多チャンネル変分自己符号化器を用いた音源分離と残響除去の統合的アプローチ *

☆井上翔太 [1]　亀岡弘和 [2]　李莉 [1]　関翔悟 [3]　牧野昭二 [1]

[1] 筑波大学　[3] 名古屋大学
[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）は，音源に関する情報や音源とマイク間の伝達関数等の事前情報を用いずに観測された信号のみから個々の信号を推定する技術であり，ロボット聴覚，補聴器，音声認識や異常音検知等様々なアプリケーションの精度向上に貢献する．マイクロホンの数が音源数を上回る優決定条件下においては，音源信号間の独立性を最大化するように分離フィルタを推定することを目的とする独立成分分析 (Independent Component Analysis: ICA) [1] が有用であることが知られており，それに基づく時間周波数領域での分離手法が数多く提案されている [2–5]．これらの手法は，時間周波数領域で成立する音源に関する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を有効に活用できるという利点がある．例えば，独立低ランク行列分析 (Independent Low-Rank Matrix Analysis: ILRMA) [3,4] は，各音源信号のパワースペクトログラムを非負値行列とみなし，非負値行列因子分解 (Non Negative Matrix Factorization: NMF) [6] で近似表現する手法である．これは，各時間フレームにおけるパワースペクトルを時間的に変化する振幅によってスケーリングされた基底スペクトルの線形和で近似することに相当する．ILRMA は音源のスペクトル構造を手がかりとしてパーミュテーション整合と呼ぶ問題を解決しながら高精度な音源分離を実現する．この手法は限られた数の基底の線形和で表現できるような特定の音源に対して有効である一方，表現が困難である音源に対しての分離性能が制限されるという課題があった．

近年，深層学習の発展に伴い，深層ニューラルネットワーク (Deep Neural Network: DNN) を用いることで柔軟かつ高精度に音源信号のスペクトログラムをモデル化することが可能になった．DNN の豊かな関数表現力に着目し，我々は条件付き変分自己符化器 (Conditional Variational Autoencoder: CVAE) [7] を用いて各音源信号のスペクトログラムの生成過程をモデル化した多チャンネル変分自己符号化器 (Multichannel VAE: MVAE) を提案した [5]．MVAE 法は，CVAE のデコーダ入力を音源モデルのパラメータとみなし，分離フィルタとともに推定する手法である．MVAE 法はより柔軟な音源モデルを用いたことで ILRMA より高い分離性能が得られることが実験的に示されている [5]．しかし，従来のモデルでは，室内インパルス応答長が時間周波数展開における時間窓長よりも十分に短い場合を仮定しており，残響が長い場合を考慮していない．従って，従来の MVAE 法は観測信号に長い残響が含まれる場合に分離性能が劣化する問題点がある．

そこで本稿では，MVAE 法を拡張し，残響下での観測信号を時間周波数領域における畳み込み混合の形で表現し，パラメータ推定により音源分離と残響除去を統合した手法を提案する．さらに，長い残響を含んだ音声信号を用いて音源分離実験を行い，提案手法の有効性を確認する．

2　MVAE 法を用いた音源分離

2.1　瞬時混合近似に基づく定式化

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．i 番目のマイクで観測される信号と j 番目の音源信号の時間周波数成分をそれぞれ $x_i(f,n)$ と $s_j(f,n)$ とする．ただし，f と n は周波数と時間フレームのインデックスである．優決定条件下において $I=J$ とする．音源とマイクの間の室内インパルス応答長が時間周波数展開における窓長よりも十分短い場合には，音源信号 $\boldsymbol{s}(f,n)=[s_1(f,n),\ldots,s_J(f,n)]^{\mathsf{T}}\in\mathbb{C}^J$ と観測信号 $\boldsymbol{x}(f,n)=[x_1(f,n),\ldots,x_I(f,n)]^{\mathsf{T}}\in\mathbb{C}^I$ の関係性は瞬時混合系を用いて

$$\boldsymbol{s}(f,n)=\boldsymbol{W}^{\mathsf{H}}(f)\boldsymbol{x}(f,n) \tag{1}$$

$$\boldsymbol{W}^{\mathsf{H}}(f)=[\boldsymbol{w}_1(f),\ldots,\boldsymbol{w}_I(f)]\in\mathbb{C}^{I\times I} \tag{2}$$

と表せる．ここで，$\boldsymbol{W}^{\mathsf{H}}(f)$ は分離フィルタを表し，$(\cdot)^{\mathsf{T}}$ は行列の転置であり，$(\cdot)^{\mathsf{H}}$ はエルミート転置である．

次に，観測信号が生成されるプロセスを生成モデルにより記述する．音源 j の複素スペクトログラム $s_j(f,n)$ を

$$s_j(f,n)\sim\mathcal{N}_{\mathbb{C}}(s_j(f,n)|0,v_j(f,n)) \tag{3}$$

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確率変数と仮定する．各音源が統計的に独立である場合，$\boldsymbol{s}(f,n)$ は

$$\boldsymbol{s}(f,n)\sim\mathcal{N}_{\mathbb{C}}(\boldsymbol{s}(f,n)|0,\boldsymbol{V}(f,n)) \tag{4}$$

に従う．ここで，$\boldsymbol{V}(f,n)$ は $v_1(f,n),\ldots,v_I(f,n)$ を対角成分に持つ対角行列である．式 (1) と式 (4) より，観測信号 $\boldsymbol{x}(f,n)$ は

$$\boldsymbol{x}(f,n)\sim\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}(f,n)|0,(\boldsymbol{W}^{\mathsf{H}}(f))^{-1}\boldsymbol{V}(f,n)\boldsymbol{W}(f)^{-1}) \tag{5}$$

に従う．従って，混合信号 $\mathcal{X}=\{\boldsymbol{x}(f,n)\}_{f,n}$ が観測されたとき，分離フィルタ $\mathcal{W}=\{\boldsymbol{W}(f)\}_f$ と各音源のパワースペクトログラム $\mathcal{V}=\{\boldsymbol{V}(f,n)\}_{f,n}$ についての対数尤度関数は

---

* Unified approach for determined BSS and dereverberation using multichannel variational autoencoder. by Shota Inoue (University of Tsukuba), Hirokazu Kameoka (NTT Communication Science Science Laboratories), Li Li (University of Tsukuba), Shogo Seki (University of Nagoya), Shoji Makino (University of Tsukuba)

$$\mathcal{L}(\mathcal{V},\mathcal{W}|\mathcal{X}) \stackrel{c}{=} -2N\log\left|\det \boldsymbol{W}^{\mathsf{H}}(f)\right| + \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{v_j(f,n)}\right) \tag{6}$$

のように書ける．ここで，$=^c$ はパラメータに依存しない項を除いた等号を表す．

2.2 MVAE 法

MVAE 法は上記の対数尤度関数を大きくするように分離行列 $\mathcal{W}$ を推定する手法である．式 (6) のとおり，音源のパワースペクトログラム $v_j(f,n)$ に制約がない場合には音源分離問題が周波数 f ごとに分解されるため，分離信号のインデックスにパーミュテーションの任意性が生じる．$s_j(f,n)$ が異なる周波数間で相関を持つ場合，その相関関係を手がかりとすることでパーミュテーション整合と周波数ごとの音源分離の問題を同時解決できる場合がある．独立ベクトル分析 (Independent Vector Analysis: IVA)，ILRMA，MVAE 法などがその例であり，MVAE 法では $\boldsymbol{S} = \{s_j(f,n)\}_{f,n}$ の同時分布を CVAE のデコーダで記述することによりこれを実現する．

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 (3) の局所ガウス音源モデルと同形の確率モデル

$$p_\theta(\boldsymbol{S}|\boldsymbol{z},c,g) = \prod_{f,n}\mathcal{N}_{\mathbb{C}}(s(f,n)|0,v(f,n)), \tag{7}$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\boldsymbol{z},c) \tag{8}$$

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リング係数を表す．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

$$q_\phi(\boldsymbol{z}|\boldsymbol{S},c) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_\phi(\boldsymbol{S},c),\mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\boldsymbol{S},c))) \tag{9}$$

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

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{S},c)}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},c)] - KL[q_\phi(\boldsymbol{z}|\boldsymbol{S},c)||p(\boldsymbol{z})]] \tag{10}$$

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解釈できる．

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を式 (6) と同形の尤度関数に帰着できる．従って，式 (6) が大きくなるように分離フィルタ $\mathcal{W}$，CVAE 音源モデルパラメータ $\Psi = \{\boldsymbol{z}_j,c_j\}_j$，スケーリング係数 $\mathcal{G} = \{g_j\}_j$ を反復更新することで，式 (6) の停留点の探索が可能である．式 (6) を上昇させる $\mathcal{W}$ の更新には IVA，ILRMA と同様に反復射影法 (Iterative Projection: IP)

$$\boldsymbol{w}_j(f) \leftarrow (\mathbf{W}^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f))^{-1}\boldsymbol{e}_j, \tag{11}$$

$$\boldsymbol{w}_j(f) \leftarrow \frac{\boldsymbol{w}_j(f)}{\sqrt{\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j(f)\boldsymbol{w}_j(f)}} \tag{12}$$

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 列のベクトルである．また，式 (6) を上昇させる Ψ の更新は誤差逆伝播法を用いて，$\mathcal{G}$ の更新は

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z}_j,c_j)} \tag{13}$$

を用いて行う．ただし，式 (13) は $\mathcal{W}$ と Ψ が固定された下で式 (6) を最大化する更新式である．

[5] に示されているように，MVAE 法は残響時間が短い場合には極めて強力である一方，残響が長い場合には従来の瞬時混合近似が成り立つことを仮定した手法と同様に分離性能が劣化する傾向にある．

3 提案手法

この問題に対して本稿では，長い残響環境下での観測信号を時間周波数領域における畳み込み混合の形 [3,8,9] で表現し，パラメータ推定を通して音源分離と残響除去を統合的に行う手法を提案する．

3.1 時間周波数領域における畳み込み混合モデルに基づく定式化

多チャンネル有限インパルス応答で表されるような時間周波数領域における畳み込み混合モデルを用いると，音源信号と観測信号の関係は次式で記述できる．

$$\boldsymbol{s}(f,n) = \sum_{n'=0}^{N'}\boldsymbol{W}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n') \tag{14}$$

ここで，$\boldsymbol{W}(f,n')$, $0\le n'\le N'$ は $I\times I$ 行列の分離フィルタであり，音源分離とともにフレーム外に及ぶ残響成分を除去する役割を担ったパラメータである．$\boldsymbol{W}^{\mathsf{H}}(f,0)$ は瞬時混合音源を分離する分離行列に対応する．$\boldsymbol{W}^{\mathsf{H}}(f,0)$ を正則であると仮定すると，式 (14) は

$$\boldsymbol{y}(f,n) = \boldsymbol{x}(f,n) - \sum_{n'=1}^{N'}\boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n'), \tag{15}$$

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\mathsf{H}}(f,0)\boldsymbol{y}(f,n) \tag{16}$$

のように書き直せる．ここで，$\boldsymbol{D}^{\mathsf{H}}(f,n') = -(\boldsymbol{W}^{\mathsf{H}}(f,0))^{-1}\boldsymbol{W}^{\mathsf{H}}(f,n')$, $(1\le n'\le N')$ であ

る．式 (15) は観測された混合信号 $\boldsymbol{x}(f,n)$ に含まれる残響成分を除去するプロセスに相当し，$\mathcal{D} = \{\boldsymbol{D}^{\mathsf{H}}(f,n')\}_{f,n'}$ は残響除去フィルタと見なせる．式 (16) は残響を除去した信号 $\boldsymbol{y}(f,n)$ に対して，周波数ごとの音源分離を行うプロセスとして解釈できる．この関係式に 2.1 節の $\boldsymbol{S}$ の生成モデルを組み込むことで，分離行列 $\mathcal{W}$，残響除去フィルタ $\mathcal{D}$ の尤度関数を得ることができる．観測信号 $\mathcal{X}$ が与えられたときの残響除去フィルタ $\mathcal{D}$，分離行列 $\mathcal{W}$，CVAE のデコーダ入力 Ψ 及びスケーリング係数 $\mathcal{G}$ についての尤度関数は以下のように記述できる．

$$\mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}) \stackrel{c}{=} -2N\log\left|\det \boldsymbol{W}^{\mathsf{H}}(f)\right| + \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{y}(f,n)|^2}{v_j(f,n)}\right) \tag{17}$$

3.2 先行研究

[9] では，時間周波数領域における畳み込み混合の形を ILRMA に導入し，分離フィルタ，音源モデルのパラメータ，残響除去フィルタを順に反復更新することでパーミュテーション整合，周波数ごとの音源分離と残響除去を同時に行う手法が提案されている．本稿ではこの手法を ILRMA+と呼ぶ．ILRMA+と我々の提案手法の相違点は音源 $s_j(f,n)$ の生成モデルの与え方にある．ILRMA+が通常の ILRMA と同様，$v_j(f,n)$ に NMF 型のモデルを用いるのに対し，提案手法では CVAE 音源モデルを用いている．

3.3 最適化アルゴリズム

本節では，観測信号 $\mathcal{X}$ が与えられたとき，残響除去フィルタ $\mathcal{D}$，分離行列 $\mathcal{W}$，CVAE のデコーダ入力 Ψ 及びスケーリング係数 $\mathcal{G}$ についての対数尤度関数式 (17) を探索するアルゴリズムについて述べる．この最適化問題の大域最適解は解析的に求めることはできないが，局所最適解は

$$\hat{\mathcal{D}} \leftarrow \operatorname*{argmin}_{\mathcal{D}} \mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}), \tag{18}$$

$$\hat{\mathcal{W}} \leftarrow \operatorname*{argmin}_{\mathcal{W}} \mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}), \tag{19}$$

$$\hat{\Psi} \leftarrow \operatorname*{argmin}_{\Psi} \mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}), \tag{20}$$

$$\hat{\mathcal{G}} \leftarrow \operatorname*{argmin}_{\mathcal{G}} \mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}) \tag{21}$$

を繰り返すことで数値探索することができる．まず，式 (17) において残響除去フィルタ $\mathcal{D}$ に関する項のみを考えると

$$\mathcal{I}(\mathcal{D}) = \sum_{f,n}\left|\boldsymbol{x}(f,n) - \sum_{n'=1}^{N'}\boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n')\right|^2_{\boldsymbol{\Sigma}_{w/v(f,n)}} \tag{22}$$

が得られる．ここで，$|x|_{\boldsymbol{\Sigma}_{w/v(f,n)}}$ は $\sqrt{\boldsymbol{x}^{\mathsf{H}}\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{x}}$ を表し，$\boldsymbol{\Sigma}_{w/v(f,n)} = \sum_j \frac{\boldsymbol{w}_j(f)\boldsymbol{w}_j^{\mathsf{H}}(f)}{v_j(f,n)}$ である．各 f に関して独立な更新式を得るため，$\boldsymbol{D}(f,n')$ の第 i 列のベクトルを $\boldsymbol{d}_i(f,n')$ として，$\{\boldsymbol{D}(f,n')\}_{n'}$ をベクトル形式に変形すると，

$$\boldsymbol{d}(f) = [\boldsymbol{d}_1^{\mathsf{T}}(f,1),\ldots,\boldsymbol{d}_I^{\mathsf{T}}(f,1),\boldsymbol{d}_1^{\mathsf{T}}(f,2),\ldots,\boldsymbol{d}_I^{\mathsf{T}}(f,2),\ldots,\boldsymbol{d}_1^{\mathsf{T}}(f,N'),\ldots,\boldsymbol{d}_I^{\mathsf{T}}(f,N')]^{\mathsf{T}} \in \mathbb{C}^{I^2N'} \tag{23}$$

が得られる．式 (23) を用いると，式 (22) は

$$\sum_{n'=1}^{N'}\boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n') = \boldsymbol{X}(f,n)\boldsymbol{d}^*(f) \tag{24}$$

のように書き直せる．ここで，$\boldsymbol{d}^*(f)$ は $\boldsymbol{d}(f)$ の複素共役であり，

$$\boldsymbol{X}(f,n) = [\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-1),\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-2),\ldots,\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-N')] \in \mathbb{C}^{I\times I^2N'} \tag{25}$$

である．$\boldsymbol{I}$ と $\otimes$ はそれぞれ $I\times I$ の単位行列とクロネッカー積を表す．式 (25) を式 (22) に代入することで次式が得られる．

$$\mathcal{I}(\mathcal{D}) = \sum_{f,n}\left(\boldsymbol{x}(f,n) - \boldsymbol{X}(f,n)\boldsymbol{d}^*(f)\right)^H \times \boldsymbol{\Sigma}_{w/v(f,n)}\left(\boldsymbol{x}(f,n) - \boldsymbol{X}(f,n)\boldsymbol{d}^*(f)\right) \tag{26}$$

式 (26) は $\partial\mathcal{I}/\partial\boldsymbol{d}^*(f) = 0$ を解析的に求めることができ，$\boldsymbol{d}^*(f)$ の更新式として

$$\boldsymbol{d}^*(f) \leftarrow \left(\sum_n \boldsymbol{X}^{\mathsf{H}}(f,n)\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{X}(f,n)\right)^{-1} \times \left(\sum_n \boldsymbol{X}^{\mathsf{H}}(f,n)\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{x}(f,n)\right) \tag{27}$$

が得られる．分離フィルタ $\mathcal{W}$ の更新は ILRMA や MVAE と同様に，式 (11)，式 (12) を用いて更新する．Ψ と $\mathcal{G}$ は MVAE と同様に，誤差逆伝搬法と式 (13) を用いて更新する．従って，CVAE 音源モデルの事前学習から各変数の最適化までの流れは以下のようにまとめられる．

1. 式 (10) に従い θ，ϕ の学習を行う．
2. Ψ，$\mathcal{G}$，$\mathcal{W}$ と $\mathcal{D}$ を初期化する．
3. 各 j，f について下記の更新を繰り返す．
   (a) 式 (11), (12) を用いた $\boldsymbol{w}_j(f)$ の更新．
   (b) 誤差逆伝搬法を用いた $\mathbf{z}_j$，c_j の更新．
   (c) 式 (13) を用いた g_j の更新．
   (d) 式 (27) を用いた $\boldsymbol{d}^*(f)$ の更新．

4 評価実験

提案手法の有効性を評価するため，既存手法の ILRMA，MVAE，ILRMA+ [9]，及び提案手法の 4 手法について，長い残響環境下における音声信号を対象とした 2 音源の分離性能を比較した．音声信号には VCC2018 [10] の男性 2 話者（SM1，SM2）と女性 2 話者（SF1，SF2）の発話データを用いた．各話者の発話データ 116 音源のうち 81 音源を CVAE の学習データとし，残りの 35 音源の中から 4 s 以上の長さの音源を評価データとして使用した．多チャンネル観測信号は RWCP データベース [11] に収録されている OFC，JR1 インパルス応答と発話データを畳み込むことで生成した．インパルス応答の残響時間（T_{60}）はそれぞれ 0.60 s と 0.78 s である．Fig. 1 に多チャンネル観測信号の生成に用いたインパルス応答におけるマイクと音源の位置関係を示す．話者の組み合わせは SF1 と

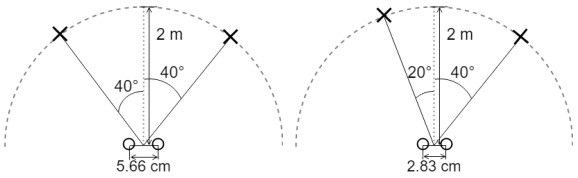

Fig. 1　実験に用いたインパルス応答におけるマイクと音源の位置関係．○はマイク位置，×は音源位置を表す．

SF2，SM1 と SM2，SM1 と SF1，SM2 と SF2 の 4 種類とし，各組み合わせの評価用データからランダムに 10 通り選ぶことで，各インパルス応答について合計 80 データの多チャンネル観測信号を作成した．音源のサンプリング周波数を 16 kHz とし，短時間 Fourier 変換の条件を窓長 256 ms，シフト量 64 ms として観測信号の複素スペクトログラムを計算した．ILRMA と ILRMA+の基底数を 20 とし，ILRMA+と提案手法の残響除去フィルタの次数 N' は JR1 を用いた音源では 3，OFC を用いた音源では 4 とした．ILRMA と ILRMA+は反復更新を 100 回行い，MVAE 法と提案手法は反復更新を 60 回行った．また，MVAE と提案手法では，ILRMA と ILRMA+を 30 回反復更新して得られた $\boldsymbol{W}^{\mathsf{H}}(f)$ と $\boldsymbol{D}^{\mathsf{H}}(f,n')$ を初期値として用いた．提案手法におけるエンコーダとデコーダのネットワーク構造は [5] と同様に，それぞれゲート付き畳み込み層 3 層とゲート付き逆畳み込み層 3 層で構築したネットワークを用いた．CVAE の学習及び分離時における Ψ の更新には Adam [12] を用いた．c_j の更新時に $\sum_j c_j = 1$ の制約を加えるため，誤差逆伝搬法により得られた出力に対してソフトマックス層を加えた．分離性能の客観評価尺度としては signal-to-distortion ratio(SDR)，signal-to-interference ratio(SIR)，及び signal-to-artifacts ratio(SAR) [13] を用いた．

各手法の SDR，SIR 及び SAR の平均改善量を Table 1 に示す．また，各条件における最良のスコアを太字で示す．結果より，提案手法の SDR，SIR 及び SAR の平均改善量が既存手法である ILRMA，ILRMA+及び MVAE を上回っており，時間周波数領域における畳み込み混合モデルと CVAE 音源モデルの両方を用いることによる，長い残響環境下での高精度な音源分離及び残響除去の同時実行が可能であることを確認した．

5　まとめ

本稿では，CVAE のデコーダ分布を音源の生成モデルに用いた音源分離手法である MVAE に対して残響除去を組み込んだ統合的な手法を提案した．提案手法は，分離系の時間周波数領域における畳み込みの形による表現と音源信号の CVAE 音源モデルによる表現に基づいた分離フィルタと残響除去フィルタの同時推定により，長い残響環境下での分離性能を向上させることができる．長い残響環境下における音源分離実験により，CVAE を用いた音源モデルの有効性と残響除去フィルタの同時推定の有効性を示した．

謝辞　本研究の一部は SECOM 科学技術支援財団，JSPS 科研費 17H01763 及び 18J20059 の助成を受けたものである．

Table 1　従来手法と提案手法による SDR，SIR 及び SAR の平均改善量 [dB]

RIRs	Methods	Improvement		
		SDR	SIR	SAR
JR1 $T_{60} = 0.6$ s	ILRMA	2.09	6.82	-1.17
	ILRMA+	4.53	9.72	1.02
	MVAE	3.63	10.68	-0.45
	proposed	**6.32**	**13.78**	**2.14**
OFC $T_{60} = 0.78$ s	ILRMA	1.90	6.47	-1.32
	ILRMA+	5.01	10.25	1.56
	MVAE	3.58	10.30	-0.49
	proposed	**6.74**	**14.29**	**2.51**

参考文献

[1] A. Hyvärinen *et al.*, “Independent component analysis: algorithms and applications”, Neural networks, Elsevier, vol. 13, no. 4, pp. 411–430, 2000.
[2] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique”, in Proc. WASPAA, pp. 189–192, 2011.
[3] H. Kameoka *et al.*, “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 *et al.*, “Determined blind source separation with independent low-rank matrix analysis”, *Audio Source Separation*, Springer, pp. 125–155, 2018.
[5] H. Kameoka *et al.*, “Semi-blind source separation with multichannel variational autoencoder”, arXiv preprint arXiv:1808.00892, 2018.
[6] D. D. Lee *et al.*, “Algorithms for non-negative matrix factorization”, in Proc. NIPS, pp. 556–562, 2001.
[7] D. P. Kingma *et al.*, “Semi-supervised learning with deep generative models”, in Proc. NIPS, pp. 3581–3589, 2014.
[8] T. Yoshioka *et al.*, “Blind separation and dereverberation of speech mixtures by joint optimization”, IEEE Trans. ASLP, vol. 19, no. 1, pp. 69–84, 2011.
[9] H. Kagami *et al.*, “Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization”, in Proc. ICASSP, pp. 31–35, 2018.
[10] J. Lorenzo-Trueba *et al.*, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods”, arXiv preprint arXiv:1804.04262, 2018.
[11] S. Nakamura *et al.*, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition”, in Proc. LREC, pp. 965–968, 2000.
[12] D. P. Kingma *et al.*, “Adam: A Mmthod for stochastic optimization”, in Proc. ICLR, 2015.
[13] E. Vincent *et al.*, “Performance measurement in blind audio source separation.”, IEEE Trans. ASLP, vol. 14, no. 4, pp. 1462–1469, 2006.