多チャンネル階乗隠れマルコフモデルのスペクトル包絡事前学習によるセミブラインド音源分離*

◎樋口卓哉 (東大院情報理工)，亀岡弘和 (東大院情報理工，NTT)

1 はじめに

ブラインド音源分離 (Blind Source Separation; BSS) の問題とは，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．一般的に BSS の問題を不良設定問題であり，解くためには，音源信号に対してなんらかの仮定を立て解を限定する必要がある．例えば，非負値行列因子分解 (Non-negative Matrix Factorization; NMF)[1, 2] では，観測信号のパワースペクトログラムが限られた数のパワースペクトルによって構成されていることを仮定することで，観測信号のパワースペクトログラムを 2 つの非負値行列の積に分解する．また筆者らは，音響イベント，残響，音源の到来方向などが BSS の問題を解くための手がかりと成りうることに着目し，多チャンネル階乗隠れマルコフモデルと呼ぶモデルを用いて，音響イベント，残響，音源の到来方向推定などを音源分離と統合的に行う手法を提案してきた [3, 4, 5]．この手法は，音源信号のパワースペクトルが限られた数のパワースペクトルの遷移によって表現できると仮定することで，BSS の問題において解を限定し，音源分離の手がかり (の一部) とする手法となっていた．しかし，NMF や多チャンネル階乗隠れマルコフモデルなどの，音源信号のスペクトルの数を制限するアプローチでは，音源信号をより詳細にモデル化しようとしてモデルにおけるパワースペクトルの数を増やせば増やすほど，BSS の問題において立てられていた仮定を弱め，解を限定することができなくなるというトレードオフが生じてしまう．そこで本稿では，[6] や [7] の手法のように，ソースフィルタモデルに基づいた新たな仮定を導入することを考える．ソースフィルタモデルとは，音声スペクトルを声帯駆動音源 (ソース) のスペクトルと声道 (フィルタ) スペクトルの積で表せるとしたモデルである．この仮定が成り立つならば，音声スペクトルは，複数個のソーススペクトルと複数個のフィルタスペクトルの組み合わせでで表すことができる．これにより，$Q^{(s)}$ 個のソーススペクトルと $Q^{(f)}$ 個のフィルタスペクトルを用いることで，$Q^{(s)}+Q^{(f)}$ の自由度で $Q^{(s)}\times Q^{(f)}$ 個のスペクトルを表現することができる．本稿では，ソースフィルタモデルに基づいて音源信号をモデル化し，フィルタスペクトルを事前学習することで，音源信号をより詳細にモデル化しながらも，BSS の問題における解を十分に限定し，音源分離性能を向上させる手法を提案する．

2 多チャンネル階乗隠れマルコフモデル

2.1 畳み込み混合近似による観測信号の混合モデル

I 個の音源から到来する信号を M 個のマイクロフォンで観測する場合を考える．室内インパルス応答長が時間周波数展開の時間窓長に対して十分に短いと限らず，瞬時混合近似が成り立たない場合 (残響がある場合) を考え，時間周波数領域における畳み込み混合近似によって時間周波数領域の観測信号を近似する．

$$\boldsymbol{y}(\omega_k,t_l)\approx\sum_{i=1}^{I}\sum_{\tau=0}^{\mathcal{T}}\boldsymbol{a}_i(\omega_k,t_\tau)s_i(\omega_k,t_l-t_\tau). \quad (1)$$

ただし，$\boldsymbol{y}(\omega_k,t_l)=(y_1(\omega_k,t_l),\ldots,y_M(\omega_k,t_l))^{\mathsf{T}}\in\mathbb{C}^M$，$\boldsymbol{a}_i(\omega_k,t_\tau)=(a_{i,1}(\omega_k,t_\tau),\ldots,a_{i,M}(\omega_k,t_\tau))^{\mathsf{T}}\in\mathbb{C}^M$ である．ここで $y_m(\omega_k,t_l)\in\mathbb{C}$ は m 番目のマイクで観測された観測信号の周波数 ω_k，時刻 t_l における時間周波数成分であり，$s_i(\omega_k,t_l)\in\mathbb{C}$ は i 番目の音源信号の周波数 ω_k，時刻 t_l における時間周波数成分である．また $\boldsymbol{a}_i(\omega_k,t_\tau)$ は i 番目の音源信号に対する周波数 ω_k における伝達周波数特性の時刻 t_τ の成分であり，$0\leq\tau\leq\mathcal{T}$ は伝達周波数特性の時間周波数領域における時間インデックスである．以下では ω_k，t_l をそれぞれ k，l の添え字で表す．

2.2 ソースフィルタモデルに基づく音源信号の生成モデル

では，音源信号の生成プロセスを確率的に記述する．まず音源信号が区分的に定常であることを仮定し，各時間周波数点で $s_{i,k,l}$ が平均 0，分散 $\sigma^2_{i,k,l}$ の複素正規分布に従うとすると，音源信号の生成プロセスは，

$$s_{i,k,l}|\sigma_{i,k,l}\sim\mathcal{N}_{\mathbb{C}}(s_{i,k,l};0,\sigma^2_{i,k,l}), \quad (2)$$

と書き表せる．ここで $\sigma^2_{i,k,l}$ は周波数 k，時刻 l における i 番目の音源のパワースペクトル密度を表す．上記のモデルに NMF の仮定を組み込むと，

$$\sigma^2_{i,k,l}=w_{i,k}h_{i,l}, \quad (3)$$

となる．上記の式による多チャンネル観測信号の生成モデルは，多チャンネル NMF[8] と呼ばれている．ここで，ソースフィルタモデルに基づき，i 番目の音源の周波数 k における基底パワースペクトルが，ソーススペクトル $w^{(s)}_{i,k}$ と，フィルタスペクトル $w^{(f)}_{i,k}$ の積によって構成されると仮定すると，

$$w_{i,k}=w^{(s)}_{i,k}w^{(f)}_{i,k}, \quad (4)$$

となる．しかしながら，ソーススペクトルやフィルタスペクトルは，発話内容によって時間変化すると考えられるので，ソーススペクトルの状態を表す隠れ変数 $z^{(s)}_{i,l}\in 1,\ldots,Q^{(s)}$，フィルタスペクトルの状態を表す隠れ変数 $z^{(f)}_{i,l}\in 1,\ldots,Q^{(f)}$ を導入し，それぞれの状態の時系列がマルコフ連鎖に従うと仮定する．

$$z^{(s)}_{i,l}|z^{(s)}_{i,l-1}\sim\mathrm{Categorical}(z^{(s)}_{i,l};\boldsymbol{\rho}^{(s)}_{i,z^{(s)}_{i,l-1}}), \quad (5)$$

$$z^{(f)}_{i,l}|z^{(f)}_{i,l-1}\sim\mathrm{Categorical}(z^{(f)}_{i,l};\boldsymbol{\rho}^{(f)}_{i,z^{(f)}_{i,l-1}}). \quad (6)$$

ここで $\mathrm{Categorical}(x;\boldsymbol{y})=y_x$ であり，$\boldsymbol{\rho}^{(s)}_{i,q^{(s)}}=(\rho^{(s)}_{i,q^{(s)},1},\ldots,\rho^{(s)}_{i,q^{(s)},Q^{(s)}})$ は i 番目の音源におけるソー

*Semi-blind source separation with multichannel factorial hidden Markov model using pretrained spectral envelope templates. by HIGUCHI, Takuya (The University of Tokyo), KAMEOKA Hirokazu (The University of Tokyo, NTT)

スの状態 $q^{(s)}$ から各状態 $1,\ldots,Q^{(s)}$ への遷移確率を表し，$\boldsymbol{\rho}^{(f)}_{i,q^{(f)}} = (\rho^{(f)}_{i,q^{(f)},1},\ldots,\rho^{(f)}_{i,q^{(f)},Q^{(f)}})$ は i 番目の音源におけるフィルタの状態 $q^{(f)}$ から各状態 $1,\ldots,Q^{(f)}$ への遷移確率を表す．状態 $q^{(s)}$ である i 番目の音源のソーススペクトルを $w^{(s)}_{i,k,q^{(s)}}$，状態 $q^{(f)}$ である i 番目の音源のフィルタスペクトルを $w^{(f)}_{i,k,q^{(f)}}$ と表すとすると，時刻 l における i 番目の音源信号のソーススペクトルとフィルタスペクトルは $z^{(s)}_{i,l}$ と $z^{(f)}_{i,l}$ にそれぞれ依存し，$s_{i,k,l}$ の生成モデルは以下のように書き直せる．

$$\begin{aligned}\boldsymbol{s}_{i,k,l}|w^{(s)}_{i,k,1:Q^{(s)}}, w^{(f)}_{i,k,1:Q^{(f)}}, h_{i,l}, z^{(s)}_{i,l}, z^{(f)}_{i,l}\\ \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{i,k,l}; 0, w^{(s)}_{i,k,z^{(s)}_{i,l}} w^{(f)}_{i,k,z^{(f)}_{i,l}} h_{i,l}).\end{aligned} \tag{7}$$

次に，音源の音量に着目すると，無音状態と有音状態では当然音量の大きな値の取りやすさが異なると考えられるので，音量もまた音源の状態 (音響イベント) に依存して異なる振る舞いをするといえる．そこで，まず音量の状態を表す隠れ変数 $z^{(j)}_{i,l} \in 1,\ldots,J$ を導入し，状態の時系列 $z^{(j)}_{i,1},\ldots,z^{(j)}_{i,L}$ がマルコフ連鎖に従うと仮定する．

$$z^{(j)}_{i,l}|z^{(j)}_{i,l-1} \sim \mathrm{Categorical}(z^{(j)}_{i,l}; \boldsymbol{\rho}^{(j)}_{i,z^{(j)}_{i,l-1}}). \tag{8}$$

このとき，$h_{i,l}$ が音量の状態 $z^{(j)}_{i,l} \in 1,\ldots,J$ によって異なるハイパーパラメータを持つガンマ分布に従うと仮定すると，

$$h_{i,l}|z^{(j)}_{i,l} \sim \mathrm{Gamma}(h_{i,l}; \alpha_{z^{(j)}_{i,l}} \beta_{z^{(j)}_{i,l}}), \tag{9}$$

となる．ここで $\alpha_{1:J}$ と $\beta_{1:J}$ はそれぞれガンマ分布の形状パラメータとスケールパラメータであり，$\mathrm{Gamma}(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$，ただし $\Gamma(\cdot)$ はガンマ関数である．$z^{(j)}_{i,l}$ が無音状態に対応するときは $h_{i,l}$ は小さな値をとってほしいので，小さな値をとる確率が高くなるようにガンマ分布のハイパーパラメータを設定し，$z^{(j)}_{i,l}$ が有音状態に対応するときは一様分布に近くなるようにガンマ分布のハイパーパラメータを設定すればよい．

2.3 混合 DOA モデルによる空間相関行列の生成モデル

次に空間相関行列の生成プロセスを確率的にモデル化する．点音源と平面波到来を仮定すると，空間相関行列は音源の到来方向に応じてある特定の構造を持つ．例えば，マイクロフォンの数 $M = 2$ の場合では，方向 $\theta(0 \le \theta \le \pi)$ にある音源の空間相関行列は，以下のように陽に記述できる．

$$\boldsymbol{J}(\theta,\omega) = \begin{bmatrix} 1 \\ e^{\mathrm{J}\omega B\cos\theta/C} \end{bmatrix} \begin{bmatrix} 1 & e^{\mathrm{J}\omega B\cos\theta/C} \end{bmatrix}^{*} \tag{10}$$

ここで J は虚数単位，B [m] はマイクロフォン間の距離，C [m/s] は音速である．i 番目の音源の到来方向 θ_i が既知の場合では，直接波に対応する空間相関行列は $\boldsymbol{J}(\theta_i,\omega_k)$ と等しくなることが期待される．しかしながら実際には，音源の到来方向は音響信号から直接観測できないばかりでなく，フレーム内の残響やノイズなどによって，空間相関行列は理想的な構造から逸脱することがある．

そこでまず，[9] の手法と同様に，$\theta_i \in \{\vartheta_1,\ldots,\vartheta_O\}$ を i 番目の音源の到来方向とし，各音源の DOA がこの DOA 候補値の中から決定されると仮定することで，音源 i の到来方向 θ_i が生成されるプロセスを以下のように記述する。

$$z^{(o)}_i|\boldsymbol{\rho}^{(o)}_i \sim \mathrm{Categorical}(z^{(o)}_i; \boldsymbol{\rho}^{(o)}_i), \tag{11}$$

$$\theta_i = \vartheta_{z^{(o)}_i}. \tag{12}$$

$z^{(o)}_i \in \{1,\ldots,O\}$ は i 番目の音源にどの DOA 候補値が割り当てられるかを表すインジケータ変数であり，上式はこれが離散分布 (各確率値が $\rho^{(o)}_1,\ldots,\rho^{(o)}_O$) から生成されることを意味している。そして，直接波に対応する空間相関行列 $\boldsymbol{C}_{i,k,0}$ が，$z^{(o)}_i$ が既知の条件下で，以下のようなウィッシャート分布に従うと仮定する．

$$\boldsymbol{C}_{i,k,0}|z^{(o)}_i \sim \mathcal{W}_{\mathbb{C}}(\boldsymbol{C}_{i,k,0}; \gamma, \boldsymbol{J}_{\vartheta_{z^{(o)}_i},k} + \epsilon\boldsymbol{I}). \tag{13}$$

ただし $\mathcal{W}_{\mathbb{C}}(\boldsymbol{X};\gamma,\boldsymbol{Y}) \propto |\boldsymbol{X}|^{(\gamma-M)/2}\exp(-\frac{1}{2}\mathrm{tr}(\boldsymbol{X}\boldsymbol{Y}^{-1}))$ であり，γ は $\gamma \ge M+1$ を満たすハイパーパラメータである．ここでは，逆行列演算を可能にするために，近似的に $\boldsymbol{J}_{\vartheta_{z^{(o)}_i},k}$ に微小値 ϵ を用いて単位行列 $\boldsymbol{I}$ を足してある．ここで隠れ変数 $z^{(o)}_i$ を推定することが，音源の到来方向推定を行うことに相当している．

観測信号の最終的な生成モデルは $\boldsymbol{a}_{1:I,k,0:\mathcal{T}}$, $w^{(s)}_{1:I,k,1:Q^{(s)}}$, $w^{(f)}_{1:I,k,1:Q^{(f)}}$, $h_{1:I,l-\mathcal{T}:l}$, $z^{(s)}_{1:I,l-\mathcal{T}:l}$, $z^{(f)}_{1:I,l-\mathcal{T}:l}$ が既知の条件下で，式 (5)，式 (6)，式 (8)，式 (9)，式 (13) と合わせて以下のように書き直せる．

$$\begin{aligned}\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k,0:\mathcal{T}}, w^{(s)}_{1:I,k,1:Q^{(s)}}, w^{(f)}_{1:I,k,1:Q^{(f)}},\\ h_{1:I,l-\mathcal{T}:l}, z^{(s)}_{1:I,l-\mathcal{T}:l}, z^{(f)}_{1:I,l-\mathcal{T}:l}\\ \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; 0, \sum_{i,\tau} \boldsymbol{C}_{i,k,\tau} w^{(s)}_{i,k,z^{(s)}_{i,l}} w^{(f)}_{i,k,z^{(f)}_{i,l}} h_{i,l-\tau}).\end{aligned} \tag{14}$$

この生成モデルに基づいて最適なパラメータを求めることは，ソースフィルタモデルによって音源信号の自由度を限定しながら，音源の到来方向推定・残響除去・音響イベント検出・音源分離の問題を統合的に解くことに相当している．

3 補助関数法に基づくパラメータ推論アルゴリズム

目的関数である対数事後確率では各パラメータが複雑に関係し合っており，一般的に最適化が困難である．そこで効率的に最適なパラメータを求める反復アルゴリズムを，補助関数法の原理に基づき導出する．モデルにおける推定したい変数は $\boldsymbol{W}^{(s)} = w^{(s)}_{1:I,1:K,1:Q^{(s)}}$, $\boldsymbol{W}^{(f)} = w^{(f)}_{1:I,1:K,1:Q^{(f)}}$, $\boldsymbol{H} = h_{1:I,1:L}$, $\boldsymbol{C} = \boldsymbol{C}_{1:I,1:K,0:\mathcal{T}}$ である．上記の変数の集合を $\boldsymbol{\Theta}$ で表す．$\boldsymbol{Z}^{(j)} = z^{(j)}_{1:I,1:L}$, $\boldsymbol{Z}^{(o)} = z^{(o)}_{1:I}$, $\boldsymbol{Z}^{(s)} = z^{(s)}_{1:I,1:L}$, $\boldsymbol{Z}^{(f)} = z^{(f)}_{1:I,1:L}$ は隠れ変数とする．以下では $\boldsymbol{\rho}^{(s)}$, $\boldsymbol{\rho}^{(f)}$, $\boldsymbol{\rho}^{(j)}$, $\boldsymbol{\rho}^{(o)}$ は実験的に定められた定数とする．補助変数の集合を $\boldsymbol{\Lambda}$ とすると，目的関数 $L(\boldsymbol{\Theta}) = \log p(\boldsymbol{\Theta}|\boldsymbol{Y})$ に対する補助関数 $L^{+}(\boldsymbol{\Theta},\boldsymbol{\Lambda})$ は以下のように設計できる．

$$L^{+}(\boldsymbol{\Theta},\boldsymbol{\Lambda})$$

$$
\begin{aligned}
&= -\frac{1}{2}\sum_{k,l}\Bigg[\sum_{i,q^{(s)},q^{(f)},\tau}\lambda^{(s)}_{q^{(s)},i,l-\tau}\lambda^{(f)}_{q^{(f)},i,l-\tau}\Bigg(\mathrm{tr}\Big(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\\
&\quad \boldsymbol{R}_{i,k,l,\tau,q^{(s)},q^{(f)}}\boldsymbol{C}^{-1}_{i,k,\tau}\\
&\quad \boldsymbol{R}_{i,k,l,\tau,q^{(s)},q^{(f)}}\Big)/w^{(s)}_{i,k,q^{(s)}}w^{(f)}_{i,k,q^{(f)}}h_{i,l-\tau}\\
&\quad +\mathrm{tr}\Big(\boldsymbol{U}^{-1}_{k,l}\boldsymbol{C}_{i,k,\tau}\Big)w^{(s)}_{i,k,q^{(s)}}w^{(f)}_{i,k,q^{(f)}}h_{i,l-\tau}\Bigg)+\log|\boldsymbol{U}_{k,l}|\Bigg]\\
&\quad +\sum_{j,i,l}\lambda^{(j)}_{j,i,l}\Big((\alpha_j-1)\log h_{i,l}-h_{i,l}/\beta_j\\
&\quad -\alpha_j\log\beta_j-\log\Gamma(\alpha_j)\Big)\\
&\quad +\sum_{i,k,o}d_{i,o}\Big[-\frac{1}{2}\mathrm{tr}\Big(\boldsymbol{C}_{i,k,0}(\boldsymbol{J}_{k,o}+\epsilon\boldsymbol{I})^{-1}\Big)\\
&\quad +\frac{\gamma-M}{2}\Big(\mathrm{tr}\Big(\tilde{\boldsymbol{U}}^{-1}_{i,k}\boldsymbol{C}_{i,k,0}\Big)+\log|\tilde{\boldsymbol{U}}_{i,k}|\Big)\Big]\\
&\quad +\sum_{q^{(s)},i,l}\lambda^{(s)}_{q^{(s)},i,l}\log p(\boldsymbol{Z}^{(s)})+\sum_{q^{(f)},i,l}\lambda^{(f)}_{q^{(f)},i,l}\log p(\boldsymbol{Z}^{(f)})\\
&\quad +\sum_{j,i,l}\lambda^{(j)}_{j,i,l}\log p(\boldsymbol{Z}^{(j)})+C_{L^+},
\end{aligned}
\tag{15}
$$

ここで $\boldsymbol{R}_{i,k,l,\tau,q^{(s)},q^{(f)}}$, $\boldsymbol{U}_{k,l}$, $\tilde{\boldsymbol{U}}_{i,k}$ は $\sum_{i,\tau}\boldsymbol{R}_{i,k,l,\tau,q^{(s)},q^{(f)}}=\boldsymbol{I}$ を満たすエルミート正定値行列であり，$\lambda^{(s)}_{q^{(s)},i,l}$, $\lambda^{(f)}_{q^{(f)},i,l}$, $\lambda^{(j)}_{j,i,l}$, $d_{i,o}$ は $\sum_{q^{(s)}}\lambda^{(s)}_{q^{(s)},i,l}=1$, $\sum_{q^{(f)}}\lambda^{(f)}_{q^{(f)},i,l}=1$, $\sum_j\lambda^{(j)}_{j,i,l}=1$, $\sum_o d_{i,o}=1$ をそれぞれ満たす非負値のスカラー値である．C_{L^+} は定数項をまとめたものである．式 (15) の等号成立条件は，$\boldsymbol{R}_{i,k,l,\tau,q^{(s)},q^{(f)}}$ と $\boldsymbol{U}_{k,l}$ に関しては

$$
\boldsymbol{R}_{i,k,l,\tau,q^{(s)},q^{(f)}}=\boldsymbol{C}_{i,k,\tau}w^{(s)}_{i,k,q^{(s)}}w^{(f)}_{i,k,q^{(f)}}h_{i,l-\tau}\hat{\boldsymbol{X}}^{-1}_{k,l},
\tag{16}
$$

$$
\boldsymbol{U}_{k,l}=\hat{\boldsymbol{X}}_{k,l},
\tag{17}
$$

である．ただし $\hat{\boldsymbol{X}}_{k,l}=\sum_{i,\tau}\boldsymbol{C}_{i,k,\tau}w^{(s)}_{i,k,z^{(s)}_{i,l-\tau}}w^{(f)}_{i,k,z^{(f)}_{i,l-\tau}}h_{i,l-\tau}$ である．$\lambda^{(s)}_{q^{(s)},i,l}$ の等号成立条件に関しては Forward-Backward アルゴリズムを用いて，

$$
\lambda^{(s)}_{q^{(s)},i,l}=F^{(s)}_{q^{(s)},i,l}B^{(s)}_{q^{(s)},i,l}/\sum_{q^{(s)}}F^{(s)}_{q^{(s)},i,l}B^{(s)}_{q^{(s)},i,l},
\tag{18}
$$

$$
F^{(s)}_{q^{(s)},i,l}=p(\boldsymbol{\Theta}|z^{(s)}_{i,l}=q^{(s)})\sum_{q^{(s)\prime}}F^{(s)}_{q^{(s)},i,l-1}\rho^{(s)}_{i,q^{(s)\prime},q^{(s)}},
\tag{19}
$$

$$
B^{(s)}_{q^{(s)},i,l}=\sum_{q^{(s)\prime}}B^{(s)}_{q^{(s)\prime},i,l+1}p(\boldsymbol{\Theta}|z^{(s)}_{i,l+1}=q^{(s)\prime})\rho^{(s)}_{i,q^{(s)},q^{(s)\prime}},
\tag{20}
$$

ただし

$$
\begin{aligned}
&p(\boldsymbol{\Theta}|z^{(s)}_{i,l}=q^{(s)})\\
&\propto\ \exp\Bigg[-\frac{1}{2}\sum_{k,\tau,q^{(f)}}\lambda^{(f)}_{q^{(f)},i,l}\Bigg(\mathrm{tr}\Big(\boldsymbol{y}_{k,l+\tau}\boldsymbol{y}_{k,l+\tau}{}^{\mathsf{H}}\\
&\quad \boldsymbol{R}_{i,k,l+\tau,\tau,q^{(s)},q^{(f)}}\boldsymbol{C}^{-1}_{i,k,\tau}
\end{aligned}
$$

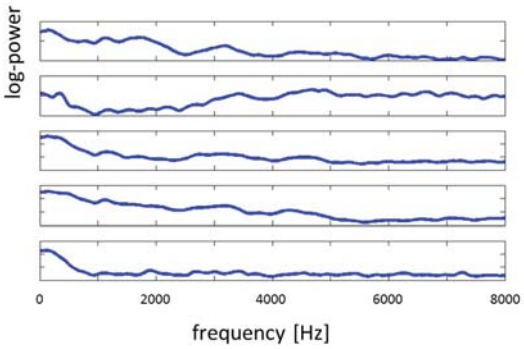

Fig. 1　クリーンな音声から STRAIGHT[12] によって抽出し実験に用いたフィルタスペクトルの例.

$$
\begin{aligned}
&\quad \boldsymbol{R}_{i,k,l+\tau,\tau,q^{(s)},q^{(f)}}\Big)/w^{(s)}_{i,k,q^{(s)}}w^{(f)}_{i,k,q^{(f)}}h_{i,l}\\
&\quad +\mathrm{tr}\Big(\boldsymbol{U}^{-1}_{k,l+\tau}\boldsymbol{C}_{i,k,\tau}\Big)w^{(s)}_{i,k,q^{(s)}}w^{(f)}_{i,k,q^{(f)}}h_{i,l}\Bigg)\Bigg],
\end{aligned}
\tag{21}
$$

である．さらに $\lambda^{(f)}_{q^{(f)},i,l}$ の等号成立条件に関しても Forward-Backward アルゴリズムを用いて，

$$
\lambda^{(f)}_{q^{(f)},i,l}=F^{(f)}_{q^{(f)},i,l}B^{(f)}_{q^{(f)},i,l}/\sum_{q^{(f)}}F^{(f)}_{q^{(f)},i,l}B^{(f)}_{q^{(f)},i,l},
\tag{22}
$$

$$
F^{(f)}_{q^{(f)},i,l}=p(\boldsymbol{\Theta}|z^{(f)}_{i,l}=q^{(f)})\sum_{q^{(f)\prime}}F^{(f)}_{q^{(f)},i,l-1}\rho^{(f)}_{i,q^{(f)\prime},q^{(f)}},
\tag{23}
$$

$$
B^{(f)}_{q^{(f)},i,l}=\sum_{q^{(f)\prime}}B^{(f)}_{q^{(f)\prime},i,l+1}p(\boldsymbol{\Theta}|z^{(f)}_{i,l+1}=q^{(f)\prime})\rho^{(f)}_{i,q^{(f)},q^{(f)\prime}},
\tag{24}
$$

ただし

$$
\begin{aligned}
&p(\boldsymbol{\Theta}|z^{(f)}_{i,l}=q^{(f)})\\
&\propto\ \exp\Bigg[-\frac{1}{2}\sum_{k,\tau,q^{(s)}}\lambda^{(s)}_{q^{(s)},i,l}\Bigg(\mathrm{tr}\Big(\boldsymbol{y}_{k,l+\tau}\boldsymbol{y}_{k,l+\tau}{}^{\mathsf{H}}\\
&\quad \boldsymbol{R}_{i,k,l+\tau,\tau,q^{(s)},q^{(f)}}\boldsymbol{C}^{-1}_{i,k,\tau}\\
&\quad \boldsymbol{R}_{i,k,l+\tau,\tau,q^{(s)},q^{(f)}}\Big)/w^{(s)}_{i,k,q^{(s)}}w^{(f)}_{i,k,q^{(f)}}h_{i,l}\\
&\quad +\mathrm{tr}\Big(\boldsymbol{U}^{-1}_{k,l+\tau}\boldsymbol{C}_{i,k,\tau}\Big)w^{(s)}_{i,k,q^{(s)}}w^{(f)}_{i,k,q^{(f)}}h_{i,l}\Bigg)\Bigg],
\end{aligned}
\tag{25}
$$

である．$\tilde{\boldsymbol{U}}_{i,k}$, $d_{i,o}$, $\lambda^{(j)}_{j,i,l}$ の等号成立条件は，[5] と同様である．紙面の都合上詳細は省略するが，その他のパラメータの更新則は L^+ の偏微分を 0 と置くことで導ける．

4　評価実験

4.1　音源分離性能の評価実験

提案法の音源分離性能の評価のために，残響下での半教師あり音源分離実験を行った．ATR 音声データベース [10] の中の 3 人の発話者 (女性 2 人，男性 1 人) による 15 種類の発話に対して，RWCP データベース [11] のインパルス応答 (残響時間 380 ms，マイク間距離 11.48 cm，マイクの数 $M=2$) を畳み込み，人工

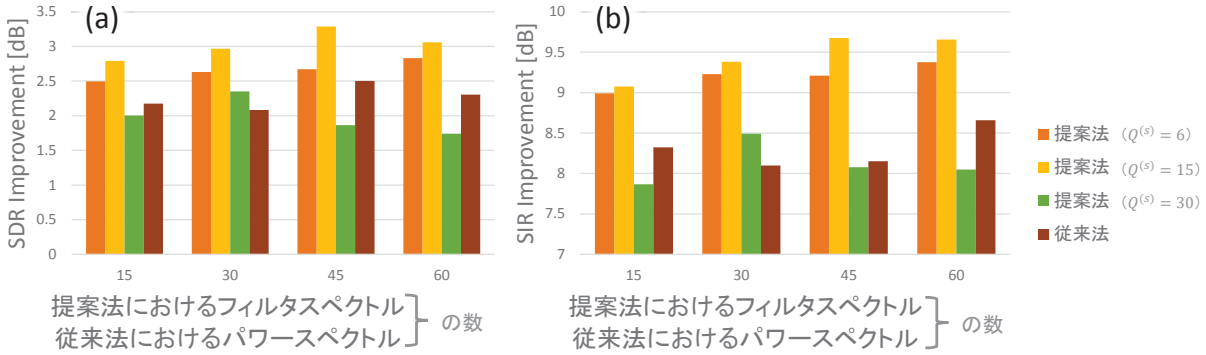

Fig. 2　提案法におけるフィルタスペクトルの数や従来法 [5] における基底パワースペクトルの数を変えて実験を行ったときに得られた分離音の (a)SDR の改善量と (b)SIR の改善量.

的に多チャンネルの混合信号を 5 つ作成した．3 つの音源はインパルス応答を畳み込まれることでマイクから 30°，90°，130° の位置に人工的にそれぞれ配置された．サンプリング周波数は 16 kHz とした．フレーム長 64 ms，フレームシフト長 16 ms で STFT を行い，時間周波数展開を行った．ウィッシャート分布のハイパーパラメータ γ は 3 とした．ソーススペクトルの状態数 $Q^{(s)}$ を 6，15，30，フィルタスペクトルの状態数 $Q^{(f)}$ を 15，30，45，60 とそれぞれ変えて実験した．$\epsilon = 10^{-3}$，到来方向の分割数 $O = 30$ とした．音量の状態数 $J = 2$ とし，α_1 と β_1 を 1，10^{-3} とそれぞれ設定し，α_2 と β_2 を 1 と 10^{10} と設定することで，$j = 1$ を無音状態とみなした．$\boldsymbol{\rho}^{(j)}$ は $i = 1, \ldots, 3$ において $\boldsymbol{\rho}_{i,1}^{(j)} = (0.9, 0.1)$，$\boldsymbol{\rho}_{i,2}^{(j)} = (0.1, 0.9)$ とし，自己遷移の確率を高めに設定した．$\mathcal{T} = 3$ とした．$\boldsymbol{C}$ の初期値は，$\tau = 0$ の成分に関しては $1/\sqrt{M} \times \boldsymbol{I}$ とし，$\tau = 1 : \mathcal{T}$ では $10^{-2} \times \boldsymbol{I}$ とした．$\boldsymbol{H}$ の初期値はすべての要素を 1 として与えた．$\lambda^{(q)}$ の初期値は $1/Q$，$\lambda^{(j)}$ の初期値は $1/J$ とした．$\boldsymbol{W}^{(f)}$ は，STRAIGHT[12] を用いて混合音声とは別の発話から抽出したスペクトル包絡に対して HMM 学習することで 3 人の話者の発話から $Q^{(f)}/3$ つずつ学習し，固定した．Fig. 1 に学習したフィルタスペクトルの一例を示す．$\boldsymbol{W}^{(s)}$ は，STRAIGHT を用いて抽出したスペクトル包絡系列でスペクトログラムを割ったものをソース成分と見なし HMM 学習をしたものを初期値とし，反復アルゴリズムにより更新した．$\boldsymbol{\rho}^{(s)}$，$\boldsymbol{\rho}^{(f)}$ は一様とした．d の初期値は到来方向を三等分するように設定した．また，θ_i の取りうる範囲を $\vartheta_{1+10(i-1)}$ から ϑ_{10i} に限定し，各音源ごとに推定する到来方向が重ならないように $\boldsymbol{\rho}^{(o)}$ を設定した．パラメータ推論アルゴリズムは 50 回反復した．前章と同様に望ましくない局所解を避けるため，最初の 25 回の反復は $\mathcal{T} = 0$ とし，その後徐々に $\mathcal{T}$ を 3 となるまで増やしながら反復した．また，最初の 25 回の反復は $\boldsymbol{W}$ を更新し，その後 $\mathcal{T}$ を増やし $\mathcal{T} > 0$ となってからは $\boldsymbol{W}$ を更新せずに固定した．比較対象には [5] の手法を用いた．比較手法では，分離に用いる混合音とは別の発話から学習した基底パワースペクトルを $\boldsymbol{W}$ の初期値として用いた．また基底パワースペクトルの状態数 Q を 15，30，45，60 とそれぞれ変えて実験した．分離音は多チャンネルウィーナーフィルタによって得た．音源分離，残響除去の客観評価基準として，残響除去済み分離音の振幅スペクトログラムに対して，室内インパルス応答を畳み込む前の音源信号の振幅スペクトログラムを参照することで得られた SDR，SIR[13] を用いた．分離処理前の SDR，SIR はそれぞれ-4.77, -0.96[dB] であった．

Fig. 2 に，本実験によって得られた分離音の SDR，SIR の改善量をそれぞれ示す．提案法におけるソーススペクトルやフィルタスペクトルの数，従来法における基底パワースペクトルの数によって変動してはいるが，提案法によって得られた分離音の客観評価値は，従来法によって得られた分離音の客観評価値を上回る傾向が見られた．特に $Q^{(s)} = 6, 15$ のときの提案法を用いて得られた分離音は，従来法によって得られた分離音より総じて高い SDR と SIR を示した．SDR は $Q^{(s)} = 15$，$Q^{(f)} = 45$ のとき最も高い値を示した．これにより，音源分離における提案法の有効性が示された．

5　おわりに

本稿では，ソースフィルタモデルに基づいて音源信号をモデル化し，多チャンネル階乗隠れマルコフモデルに組み込むことによって，従来法を上回る性能を実現する音源分離手法を提案した．

謝辞　本研究は JSPS 科研費 26730100 の助成を受けたものです．

参考文献

[1] D. D. Lee, and H. S. Seung, *Nature*, vol. 401, pp.788–791, 1999.
[2] P. Smaragdis, and J. C. Brown, *WASPAA 2003*, pp. 177–180, Oct. 2003.
[3] T. Higuchi, *et al.*, *Interspeech 2014*, pp. 850–854, 2014.
[4] T. Higuchi and H. Kameoka, "Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model," *MLSP 2014*.
[5] 樋口 他, "多チャンネル階乗隠れマルコフモデルと混合 DOA モデルによる音源分離・到来方向推定・音響イベント検出・残響除去の統合的アプローチ," 電子情報通信学会技術研究報告, 電気音響, 2015 [to appear].
[6] H. Kameoka and K. Kashino, ISCAS2009, pp. 2477–2480, May 2009.
[7] 吉井和佳, 後藤真孝, 情報処理学会研究報告, 2012–MUS–96–8, Aug. 2012.
[8] H. Sawada *et al.*, *ICASSP 2012*, pp. 261–264, 2012.
[9] 亀岡 他, 音講論 (春), 1–1–19, pp.713-716, Mar. 2012.
[10] A. Kurematsu *et al.*, *Speech Communication*, pp. 357–363, 1990.
[11] S. Nakamura *et al.*, *LREC 2000*, pp. 965–968, 2000.
[12] H. Kawahara *et al*, *Speech Communication,* vol. 27, no. 3–4, pp. 187–207, 1999.
[13] E. Vincent *et al.*, *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1462–1469, 2006.