多チャンネル Factorial hidden Markov model による音源分離・残響除去・音響イベント検出の統合的アプローチ *

◎樋口卓哉 (東大院情報理工)，亀岡弘和 (東大院情報理工，NTT CS 研)

1 はじめに

ブラインド音源分離の問題とは，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，ブラインド音源分離の問題を解くためには，音源信号に対して立てたなんらかの仮定を基に最適化基準を立て，最適化問題を解く必要がある．例えば，観測信号数が音源数よりも多い優決定問題では，音源信号間の独立性を仮定して分離する独立成分分析 (Independent Component Analysis; ICA) が有用であることが知られており，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が目的となる [1]。しかし，ICA では観測信号数が音源数よりも少ない劣決定問題を扱うことはできず，この場合は独立性よりもさらに強い仮定が必要である。

単チャンネルの観測信号に対するブラインド音源分離の有効なアプローチとして，非負値行列因子分解 (Non-negative Matrix Factorization; NMF) が知られている [2, 3]．この手法では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．分解した各行列は，いくつかの基底パワースペクトルによって構成される基底行列と，それらの基底パワースペクトルの時変な音量を表すアクティベーションによって構成されるアクティベーション行列となる．また，音源信号の空間的な情報も利用して音源分離を行うために，NMF を多チャンネルの音響信号へと拡張するアプローチがいくつか取られてきた [4, 5]．

しかし以上のアプローチでは，各音源信号のパワースペクトルが，1 つの基底パワースペクトルによって表現できることを仮定していた．しかし実際の音源信号のパワースペクトルは時変であることが多く，1 つの基底パワースペクトルで表現するのは不十分な場合が多い．また音量に着目してみても，無音状態，有音状態では大きな値のとりやすさが異なると考えられる．そこで我々は以前，音源信号の時変な性質を，音源の状態を隠れ変数とする隠れマルコフモデル (hidden Markov model; HMM) によってモデル化し，パラメータ推定を通してブラインド音源分離と音響イベント検出を統合的に行う手法を提案した [6]．

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，残響がある場合を考慮していない．従って従来の手法をそのまま利用して，残響を考慮した精度良い音源分離を行うためには，従来法とは別に，残響除去を行う必要がある．すなわち，音源分離性能は残響除去の精度に依存することになる．一方で，実環境においてどのように残響がかかるかは各音源信号によって異なるため，残響除去の精度は音源分離性能に依存する．

このような観点から，残響除去の問題と音源分離の問題は相互依存の関係にあると考えられ，本来同時に解かれるべきであるといえる．そこで本稿では，従来のモデル [6] 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，パラメータ推定を通して音源分離・音響イベント検出・残響除去を統合的に行う手法を提案する．本稿ではこのモデルを多チャンネル factorial hidden Markov model と呼ぶ．

2 多チャンネル factorial HMM

2.1 混合モデル

まず，観測信号の生成プロセスについて述べる．I 個の音源信号が M 個のマイクロフォンで観測される場合を考える．$y_m(t) \in \mathbb{R}$ を m 番目のマイクで観測される観測信号，$s_i(t) \in \mathbb{R}$ を i 番目の音源信号だとすると，観測信号は時間領域において以下のように記述できる．

$$\boldsymbol{y}(t) = \sum_{i=1}^{I} \sum_{\tau=0}^{\infty} \boldsymbol{a}_i(\tau) s_i(t-\tau). \tag{1}$$

ここで $\boldsymbol{y}(t) = (y_1(t), \ldots, y_M(t))^{\mathsf{T}} \in \mathbb{R}^M$，$\boldsymbol{a}_i(t) = (a_{i,1}(t), \ldots, a_{i,M}(t))^{\mathsf{T}} \in \mathbb{R}^M$ である．$a_{i,m}(t)$ は i 番目の音源と m 番目のマイクの間の室内インパルス応答を表す．ここで，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，瞬時混合近似を用いて観測信号は以下のように時間周波数領域において記述できる．

$$\boldsymbol{y}(\omega_k, t_l) = \sum_{i=1}^{I} \boldsymbol{a}_i(\omega_k) s_i(\omega_k, t_l), \tag{2}$$

ただし，$\boldsymbol{y}(\omega_k, t_l) = (y_1(\omega_k, t_l), \ldots, y_M(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^M$，$\boldsymbol{a}_i(\omega_k) = (a_{i,1}(\omega_k), \ldots, a_{i,M}(\omega_k))^{\mathsf{T}} \in \mathbb{C}^M$ 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．$\boldsymbol{a}_i(\omega_k)$ 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を表す．$1 \leq k \leq K$ と $1 \leq l \leq L$ はそれぞれ時間周波数領域における周波数と時間のインデックスである．しかし残響がある場合には一般に，室内インパルス応答長は時間窓長に対して十分に短いとはいえず，式 (2) の瞬時混合近似は成り立たない．そこで，時間周波数領域における畳み込み混合の形で観測信号を近似する．

$$\boldsymbol{y}(\omega_k, t_l) \approx \sum_{i=1}^{I} \sum_{n=0}^{N} \boldsymbol{a}_i(\omega_k, t_n) s_i(\omega_k, t_l - t_n). \tag{3}$$

ここで $\boldsymbol{a}_i(\omega_k, n)$ 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_n の成分であり，$0 \leq n \leq N$ 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．ここで $\boldsymbol{a}_i(\omega_k, t_n)$ は i 番目の音源信号が時間周波数領域において t_n だけ先の時刻にどれだけ影響を与えるか，すなわち時間周波数領域においてどれだけ残響がかかるかを表している．表記の都合上，以下では ω_k，t_l，t_n を k，l，n の添え字でそれぞれ表す．

2.2 観測信号の生成プロセス

次に，式 (3) に基づいて観測信号の生成プロセスを確率的に記述する．まず，音源信号 $s_{i,k,l}$ が平均 0，分散 $\sigma^2_{i,k,l}$ の複素正規分布に従うと仮定する．

$$s_{i,k,l} | \sigma_{i,k,l} \sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l}; 0, \sigma^2_{i,k,l}). \tag{4}$$

---

* Unified approach for source separation, dereverberation and source activity detection based on multi-channel factorial hidden Markov model. by HIGUCHI, Takuya (The University of Tokyo), KAMEOKA Hirokazu (The University of Tokyo, NTT)

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．式 (3) と式 (4) から，$\boldsymbol{a}_{1:I,k,0:N}$ と $\sigma_{1:I,k,l-N:l}$ が既知の条件下で観測信号 $\boldsymbol{y}_{k,l}$ は同じく複素正規分布に従う．

$$\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k,0:N},\sigma_{1:I,k,l-N:l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l};0,\sum_{i,n}\boldsymbol{C}_{i,k,n}\sigma^2_{i,k,l-n}), \quad (5)$$

ただし $\boldsymbol{C}_{i,k,n}=\boldsymbol{a}_{i,k,n}\boldsymbol{a}^{\mathsf{H}}_{i,k,n}$ は i 番目の音源に対する周波数 k における空間相関行列と呼ばれる行列であり，また，$\mathcal{N}_{\mathbb{C}}(\boldsymbol{x};\boldsymbol{\mu},\Sigma)\propto\exp(-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{H}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$ である．

2.3 多チャンネル NMF[4, 5] における生成モデル

通常の NMF のモデルにおいては ([7] 参照)，音源信号のパワースペクトルはスケールを除いて時不変であることが仮定されていた． 上記のモデルにこの仮定を組み込むと，

$$\sigma^2_{i,k,l}=w_{i,k}h_{i,l}, \quad (6)$$

となる．ここで $\sigma^2_{i,k,l}$ は時不変の基底パワースペクトル $w_{i,k}$ と時変な音量を表す $h_{i,l}$ の積の形で表現されている．これにより $s_{i,k,l}$ の生成モデルは以下のように書き直せる．

$$\boldsymbol{s}_{i,k,l}|w_{i,k},h_{i,l}\sim\mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{i,k,l};0,w_{i,k}h_{i,l}). \quad (7)$$

式 (7) による $\boldsymbol{Y}=\{\boldsymbol{y}_{k,l}\}_{k,l}$ の生成モデルは，NMF の多チャンネル観測信号への自然な拡張とみなすことができ，このモデルに基づく BSS のアプローチは多チャンネル NMF[4, 5] と呼ばれている．

2.4 HMM を用いた音源信号の生成モデル

上記のように，多チャンネル NMF はそれぞれの音源信号のパワースペクトルはスケールを除いて時不変であることが仮定されていた．しかし，多くの音源のパワースペクトルはその “状態” に応じて異なると考えられる．そこで音源信号の時変な性質を表現するため，ここで各音源信号のパワースペクトルの時系列と音量を，HMM を用いて表現する．

時刻 l における i 番目の音源の状態を表す隠れ変数 $z_{i,l}\in\{1,\dots,D\}$ を導入し，状態の時系列 $z_{i,1},\dots,z_{i,L}$ がマルコフ連鎖に従うと仮定する．

$$z_{i,l}|z_{i,l-1}\sim\mathrm{Categorical}(z_{i,l};\boldsymbol{\rho}_{z_{i,l-1}}). \quad (8)$$

ここで $\mathrm{Categorical}(x;\boldsymbol{y})=y_x$ であり，$\boldsymbol{\rho}_d=(\rho_{d,1},\dots,\rho_{d,D})$ は状態 d から各状態 $1,\dots,D$ への遷移確率を表し，$\boldsymbol{\rho}=(\rho_{d,d'})_{D\times D}$ は遷移行列である．ここで，$h_{i,l}$ が，$z_{i,l}$ 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，

$$h_{i,l}|z_{i,l}\sim\mathrm{Gamma}(h_{i,l};\alpha_{z_{i,l}}\beta_{z_{i,l}}), \quad (9)$$

となる．ここで $\alpha_{1:D}$ と $\beta_{1:D}$ 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，$\mathrm{Gamma}(x;\alpha,\beta)=\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$ である．$z_{i,l}$ 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，$z_{i,l}$ 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．状態 d である i 番目の音源の基底パワースペクトルを $w_{i,k,d}$ と表すとすると，時刻 l における i 番目の音源信号のパワースペクトルは $z_{i,l}$ に依存し，$s_{i,k,l}$ の生成モデルは以下のように書ける．

$$s_{i,k,l}|w_{i,k,1:D},h_{i,l},z_{i,l} \sim\mathcal{N}_{\mathbb{C}}(s_{i,k,l};0,w_{i,k,z_{i,l}}h_{i,l}). \quad (10)$$

$\boldsymbol{y}_{k,l}$ の生成モデルは各音源信号による HMM の足し合わせとなるので，最終的な生成モデルは factorial HMM とみなせる．観測信号の最終的な生成モデルは $\boldsymbol{a}_{1:I,k,0:N}$，$w_{1:I,k,1:D}$，$h_{1:I,l-N:l}$，$z_{1:I,l-N:l}$ が既知の条件下で，式 (8)，式 (9) と合わせて以下のように書ける．

$$\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k,0:N},w_{1:I,k,1:D},h_{1:I,l-N:l},z_{1:I,l-N:l} \sim\mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l};0,\sum_{i,n}\boldsymbol{C}_{i,k,n}w_{i,k,z_{i,l-n}}h_{i,l-n}). \quad (11)$$

3 パラメータ推定アルゴリズム

3.1 目的関数

ここでは，上記の生成モデルに対する，補助関数法に基づくパラメータ推定アルゴリズムについて述べる．モデルにおける推定したい変数は $\boldsymbol{W}=w_{1:I,1:K,1:D}$，$\boldsymbol{H}=h_{1:I,1:L}$，$\boldsymbol{C}=\boldsymbol{C}_{1:I,1:K,0:N}$，$\boldsymbol{Z}=z_{1:I,1:L}$ である．上記の変数の集合を Θ で表す．以下では $\boldsymbol{\rho}$ は実験的に定められた定数とする．パラメータ推定のためには事後確率

$$p(\Theta|\boldsymbol{Y})=\frac{p(\boldsymbol{Y},\Theta)}{p(\boldsymbol{Y})}, \quad (12)$$

を計算することが必要となる．ここで $\boldsymbol{Y}=\boldsymbol{y}_{1:K,1:L}$ は多チャンネル観測信号の時間周波数成分の集合である．ここで 2 章で定義された条件つき確率を用いて，同時確率 $p(\boldsymbol{Y},\Theta)$ は以下のように記述できる．

$$p(\boldsymbol{Y},\Theta)\propto p(\boldsymbol{Y}|\Theta)p(\boldsymbol{H}|\boldsymbol{Z})p(\boldsymbol{Z}). \quad (13)$$

目的関数は $L(\Theta)=\log p(\Theta|\boldsymbol{Y})$ と定義でき，我々の目的は以下の式を満たす $\hat{\Theta}$ を求めることである．

$$\hat{\Theta}=\operatorname*{argmax}_{\Theta}\log p(\Theta|\boldsymbol{Y}). \quad (14)$$

式 (12)，(13)，(14) から，この最適化問題は以下のように書き直せる．

$$\hat{\Theta}=\operatorname*{argmax}_{\Theta}\big(\log p(\boldsymbol{Y}|\Theta) +\log p(\boldsymbol{H}|\boldsymbol{Z})+\log p(\boldsymbol{Z})\big). \quad (15)$$

また，2 章で記述した生成モデルから，$\log p(\boldsymbol{Y}|\Theta)$ は以下のように書ける．

$$\log p(\boldsymbol{Y}|\Theta) =\sum_{k,l}(-\frac{M}{2}\log 2\pi-\frac{1}{2}\log|\hat{\boldsymbol{X}}_{k,l}|-\frac{1}{2}\boldsymbol{y}_{k,l}{}^{\mathsf{H}}\hat{\boldsymbol{X}}^{-1}_{k,l}\boldsymbol{y}_{k,l}). \quad (16)$$

ここで $\hat{\boldsymbol{X}}_{k,l}=\sum_{i,n}\boldsymbol{C}_{i,k,n}w_{i,k,z_{i,l-n}}h_{i,l-n}$ である．

3.2 補助関数法に基づく最適化アルゴリズム

目的関数 $L(\Theta)$ を Θ について解析的に最大化することは難しいが，補助関数法に基づく反復計算によって局所最適となる Θ を求めることができる [5]．今回の最適化問題に適用するため，まず $L(\Theta)=$

$\max_\Lambda L^+(\Theta,\Lambda)$ を満たす補助関数 $L^+(\Theta,\Lambda)$ を設計する．ここで Λ は補助変数である．紙面の都合上詳細は省略するが，$\Theta \leftarrow \mathrm{argmax}_\Theta L^+(\Theta,\Lambda)$ と $\Lambda \leftarrow \mathrm{argmax}_\Lambda L^+(\Theta,\Lambda)$ を繰り返すことにより，$L(\Theta)$ を局所最大化できる．このとき，Θ と Λ について解析的に最適化可能な $L^+(\Theta,\Lambda)$ を設計することが重要である．今回の場合は，以下のように補助関数 $L^+(\Theta,\Lambda)$ を設計できる．

$$
\begin{aligned}
&L(\Theta)\\
\geq& L^+(\Theta,\Lambda)\\
=&-\frac{1}{2}\sum_{k,l}\Bigg(\sum_{i,n}\Bigg(\frac{\mathrm{tr}(\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,n}\boldsymbol{C}_{i,k,n}^{-1}\boldsymbol{R}_{i,k,l,n})}{w_{i,k,z_{i,l-n}}h_{i,l-n}}\\
&+\mathrm{tr}(\boldsymbol{U}_{k,l}^{-1}\boldsymbol{C}_{i,k,n})w_{i,k,z_{i,l-n}}h_{i,l-n}\Bigg)+\log|\boldsymbol{U}_{k,l}|-M\Bigg)\\
&+\sum_{i,l}\Big((\alpha_{z_{i,l}}-1)\log h_{i,l}-h_{i,l}/\beta_{z_{i,l}}-\alpha_{z_{i,l}}\log\beta_{z_{i,l}}\Big)\\
&+\log p(\boldsymbol{Z}). \qquad (17)
\end{aligned}
$$

ここで $\boldsymbol{R}_{i,k,l,n}$ と $\boldsymbol{U}_{k,l}$ は $\sum_{i,n}\boldsymbol{R}_{i,k,l,n}=\boldsymbol{I}$ を満たすエルミート正定値行列であり，$\boldsymbol{R}$ と $\boldsymbol{U}$ の集合を Λ で表す．$\mathrm{tr}(\cdot)$ は行列のトレースを表す．式 (17) の等号成立条件は

$$\boldsymbol{R}_{i,k,l,n}=\boldsymbol{C}_{i,k,n}w_{i,k,z_{i,l-n}}h_{i,l-n}\hat{\boldsymbol{X}}_{k,l}^{-1}, \qquad (18)$$

$$\boldsymbol{U}_{k,l}=\hat{\boldsymbol{X}}_{k,l}, \qquad (19)$$

となる．

以上から，L は次の 2 つのステップを繰り返すことによって局所最大化できる．

1. $\boldsymbol{R}$ と $\boldsymbol{U}$ について L^+ を最大化．
2. $\boldsymbol{W}$，$\boldsymbol{H}$，$\boldsymbol{C}$，$\boldsymbol{Z}$ について L^+ を最大化．

ステップ 1 における $\boldsymbol{R}$ と $\boldsymbol{U}$ の更新では，式 (18) と式 (19) を使えばよい．ステップ 2 では，L^+ をそれぞれの変数で変微分して 0 となるものを求めることで，更新則が導ける．$\boldsymbol{W}$ と $\boldsymbol{H}$ に関する L^+ の変微分はそれぞれ以下のようになる．

$$
\begin{aligned}
&\frac{\partial L^+}{\partial w_{i,k,z_{i,l}}}\\
&=\frac{1}{2}\sum_{l,n}\Bigg(\frac{\mathrm{tr}(\boldsymbol{y}_{k,l+n}\boldsymbol{y}_{k,l+n}^{\mathsf{H}}\boldsymbol{R}_{i,k,l+n,n}\boldsymbol{C}_{i,k,n}^{-1}\boldsymbol{R}_{i,k,l+n,n})}{w_{i,k,z_{i,l}}^2h_{i,l}}\\
&-\mathrm{tr}(\boldsymbol{U}_{k,l+n}^{-1}\boldsymbol{C}_{i,k,n})h_{i,l}\Bigg), \qquad (20)\\
&\frac{\partial L^+}{\partial h_{i,l}}\\
&=\frac{1}{2}\sum_{k,n}\Bigg(\frac{\mathrm{tr}(\boldsymbol{y}_{k,l+n}\boldsymbol{y}_{k,l+n}^{\mathsf{H}}\boldsymbol{R}_{i,k,l+n,n}\boldsymbol{C}_{i,k,n}^{-1}\boldsymbol{R}_{i,k,l+n,n})}{w_{i,k,z_{i,l}}h_{i,l}^2}\\
&-\mathrm{tr}(\boldsymbol{U}_{k,l+n}^{-1}\boldsymbol{C}_{i,k,n})w_{i,k,z_{i,l}}\Bigg)\\
&+(\alpha_{z_{i,l}}-1)/h_{i,l}-1/\beta_{z_{i,l}}. \qquad (21)
\end{aligned}
$$

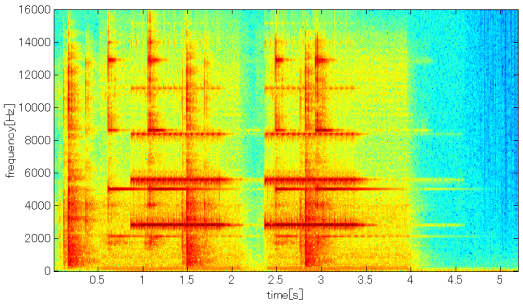

Fig. 1　作成した混合音のスペクトログラム (残響時間 600 ms)．

これらを 0 と置くことで，以下の更新則が導ける．

$$
\begin{aligned}
&w_{i,k,z_{i,l}}\\
&\leftarrow\sqrt{\frac{\sum_{l,n}\frac{\mathrm{tr}(\boldsymbol{y}_{k,l+n}\boldsymbol{y}_{k,l+n}^{\mathsf{H}}\boldsymbol{R}_{i,k,l+n,n}\boldsymbol{C}_{i,k,n}^{-1}\boldsymbol{R}_{i,k,l+n,n})}{h_{i,l}}}{\sum_{l,n}\mathrm{tr}(\boldsymbol{U}_{k,l+n}^{-1}\boldsymbol{C}_{i,k,n})h_{i,l}}}, \qquad (22)
\end{aligned}
$$

$$
\begin{aligned}
&h_{i,l}\\
&\leftarrow\frac{(\alpha_{z_{i,l}}-1)+\sqrt{(\alpha_{z_{i,l}}-1)^2+\mu_{i,l}\nu_{i,l}}}{\nu_{i,l}}, \qquad (23)
\end{aligned}
$$

ただし

$$
\begin{aligned}
&\mu_{i,l}\\
&=\sum_{k,n}\frac{\mathrm{tr}(\boldsymbol{y}_{k,l+n}\boldsymbol{y}_{k,l+n}^{\mathsf{H}}\boldsymbol{R}_{i,k,l+n,n}\boldsymbol{C}_{i,k,n}^{-1}\boldsymbol{R}_{i,k,l+n,n})}{w_{i,k,z_{i,l}}}, \qquad (24)
\end{aligned}
$$

$$
\begin{aligned}
&\nu_{i,l}\\
&=\sum_{k,n}\mathrm{tr}(\boldsymbol{U}_{k,l+n}^{-1}\boldsymbol{C}_{i,k,n})w_{i,k,z_{i,l}}+2/\beta_{z_{i,l}}, \qquad (25)
\end{aligned}
$$

である．$\boldsymbol{C}$ についての L^+ の変微分は以下のようになる．

$$
\begin{aligned}
\frac{\partial L^+}{\partial \boldsymbol{C}_{i,k,n}}=\sum_l\Bigg(&\boldsymbol{C}_{i,k,n}^{-1}\boldsymbol{R}_{i,k,l,n}\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}^{\mathsf{H}}\boldsymbol{R}_{i,k,l,n}\boldsymbol{C}_{i,k,n}^{-1}\\
&\frac{}{w_{i,k,z_{i,l-n}}h_{i,l-n}}\\
&-\boldsymbol{U}_{k,l}^{-1}w_{i,k,z_{i,l-n}}h_{i,l-n}\Bigg). \qquad (26)
\end{aligned}
$$

これを 0 と置くと，以下の Riccati 方程式が得られる．

$$\boldsymbol{C}_{i,k,n}\boldsymbol{A}_{i,k,n}\boldsymbol{C}_{i,k,n}=\boldsymbol{B}_{i,k,n}, \qquad (27)$$

ただし

$$
\begin{aligned}
\boldsymbol{A}_{i,k,n}&=\sum_l w_{i,k,z_{i,l-n}}h_{i,l-n}\hat{\boldsymbol{X}}_{k,l}^{-1},\\
\boldsymbol{B}_{i,k,n}&=\boldsymbol{C}_{i,k,n}\Big(\sum_l w_{i,k,z_{i,l-n}}h_{i,l-n}\\
&\quad\hat{\boldsymbol{X}}_{k,l}^{-1}\boldsymbol{y}_{k,l}\boldsymbol{y}_{k,l}^{\mathsf{H}}\hat{\boldsymbol{X}}_{k,l}^{-1}\Big)\boldsymbol{C}_{i,k,n}, \qquad (28)
\end{aligned}
$$

である．紙面の都合上詳細は省略するが，この Riccati 方程式を解くことで $\boldsymbol{C}$ の更新則が得られる．

L^+ を $\boldsymbol{Z}$ の関数と見ると，各音源に対応する HMM の対数事後確率の和となっている．従って最適な状態の時系列 $z_{i,1},\ldots,z_{i,L}$ をビタビアルゴリズムによって効率的に，各 i ごとに個別に求めることができる．

以上の更新則から $\boldsymbol{W}$，$\boldsymbol{H}$，$\boldsymbol{C}$，$\boldsymbol{Z}$ を反復計算により求めることは，音源分離・音響イベント・残響除去の問題を画一的最適化規準に基づき協調的に解いていることに相当している．

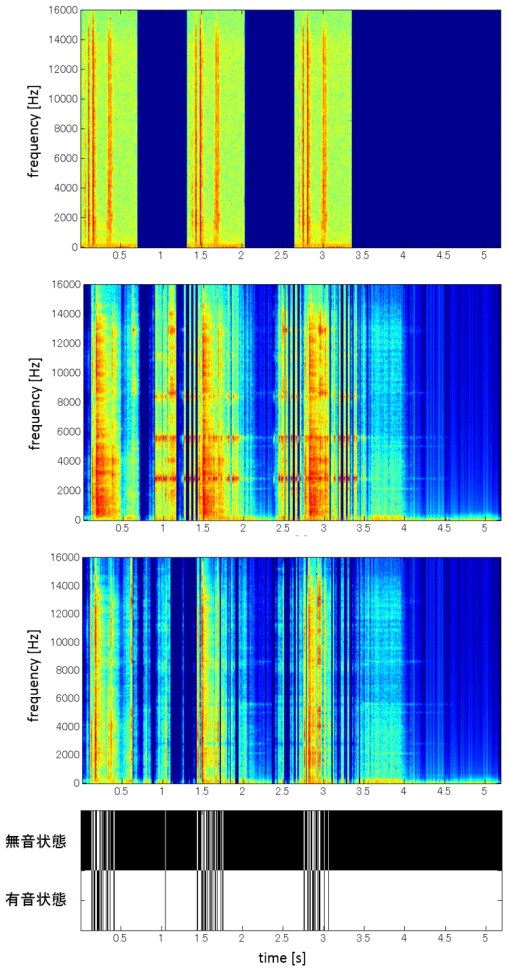

Fig. 2　無響下でのホッチキスの音のスペクトログラム (上)，従来法によって得られた分離音のスペクトログラム (中上)，提案法によって得られた分離音のスペクトログラム (中下)，音響イベント検出結果 (下) を示す．黒がその時刻で推定された状態を表す．

4　評価実験

提案法の音源分離・残響除去性能と音響イベント検出性能の評価のために，実験を行った．RWCP データベース非音声ドライソース [8] の中のホッチキスの音，ベルの音，携帯電話の着信音に対して同じく RWCP データベース [8] のインパルス応答 (残響時間 600 ms，マイク間距離 2.83 cm，マイクの数 $M = 2$) を畳み込み，人工的に残響下での多チャンネルの混合信号を作成した．Fig. 1 に作成した混合音のスペクトログラムを示す．サンプリング周波数は 32 kHz とした．フレーム長 16 ms，フレームシフト長 8 ms で STFT を行い，時間周波数展開を行った．HMM の状態数 D は 2 とした．α_1 と β_1 を 1，10^{-2} とそれぞれ設定し，$\alpha_{2:3}$ と $\beta_{2:3}$ を 1 と 10^{20} と設定することで，$d = 1$ を無音状態とみなした．

あらかじめ無響下で録音されたホッチキスの音，ベルの音，携帯電話の着信音に対して $N = 0$ としてそれぞれ提案法を適用し基底スペクトルと遷移確率を学習した後，混合音に対して $N = 20$ として提案法を適用した．$\boldsymbol{C}_{1:I,1:K,0}$ の初期値については，対角成分を $1/\sqrt{M}$，それ以外の成分を 0 とした．$\boldsymbol{C}_{1:I,1:K,1:N}$ の初期値については，対角成分を $10^{-3}/\sqrt{M}$，それ以外の成分を 0 とした．パラメータ推定アルゴリズムはまず $N = 0$ として 100 回反復し，その後徐々に N を増やしながら全体として 150 回反復した．比較対象には [6] の手法を用いた．この手法は提案法において $N = 0$ と置いたものに相当する．分離音 $\hat{\boldsymbol{y}}_{i,k,l}$ はウィーナーフィルタ

$$\hat{\boldsymbol{y}}_{i,k,l} = w_{i,k,z_{i,l}} h_{i,l} \boldsymbol{C}_{i,k,0} \hat{\boldsymbol{X}}_{k,l}^{-1} \boldsymbol{y}_{k,l}. \qquad (29)$$

によって得た．客観評価基準として，signal-to-interference ratio (SIR)[9] を用いた．高い SIR は高い音源分離性能を表す．分離前の SIR はベル，携帯電話の着信音，ホッチキスの音でそれぞれ　-16.61, 16.58, -39.17 [dB] であった．

Table 1 に提案法と従来法で得られた SIR の値を示す．提案法によって得られた分離音の SIR の平均は従来法の SIR の平均を 9.61 [dB] 上回った．Fig. 2 に無響下でのホッチキスの音のスペクトログラム (上)，従来法によって得られた分離音のスペクトログラム (中上)，提案法によって得られた分離音のスペクトログラム (中下)，音響イベント検出結果 (下) を示す．黒がその時刻で推定された状態を表す．提案法によって残響が除去されていると共に，音響イベント検出がおおむね正しく行われていることがわかる．

Table 1　提案法と従来法によって得られた SIR.

SIR [dB]	ベル	携帯電話	ホッチキス
提案法	43.35	32.05	7.19
従来法	32.37	30.15	-8.77

5　おわりに

本稿ではブラインド音源分離・残響除去・音響イベント検出を統合的に行う手法を提案した．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モデル化し，また音源信号の状態を隠れ変数とする HMM を用いて各音源信号の生成モデルを記述した．設計した生成モデルに基づくパラメータ推定を通して，音源分離・残響除去・音響イベント検出を協調的に行うことができるのが本手法のポイントである．

6　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] A. Hyvärinen *et al.*, John Wiley & Sons, 2001.

[2] D. D. Lee, and H. S. Seung, *Nature*, vol. 401, pp.788–791, 1999.

[3] P. Smaragdis, and J. C. Brown, in *Proc. WASPAA 2003*, Oct. 2003, pp. 177–180.

[4] A. Ozerov, and C. Févotte, *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550-563, Mar.2010.

[5] H. Sawada *et al.*, in *Proc. ICASSP*, pp. 261-264, 2012.

[6] 樋口　他，音講論 (秋)，2014.

[7] C. Févotte, N. Bertin, and J.L. Durrieu, *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[8] S. Nakamura *et al.*, in *Proc. LREC '00*, pp. 965–968, 2000.

[9] E. Vincent*et al.*, *IEEE Trans. ASLP*, pp. 1462–1469, 2006.