現能力を音源のパワースペクトログラムモデリングへ利用することで，MVAE は ILRMA を超える分離性能を達成している．

本稿では，優決定音源分離を想定する MVAE を拡張し，劣決定音源分離へと適用可能な一般化された MVAE を提案する (Table 1.)．劣決定条件での定式化にしたがって提案法を示すとともに，MNMF などと同様に停留点への収束性が保証された最適化アルゴリズムを導く．

Table 1 Comparison with the conventional methods

	Separation	Source model
ILRMA [6], [8], [10]	Determined	NMF
MNMF [5], [7], [9]	Underdetermined	NMF
MVAE [12]	Determined	VAE
Proposed	Underdetermined	VAE

2. 定式化

観測チャンネル数および音源数をそれぞれ I，J とし，観測信号および音源信号の時間周波数表現をそれぞれ

$$\mathbf{s}(f,n) = [s_1(f,n), \cdots, s_j(f,n), \cdots, s_J(f,n)]^{\mathsf{T}} \in \mathbb{C}^J \quad (1)$$

$$\mathbf{x}(f,n) = [x_1(f,n), \cdots, x_i(f,n), \cdots, x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I \quad (2)$$

とする．ここで $(\cdot)^{\mathsf{T}}$ は転置を表し，i，j，f，n はそれぞれチャネル，音源，周波数，時間を表すインデックスである．本稿では，周波数領域における瞬時混合表現で表される混合系

$$\mathbf{x}(f,n) = \mathbf{A}(f)\mathbf{s}(f,n) \quad (3)$$

$$\mathbf{A}(f) = [\mathbf{a}_1(f), \cdots, \mathbf{a}_j(f), \cdots, \mathbf{a}_J(f)] \in \mathbb{C}^{I\times J} \quad (4)$$

を用いる．ここで $\mathbf{A}(f)$ は混合行列である．

いま，$s_j(f,n)$ が平均 0，分散 $v_j(f,n)$ の複素ガウス分布 $\mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n))$ にしたがう Local Gaussian Model (LGM) [15] を仮定する．$j \neq j'$ において $s_j(f,n)$，$s_{j'}(f,n)$ が独立であるとき，音源信号 $\mathbf{s}(f,n)$ は

$$\mathbf{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|\mathbf{0}, \mathbf{V}(f,n)) \quad (5)$$

にしたがう．ここで $\mathbf{V}(f,n)$ は $v_1(f,n)$, $\cdots$, $v_J(f,n)$ を要素にもつ対角行列である．(3)，(5) より観測信号 $\mathbf{x}(f,n)$ は

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, \mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)) \quad (6)$$

にしたがう．ここで $(\cdot)^{\mathsf{H}}$ は共役転置を表す．

したがって，混合行列 $\mathcal{A} = \{\mathbf{A}(f)\}_f$ および音源の分散 $\mathcal{V} = \{v_j(f,n)\}_{f,n}$ を用いて観測信号 $\mathcal{X} = \{\mathbf{x}(f,n)\}_{f,n}$ に対する対数尤度は以下で与えられる．

$$\begin{aligned}&\log p(\mathcal{X}|\mathcal{A},\mathcal{V})\\ &\overset{c}{=} -\sum\nolimits_{f,n}\big[\mathrm{tr}(\mathbf{x}^{\mathsf{H}}(f,n)(\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f))^{-1}\mathbf{x}(f,n))\\ &\qquad +\mathrm{logdet}(\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f))\big]\end{aligned} \quad (7)$$

ここで $\overset{c}{=}$ はパラメータに関する等号を表す．

3. 関連研究

3.1 多チャンネル非負値行列因子分解（MNMF）

観測信号の空間共分散はステアリングベクトル $\mathbf{a}_j(f)$ および

分散 $v_j(f,n)$ を用いて以下で表される．

$$\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f) = \sum\nolimits_j \mathbf{a}_j(f)v_j(f,n)\mathbf{a}_j^{\mathsf{H}}(f) = \sum\nolimits_j v_j(f,n)\mathbf{R}_j(f) \quad (8)$$

ここで $\mathbf{R}_j(f)$ は音源 j の空間共分散を表す．MNMF では，IVA と同様に $v_j(f,n)$ に対して制約を加えることにより，周波数成分ごとの音源分離とパーミュテーション整合を可能にする．MNMF では，各音源信号の分散 $v_j(f,n)$ を K_j のスペクトルテンプレート $h_{j,1}(f), \cdots, h_{j,K_j}(f) \geq 0$ とその時変な励起パターン $u_{j,1}(n), \cdots, u_{j,K_j}(n) \geq 0$ を用いた

$$v_j(f,n) = \sum\nolimits_{k=1}^{K_j} h_{j,k}(f)u_{j,k}(n), \quad (9)$$

もしくは，$\sum_k b_{j,k} = 1$ を満たす指示変数 $b_{j,k} \in [0,1]$ を加え音源間で共有した K のスペクトルテンプレートおよびその励起パターンを用いた

$$v_j(f,n) = \sum\nolimits_{k=1}^{K} b_{j,k}h_k(f)u_k(n) \quad (10)$$

によってモデル化する．

MNMF の最適化アルゴリズムは空間共分散 $\mathcal{R} = \{\mathbf{R}_j(f)\}_{j,f}$ と音源モデル（$\mathcal{H}_1 = \{h_{j,k}(f)\}_{j,k,f}$, $\mathcal{U}_1 = \{u_{j,k}(n)\}_{j,k,n}$ もしくは $\mathcal{B} = \{b_{j,k}\}_{j,k}$, $\mathcal{H}_2 = \{h_k(f)\}_{k,f}$, $\mathcal{U}_2 = \{u_k(n)\}_{k,n}$）の反復更新によって構成され，それぞれの更新則は Majorization-Minimization (MM) アルゴリズムにより得られる [16].

3.2 独立低ランク行列分析（ILRMA）

ILRMA は優決定音源分離を扱う MNMF の特殊なケースである．混合系を仮定する MNMF とは異なり，ILRMA では混合行列が正則な場合である分離系

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f)\mathbf{x}(f,n) \quad (11)$$

$$\mathbf{W}(f) = [\mathbf{w}_1(f), \cdots, \mathbf{w}_i(f), \cdots, \mathbf{w}_I(f)] \in \mathbb{C}^{I\times J} \quad (12)$$

を仮定する．ここで混合行列の逆行列 $\mathbf{W}^{\mathsf{H}}(f)$ は分離行列とよばれ，(5)，(11) より観測信号 $\mathbf{x}(f,n)$ は

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, (\mathbf{W}^{\mathsf{H}}(f))^{-1}\mathbf{V}(f,n)(\mathbf{W}(f))^{-1}) \quad (13)$$

にしたがう．よって，分離行列 $\mathcal{W} = \{\mathbf{W}(f)\}_f$ および $\mathcal{V}$ を用いて対数尤度は次式で与えられる．

$$\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) \stackrel{c}{=} 2N\sum\nolimits_f \log|\det\mathbf{W}^{\mathsf{H}}(f)| - \sum\nolimits_{f,n,j}\left[\log v_j(f,n) + \frac{|\mathbf{w}_j^{\mathsf{H}}(f)\mathbf{x}(f,n)|^2}{v_j(f,n)}\right] \quad (14)$$

ILRMA において，音源の分散 $v_j(f,n)$ は MNMF と同様に (9) または (10) としてモデル化される．

MNMF と同様に $\mathcal{H}_1, \mathcal{U}_1$ または $\mathcal{B}, \mathcal{H}_2, \mathcal{U}_2$ について，MM アルゴリズムに基づく更新則が得られる．また，ILRMA は IVA の自然な拡張であることから，IVA において提案された反復射影（Iterative Projection）法 [17] とよばれる高速な更新アルゴリズムにより分離行列を更新することが可能である．

3.3 多チャンネル変分自己符号化器（MVAE）

ILRMA を含めた MNMF の枠組みでは，音源モデル $v_j(f,n)$ の表現が (9) または (10) に制限されることにより，実際にはモデル化が困難な音源に対してはその分離性能が限定される．MVAE は NMF の代わりに学習済みの CVAE を用いた ILRMA の拡張である．$\tilde{\mathbf{S}} = \{s(f,n)\}_{f,n}$ をある音源の複素スペクトログラムとする．MVAE は補助特徴量 c を伴う CVAE により $\tilde{\mathbf{S}}$ の生成モデルを学習する．CVAE はエンコーダ・デコーダ型のネットワークで構成され，ラベル付きの学習データ $\{\tilde{\mathbf{S}}_m, c_m\}_{m=1}^M$ を用いて事前に学習される．エンコーダの分布 $q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c)$ はガウス分布

$$q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c) = \prod\nolimits_k \mathcal{N}(z(k)|\mu_\phi(k;\tilde{\mathbf{S}}, c), \sigma_\phi^2(k;\tilde{\mathbf{S}}, c)) \quad (15)$$

デコーダの分布 $p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g)$ は平均 0 の複素ガウス分布

$$p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g) = \prod\nolimits_{f,n} \mathcal{N}_{\mathbb{C}}(s(f,n)|0, v(f,n)) \quad (16)$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\mathbf{z},c) \quad (17)$$

を表すようにネットワークを学習する．ここで，$\mathbf{z}$ は潜在変数，$\mu_\phi(k;\tilde{\mathbf{S}}, c)$, $\sigma_\phi^2(k;\tilde{\mathbf{S}}, c)$ はそれぞれ要素 k のエンコーダ出力 $\mu_\phi(\tilde{\mathbf{S}}, c)$, $\sigma_\phi^2(\tilde{\mathbf{S}}, c)$, $\sigma_\theta^2(f,n;\mathbf{z},c)$ は要素 (f,n) のデコーダ出力 $\sigma_\theta^2(\mathbf{z},c)$ であり，g は生成されたスペクトログラムに対するグローバルスケールである．エンコーダ，デコーダのネットワークパラメータ ϕ, θ は以下の目的関数により学習される．

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\tilde{\mathbf{S}},c)\sim p_{\mathrm{D}}(\tilde{\mathbf{S}},c)}\left[\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\tilde{\mathbf{S}},c)}[\log p(\tilde{\mathbf{S}}|\mathbf{z},c)] - \mathrm{KL}[q(\mathbf{z}|\tilde{\mathbf{S}},c)||p(\mathbf{z})]\right] \quad (18)$$

ここで，$\mathbb{E}_{(\tilde{\mathbf{S}},c)\sim p_{\mathrm{D}}(\tilde{\mathbf{S}},c)}[\cdot]$ は学習サンプルについての標本平均，$\mathrm{KL}[\cdot||\cdot]$ は Kullback-Leibler（KL）ダイバージェンスを表す．

学習されたデコーダの分布 $p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g)$ は，学習サンプルに含まれるスペクトログラムを生成するユニバーサルな生成モデルとして考えられる．MVAE は CVAE のデコーダを (14) における音源モデルとして利用し，デコーダの入力である潜在表現 $\mathbf{z}$ および潜在コード c を推定すべきモデルパラメータとして扱う．したがって，MVAE の最適化アルゴリズムは IP を用いた分離行列の推定，MM アルゴリズムに基づくグローバルスケールの推定，およびバックプロパゲーションを用いたデコーダ入力の推定によって構成される．MVAE では，ニューラルネットワークである VAE の強力な表現能力を音源のパワースペクトログラムモデリングへ利用することが可能である．

4. 提案法：一般化 MVAE

Fig. 1 に一般化 MVAE の概要を (9) で与えられる音源モデルをもつ MNMF とともに示す．提案法は分離系において提案されている MVAE を一般化し，劣決定音源分離へと適用する．従来の MVAE と同様に，一般化 MVAE においても，学習された CVAE のデコーダをパワースペクトログラムの生成モデルとして利用する．

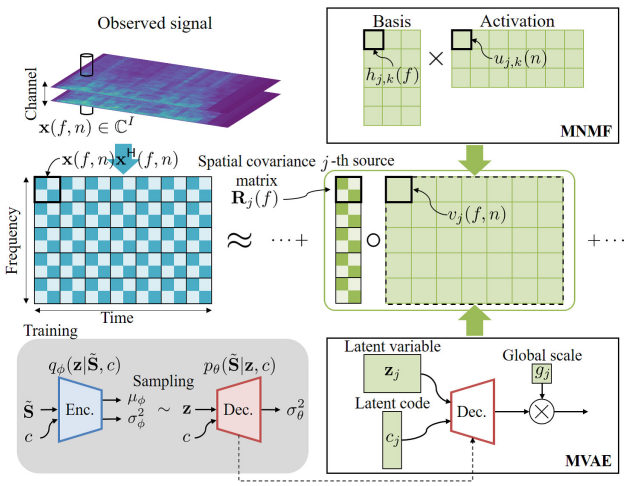


Fig. 1　Illustration of Generalized MVAE

デコーダの分布が LGM と同形で与えられることから，(7) で表される対数尤度に対して，デコーダの分散 $g_j \cdot p_\theta(\tilde{\mathbf{S}}_j|\mathbf{z}_j, c_j)$ を用いることが可能である．したがって，MM アルゴリズムに基づく MNMF の最適化アルゴリズム導出と同様の方法で，$\mathcal{R}$, $\mathcal{G} = \{g_j\}_j$，および $\Psi = \{\mathbf{z}_j, c_j\}_j$ に対する反復アルゴリズムが導出可能である．$\mathsf{X}(f,n)$，$\hat{\mathsf{X}}(f,n)$ をそれぞれ

$$\mathsf{X}(f,n) = \mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n) \tag{19}$$

$$\hat{\mathsf{X}}(f,n) = \mathbf{A}_j(f)\mathbf{V}(f,n)\mathbf{A}_j(f) \tag{20}$$

とする．負の対数尤度に対して，以下の不等式が成立する [16].

$$\begin{aligned}\mathcal{L} &= -\log p(\mathcal{X}|\mathcal{A},\mathcal{V}) \\ &\overset{c}{\leq} \sum_j \sum_{f,n} \left[\frac{\mathrm{tr}(\mathsf{X}(f,n)\mathbf{P}_j(f,n)\mathbf{R}_j^{-1}(f,n)\mathbf{P}_j(f,n))}{v_j(f,n)} \right. \\ &\qquad \left. + v_j(f,n)\mathrm{tr}(\mathbf{K}^{-1}(f,n)\mathbf{R}_j(f,n)) \right] \end{aligned} \tag{21}$$

ここで，$\mathcal{P} = \{\mathbf{P}_j(f,n)\}_{j,f,n}$，$\mathcal{K} = \{\mathbf{K}(f,n)\}_{f,n}$ は補助変数であり，等号成立は以下の場合である．

$$\mathbf{P}_j(f,n) = v_j(f,n)\mathbf{R}_j(f,n)\left(\sum\nolimits_j v_j(f,n)\mathbf{R}_j(f,n)\right)^{-1} \tag{22}$$

$$\mathbf{K}(f,n) = \mathsf{X}(f,n) \tag{23}$$

このとき，(21) の右辺を負の対数尤度 $\mathcal{L}$ に対する majorizer として利用することができる．したがって，$\mathcal{R}$, $\mathcal{G}$ および，Ψ について majorizer を最小化し，(22) および (23) を用いて $\mathcal{P}$, $\mathcal{K}$ をそれぞれ更新する反復アルゴリズムは，$\mathcal{L}$ に対して停留点への収束が保証される．空間共分散 $\mathcal{R}$ の更新は MNMF と同様に次式で与えられる [18].

$$\mathbf{R}_j(f) \leftarrow \mathbf{\Lambda}_j^{-1}(f)\#(\mathbf{R}_j(f)\mathbf{\Omega}_j(f)\mathbf{R}_j(f)) \tag{24}$$

ただし，$\#$ は半正定値行列に対する幾何平均

$$\mathbf{G}\#\mathbf{H} = \mathbf{G}^{\frac{1}{2}}(\mathbf{G}^{-\frac{1}{2}}\mathbf{H}\mathbf{G}^{-\frac{1}{2}})^{\frac{1}{2}}\mathbf{G}^{\frac{1}{2}} \tag{25}$$

Algorithm 1 MVAE algorithm

```
Train φ and θ with (18)
Initialize R, Ψ, and G
repeat
  for each j do
    Update R_j = {R_j(f)}_f using (25)
    Update ψ_j = {z_j, c_j} with (21) using Backpropagation
    Update g_j using (29)
  end for
until converge
```

であり $\mathbf{\Lambda}_j(f)$, $\mathbf{\Omega}_j(f)$ はそれぞれ以下である．

$$\mathbf{\Lambda}_j(f) = \sum\nolimits_n v_j(f,n)\hat{\mathsf{X}}^{-1}(f,n), \tag{26}$$

$$\mathbf{\Omega}_j(f) = \sum\nolimits_n v_j(f,n)\hat{\mathsf{X}}^{-1}(f,n)\mathsf{X}(f,n)\hat{\mathsf{X}}^{-1}(f,n) \tag{27}$$

また，majorizer は各音源に対する目的関数へ分離可能であることから，Ψ についてバックプロパゲーションによる更新が並列可能となる．ここで，c_j を更新においては one-hot 表現の総和が 1 となることを保証する必要があるが，c_j の出力層にソフトマックス関数

$$c_j = \mathrm{softmax}(d_j) \tag{28}$$

を挿入し d_j を代わりに推定するネットワークパラメータとして扱うことで保証される．$\mathcal{G}$ の更新は以下のように得られる．

$$g_j \leftarrow g_j \times \sqrt{\frac{\sum_{f,n} \sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\hat{\mathsf{X}}^{-1}(f,n)\mathsf{X}(f,n)\hat{\mathsf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_{f,n} \sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\hat{\mathsf{X}}^{-1}(f,n)\mathbf{R}_j(f))}} \tag{29}$$

したがって，提案法のアルゴリズムは Algorithm 1 となる．

5.　実験的評価

5.1　実 験 条 件

提案法の有効性を調査するために 2 チャンネルのマイクロフォンアレイを用いた 3 話者分離による実験的評価を行った．実験データとして，Voice Conversion Challenge (VCC) 2018 データセットの音声サンプルを用いた [19]．データセット内には男女各 6 話者の英語発話音声が収録されており，実験においては男女各 2 話者（SF1, SF2, SM1, SM2）の発話を利用した．各話者について学習，評価にそれぞれ 81 発話，35 発話を用いた．

Fig. 2 に実験環境を示す．ここで，○，× はそれぞれマイクロフォン，音源の位置を表す．各話者の評価音声を用いて，全 3 話者のパターン（SF1+SF2+SM2，SF1+SM1+SF2，SF1+SM1+SM2，SF1+SM1+SM2）の混合音声を作成した．各パターンについて，話者ごとにランダムに評価音声を選択，配置することによって計 10 サンプルの混合音声を作成した．全ての混合音声は，残響環境がそれぞれ $T_{60} = 78$ ms, $T_{60} = 351$ ms の条件で作成した．

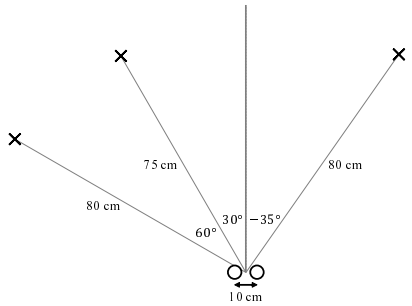


Fig. 2　Position of microphones and sources

観測信号についてサンプリング周波数は 16 kHz とし，STFT のフレームサイズ，シフトサイズをそれぞれ 256 ms, 128 ms とした．CVAE のエンコーダ，デコーダには Fig. 3 に示されるネットワーク構造をそれぞれ利用した．音源のクラスラベルについては話者情報を利用し，c_j は 4 次元の one-hot 表現とした．CVAE の学習には Adam [20] を用い，$\mathbf{z}_j$，c_j の最適化には SGD を用いた．

提案法を含めた以下の手法を比較した．

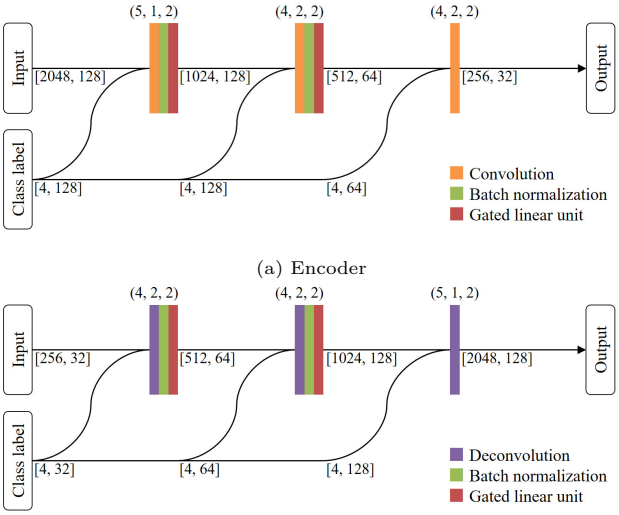
- MNMF1: MNMF w/ source model given by (9)
- MNMF2: MNMF w/ source model given by (10)
- Semi-blind MNMF: MNMF2 w/ speaker templates
- MVAE1: MVAE initialized by MNMF1
- MVAE2: MVAE initialized by MNMF2

MNMF, MVAE の反復回数はそれぞれ 300, 100 とした．MVAE の初期値については，反復回数を 200 とした MNMF を用いた．MVAE において，空間共分散 $\mathcal{R}$ については，MNMF によって得られた値を初期値とした．MNMF によって得られた音源 j の分散 $v_j(f,n)$ を CVAE におけるエンコーダへ入力することで，潜在変数 $\mathbf{z}_j$ の初期値を得た．このとき，音源のクラスラベル c_j については，全ての要素の値に要素数の逆数を初期値として与えることで，エンコーダにおいてクラスラベルによる条件付けの影響を無視するように設定した．また，g_j については，MNMF により得られる分離信号のパワーを初期値とした．MNMF の基底数は，各音源あたり 10 とした [7]．Semi-blind MNMF では，全 4 話者のスペクトルテンプレートを用いた．話者ごとのスペクトルテンプレートの学習には，CVAE と同一の学習データセットを利用し，Itakura-Saito NMF（IS-NMF）[4] を用いた．IS-NMF における反復回数は 1000 とした．

評価指標として，リファレンス信号と分離信号間の信号対歪み比（Signal-to-Distortion Ratio: SDR），信号対線形歪み比（source Image-to-Spatial distortion ratio: ISR），信号対干渉比（Signal-to-Inference Ratio: SIR），信号対非線形歪み比（Signal-to-Artifact Ratio: SAR）を用いて [21]，平均値を算出した．

5.2　実 験 結 果

実験結果を Fig. 4 に示す．ベースラインである MNMF1, MNMF2 と比較すると，いずれの場合においても提案法による高い分離性能が確認できる．ベースラインと提案法の差異は，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



(a) Encoder

(b) Decoder

Fig. 3　Network configurations of (a) encoder and (b) decoder, where $[\mathtt{c}, \mathtt{l}]$ denotes the input channel and frame length. Both convolution and deconvolution represents 1-dimensional operation. $(\mathtt{k}, \mathtt{s}, \mathtt{p})$ represents the kernel size, the stride size along frame, and the zero padding size at both ends of frame, respectively.

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．セミブラインドな条件での MNMF との比較においても，提案法が高い分離性能を有していることが確認できる．各評価指標についてみると，SIR において提案法による顕著な性能向上が確認できる．音源モデルに VAE を利用することで，空間共分散がより正確に推定され，目的音源以外の干渉音を高い精度で抑圧しているためであると考えられる．

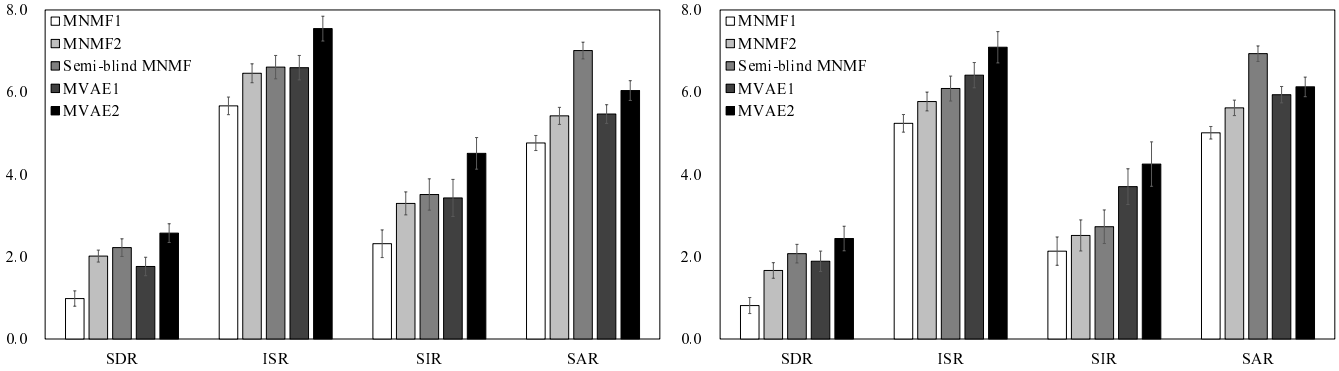
6.　お わ り に

本稿では，優決定音源分離において提案されている MVAE の拡張を行い，劣決定音源分離に適用可能な一般化 MVAE を提案した．提案法は，混合系を仮定した定式化より導かれ，分離系を想定する従来の MVAE を特殊なケースとして含む一般的な枠組みである．また，MNMF などと同様に停留点への収束性が保証された最適化アルゴリズムを導出した．2 チャンネルのマイクロフォンアレイを用いた 3 話者の音源分離を想定した実験的評価により，提案法の有効性を調査した．実験的評価より，VAE を音源モデルとして用いる提案法において高い分離性能が確認され，表現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が示された．また，セミブラインドな条件での MNMF との比較においても提案法の有効性が確認された．

今後の課題として，音楽信号に対する分離性能の調査，独立深層学習行列分析（Independent Deeply-Learned Matrix Analysis: IDLMA）[22] の劣決定音源分離を想定した DNN ベースの識別的な多チャンネル音源分離手法 [23] との比較が挙げられる．

謝　　辞

本研究の一部は，JSPS 科研費 17H01763 の助成を受け実施したものである．



(a) $T_{60} = 78$ ms (b) $T_{60} = 351$ ms

Fig. 4 Averaged separation performances [dB] at (a) $T_{60} = 78$ ms and (b) $T_{60} = 351$ ms

文　献

[1] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ica to multivariate components," International Conference on Independent Component Analysis and Signal Separation, pp.165–172 2006.

[2] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," International Conference on Independent Component Analysis and Signal Separation, pp.601–608, 2006.

[3] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.177–180, 2003.

[4] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," Neural computation, vol.21, no.3, pp.793–830, 2009.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.3, pp.550–563, 2010.

[6] 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," International Conference on Latent Variable Analysis and Signal Separation, pp.245–253, 2010.

[7] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," IEEE Transactions on Audio, Speech, and Language Processing, vol.21, no.5, pp.971–982, 2013.

[8] 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 Transactions on Audio, Speech and Language Processing, vol.24, no.9, pp.1622–1637, 2016.

[9] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," IEEE International Workshop on Acoustic Signal Enhancement, pp.1–5, 2016.

[10] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, "Independent low-rank matrix analysis based on complex student's t-distribution for blind audio source separation," IEEE International Workshop on Machine Learning for Signal Processing, pp.1–6, 2017.

[11] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," Audio Source Separation, pp.125–155, Springer, 2018.

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

[13] D.P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[14] D.P. Kingma, S. Mohamed, D.J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," Advances in Neural Information Processing Systems, pp.3581–3589, 2014.

[15] E. Vincent, M.G. Jafari, S.A. Abdallah, M.D. Plumbley, and M.E. Davies, "Probabilistic modeling paradigms for audio source separation," Machine Audition: Principles, Algorithms and Systems, pp.162–185, IGI global, 2011.

[16] H. Kameoka, H. Sawada, and T. Higuchi, "General formulation of multichannel extensions of nmf variants," Audio Source Separation, pp.95–124, Springer, 2018.

[17] N. Ono, "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.

[18] K. Yoshii, "Correlated tensor factorization for audio source separation," IEEE International Conference on Acoustics, Speech and Signal Processing, pp.731–735, 2018.

[19] 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 preprint arXiv:1804.04262, 2018.

[20] D.P. Kingma and J.L. Ba, "Adam: Amethod for stochastic optimization," International Conference on Learning Representations, 2015.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.4, pp.1462–1469, 2006.

[22] 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," IEEE European Signal Processing Conference, pp.1557–1561 2018.

[23] A.A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks.," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol.24, no.9, pp.1652–1664, 2016.