# 多チャンネル階乗隠れマルコフモデルによる音源分離・音響イベント検出・残響除去・到来方向推定の統合的アプローチとその性能評価 *

◎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．

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，音源分離，音響イベント検出，残響除去，音源の到来方向推定を統合的に行う手法を提案してきた [1, 2, 3]．本稿では，さらなる実験を行い，その性能を詳細に評価した．

## 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 $\omega_k$，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 $\omega_k$，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 $\omega_k$ における伝達周波数特性の時刻 $t_\tau$ 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 $\omega_k$，$t_l$ をそれぞれ $k$，$l$ の添え字で表す．

### 2.2 音響イベントに基づく音源信号の生成モデル

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 Non-negative Matrix Factorization(NMF)[4, 5] 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[6] と呼ばれている．しかし，多くの音源のパワースペクトルは，無音状態，音の立ち上がり，定常状態などその音源の状態 (音響イベント) に応じて異なると考えられるので，時間変化する音源の状態に応じて各音源信号が異なるパワースペクトルを持つと仮定する．時刻 $l$ における $i$ 番目の音源のパワースペクトルの状態を表す隠れ変数 $z^{(q)}_{i,l} \in \{1, \ldots, Q\}$ を導入し，状態の時系列 $z^{(q)}_{i,1}, \ldots, z^{(q)}_{i,L}$ がマルコフ連鎖に従うと仮定すると，

$$z^{(q)}_{i,l}|z^{(q)}_{i,l-1} \sim \mathrm{Categorical}(z^{(q)}_{i,l}; \boldsymbol{\rho}^{(q)}_{i,z^{(q)}_{i,l-1}}), \quad (4)$$

と書ける．ここで $\mathrm{Categorical}(x; \boldsymbol{y}) = y_x$ であり，$\boldsymbol{\rho}^{(q)}_{i,q} = (\rho^{(q)}_{i,q,1}, \ldots, \rho^{(q)}_{i,q,Q})$ は $i$ 番目の音源における状態 $q$ から各状態 $1, \ldots, Q$ への遷移確率を表し，$\boldsymbol{\rho}^{(q)}_i = (\rho^{(q)}_{i,q,q'})_{Q \times Q}$ は $i$ 番目の音源における遷移行列である．状態 $q$ である $i$ 番目の音源の基底パワースペクトルを $w_{i,k,q}$ 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 $z^{(q)}_{i,l}$ に依存し，$s_{i,k,l}$ の生成モデルは以下のように書きなおせる．

$$s_{i,k,l}|w_{i,k,1:Q}, h_{i,l}, z^{(q)}_{i,l} \sim \mathcal{N}_{\mathbb{C}}(s_{i,k,l}; 0, w_{i,k,z^{(q)}_{i,l}} h_{i,l}). \quad (5)$$

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．

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

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，

$$h_{i,l}|z^{(j)}_{i,l} \sim \mathrm{Gamma}(h_{i,l}; \alpha_{z^{(j)}_{i,l}} \beta_{z^{(j)}_{i,l}}), \quad (7)$$

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．

---

* Unified approach for source separation, audio event detection, dereverberation and DOA estimation based on multichannel factorial hidden Markov model and its performance evaluation. by HIGUCHI, Takuya (The University of Tokyo), KAMEOKA Hirokazu (The University of Tokyo, NTT)

### 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到来方向に基づいて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 $\theta (0 \leq \theta \leq \pi)$ 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．

$$\boldsymbol{J}(\theta, \omega) = \begin{bmatrix} 1 \\ e^{\mathrm{J}\omega B \cos\theta / C} \end{bmatrix} [1 \quad e^{\mathrm{J}\omega B \cos\theta / C}]^* \tag{8}$$

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 $\theta_i$ 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，[7] 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 $\theta_i$ が生成されるプロセスを以下のように記述する。

$$z_i^{(o)} | \boldsymbol{\rho}_i^{(o)} \sim \mathrm{Categorical}(z_i^{(o)}; \boldsymbol{\rho}_i^{(o)}), \tag{9}$$

$$\theta_i = \vartheta_{z_i^{(o)}}. \tag{10}$$

ただし $\boldsymbol{\rho}_i^{(o)} = (\rho_{i,1}^{(o)}, \ldots, \rho_{i,O}^{(o)})$ である。$z_i^{(o)} \in \{1, \ldots, O\}$ 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 $\rho_{i,1}^{(o)}, \ldots, \rho_{i,O}^{(o)}$) 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，$z_i^{(o)}$ 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．

$$\boldsymbol{C}_{i,k,0} | z_i^{(o)} \sim \mathcal{W}_{\mathbb{C}}(\boldsymbol{C}_{i,k,0}; \gamma, \boldsymbol{J}_{\vartheta_{z_i^{(o)}},k} + \epsilon \boldsymbol{I}). \tag{11}$$

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，$\gamma$ は $\gamma \geq M + 1$ 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 $\boldsymbol{J}_{\vartheta_{z_i^{(o)}},k}$ に微小値 $\epsilon$ を用いて単位行列 $\boldsymbol{I}$ を足してある．

観測信号の最終的な生成モデルは $\boldsymbol{a}_{1:I,k,0:\mathcal{T}}$, $w_{1:I,k,1:Q}$, $h_{1:I,l-\mathcal{T}:l}$, $z_{1:I,l-\mathcal{T}:l}^{(q)}$ が既知の条件下で，式 (4)，式 (6)，式 (7)，式 (10)，式 (11) と合わせて以下のように書き直せる．

$$\boldsymbol{y}_{k,l} | \boldsymbol{a}_{1:I,k,0:\mathcal{T}}, w_{1:I,k,1:Q}, h_{1:I,l-\mathcal{T}:l}, z_{1:I,l-\mathcal{T}:l}^{(q)} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{k,l}; 0, \sum_{i,\tau} \boldsymbol{C}_{i,k,\tau} w_{i,k,z_{i,l-\tau}^{(q)}} h_{i,l-\tau}). \tag{12}$$

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，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

目的関数である対数事後確率は各パラメータがお互いに関係し合っており，一般に最適化が困難であるが，目的関数を局所最大化する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した [3]．モデルにおける推定したい変数は $\boldsymbol{W} = w_{1:I,1:K,1:Q}$, $\boldsymbol{H} = h_{1:I,1:L}$, $\boldsymbol{C} = \boldsymbol{C}_{1:I,1:K,0:\mathcal{T}}$ である．上記の変数の集合を $\boldsymbol{\Theta}$ で表す．$\boldsymbol{Z}^{(q)} = z_{1:I,1:L}^{(q)}$, $\boldsymbol{Z}^{(j)} = z_{1:I,1:L}^{(j)}$, $\boldsymbol{Z}^{(o)} = z_{1:I}^{(o)}$ は隠れ変数とする．以下では $\boldsymbol{\rho}^{(q)}$, $\boldsymbol{\rho}^{(j)}$, $\boldsymbol{\rho}^{(o)}$ 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．

$$\begin{aligned}
L(\boldsymbol{\Theta}) \geq & L^+(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \\
= & -\frac{1}{2} \sum_{k,l} \Bigg[ \sum_{i,q,\tau} \lambda_{q,i,l-\tau}^{(q)} \\
& \Bigg( \frac{\mathrm{tr}(\boldsymbol{y}_{k,l} \boldsymbol{y}_{k,l}^{\mathsf{H}} \boldsymbol{R}_{i,k,l,q,\tau} \boldsymbol{C}_{i,k}^{-1} \boldsymbol{R}_{i,k,l,q,\tau})}{w_{i,k,q} h_{i,l-\tau}} \\
& + \mathrm{tr}(\boldsymbol{U}_{k,l}^{-1} \boldsymbol{C}_{i,k}) w_{i,k,q} h_{i,l-\tau} \Bigg) + \log |\boldsymbol{U}_{k,l}| \Bigg] \\
& + \sum_{j,i,l} \lambda_{j,i,l}^{(j)} \Big( (\alpha_j - 1) \log h_{i,l} - h_{i,l} / \beta_j \\
& - \alpha_j \log \beta_j - \log \Gamma(\alpha_j) \Big) \\
& + \sum_{i,k,o} d_{i,o} \Bigg[ -\frac{1}{2} \mathrm{tr}(\boldsymbol{C}_{i,k,0} (\boldsymbol{J}_{k,o} + \epsilon \boldsymbol{I})^{-1}) \\
& + \frac{\gamma - M}{2} \Big( \mathrm{tr}(\tilde{\boldsymbol{U}}_{i,k}^{-1} \boldsymbol{C}_{i,k,0}) \\
& + \log |\tilde{\boldsymbol{U}}_{i,k}| \Big) \Bigg] + \sum_{q,i,l} \lambda_{q,i,l}^{(q)} \log p(\boldsymbol{Z}^{(q)}) \\
& + \sum_{j,i,l} \lambda_{j,i,l}^{(j)} \log p(\boldsymbol{Z}^{(j)}) + C_{L^+},
\end{aligned} \tag{13}$$

ここで $\boldsymbol{R}_{i,k,l,\tau,q}$, $\boldsymbol{U}_{k,l}$, $\tilde{\boldsymbol{U}}_{i,k}$ は $\sum_{i,\tau} \boldsymbol{R}_{i,k,l,\tau,q} = \boldsymbol{I}$ を満たすエルミート正定値行列であり，$\lambda_{q,i,l}^{(q)}$, $\lambda_{j,i,l}^{(j)}$, $d_{i,o}$ は $\sum_q \lambda_{q,i,l}^{(q)} = 1$, $\sum_j \lambda_{j,i,l}^{(j)} = 1$, $\sum_o d_{i,o} = 1$ をそれぞれ満たす非負値のスカラー値，$C_{L^+}$ は定数項をまとめたものである．紙面の都合上詳細は省略するが，上記の補助関数 $L^+(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$ を局所最大化するように $\boldsymbol{\Lambda}$ と $\boldsymbol{\Theta}$ の更新を交互に繰り返すことで，目的関数 $L(\boldsymbol{\Theta})$ を間接的に局所最大化することができる [3]．

## 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，$\mathcal{T} = 0$，混合 DOA モデルなしの場合の提案法を用いて，教師なし/教師ありの 2 つの条件下で実験を行った．RWCP データベース非音声ドライソース [8] の中の 15 種類の音に対して同じく RWCP データベースのインパルス応答 (残響時間 0 ms，マイク間距離 11.48 cm，マイクの数 $M = 2$) を畳み込み，人工的に多チャンネルの混合信号を作成した．15 種類の音をそれぞれ 2 回ずつ使用し，ランダムに 3 つの音を混合することで，10 種類の混合音を作成した．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 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．基底パワースペクトルの状態数 $Q$ を 1, 3, 5, 10 と変えて実験を行った．音量の状態数 $J = 2$ とし，$\alpha_1$ と $\beta_1$ を 1，$10^{-3}$ とそれぞれ設定し，$\alpha_2$ と $\beta_2$ を 1

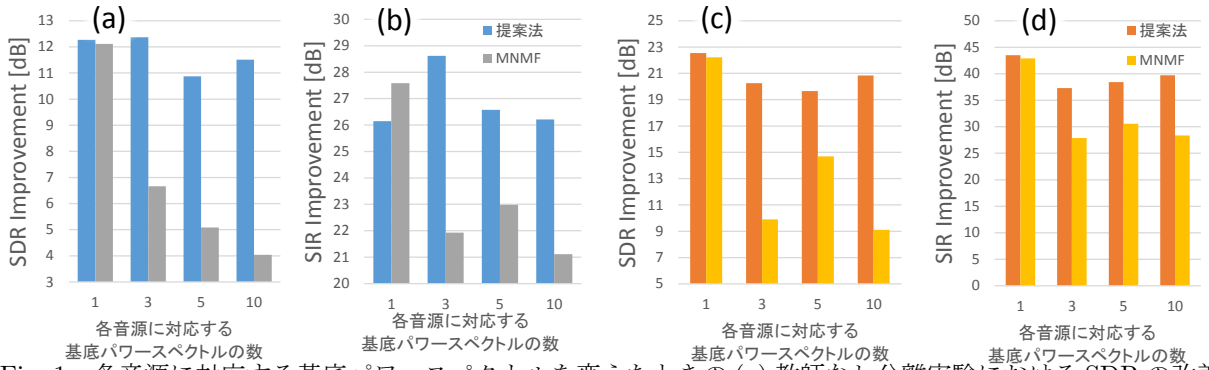


Fig. 1　各音源に対応する基底パワースペクトルを変えたときの (a) 教師なし分離実験における SDR の改善量と (b)SIR の改善量，(c) 教師あり分離実験における SDR の改善量と (d)SIR の改善量．MNMF は [6] の手法を表す．

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．$\boldsymbol{C}$ の初期値は $1/\sqrt{M} \times \boldsymbol{I}$ 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．$\boldsymbol{W}$ と $\boldsymbol{\rho}^{(q)}$ をクリーンな音から学習し，固定して用いる教師あり実験と，$\boldsymbol{W}$ を乱数を初期値として推定し，$\boldsymbol{\rho}^{(q)}$ を一様として固定する教師なし実験の 2 種類の実験を行った．パラメータ推論アルゴリズムは 100 回反復した．比較対象には [6] の手法を用い，一つの音源に対応する基底パワースペクトルの数を，1, 3, 5, 10 と変えて実験を行った．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．客観評価基準として，SDR と SIR [9] を用いた．高い SDR，SIR はそれぞれ高い音源分離性能を表す．

教師なし音源分離実験における SDR と SIR の改善量を，それぞれ Fig. 1 の (a) と (b) に，教師あり音源分離実験における SDR と SIR の改善量を，それぞれ (c) と (d) に示す．提案法は [6] の手法を上回っていることが分かる．特に，各音源に対応する基底パワースペクトルの数を増やしたときに，提案法は従来法を大きく上回った．これにより，音声などの多数のスペクトルを持つ音源信号の分離においては，[6] の手法よりも提案法が有効であることが示唆された．

### 4.2　残響除去，音響イベント検出，音源分離性能評価実験

次に，提案法の音源分離性能，残響除去性能，音響イベント検出性能の評価のために，混合 DOA モデルなしの場合の提案法において，$\mathcal{T}$ の値を $0, 3, 10$ と変えて，残響下での教師あり音源分離実験を行った．前の実験と同じ 15 種類の音に対してインパルス応答 (残響時間 600 ms，マイク間距離 11.48 cm，マイクの数 $M = 2$) を畳み込み，人工的に多チャンネルの混合信号を 10 個作成した．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 50°，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．基底パワースペクトルの状態数 $Q$ は 1 とした．$\boldsymbol{C}_{1:\mathcal{T}}$ の初期値は，$10^{-2}/\sqrt{M} \times \boldsymbol{I}$ とした．$\boldsymbol{W}$ と $\boldsymbol{\rho}^{(q)}$ をクリーンな音から学習し，固定して用いた．その他のパラメータは前の実験と同様に設定した．パラメータ推論アルゴリズムは 100 回反復した．$\mathcal{T} > 0$ のときに望ましくない局所解を避けるため，最初の 50 回の反復は $\mathcal{T} = 0$ とし，その後 $\mathcal{T}$ を 3 または 10 となるまで徐々に増やしながら反復した．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 を用いた．分離処理前の SDR，SIR はそれぞれ-7.42, -3.86[dB] であった．

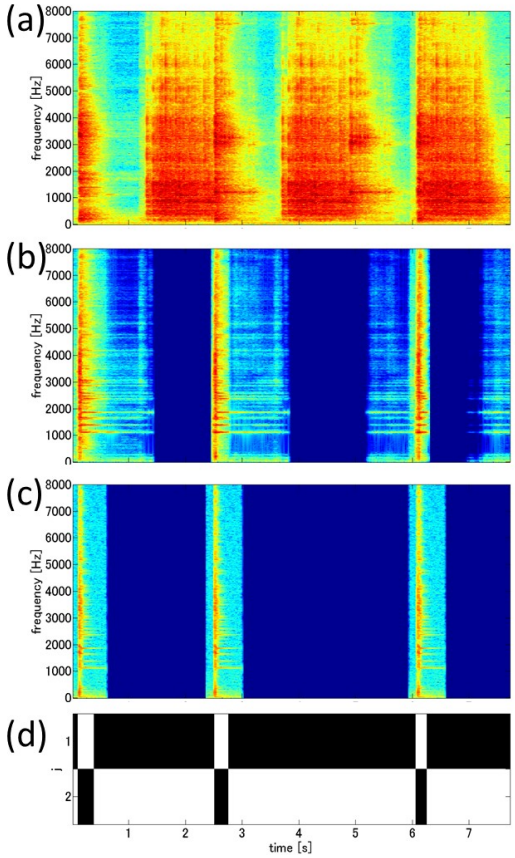


Fig. 3　残響下音源分離実験における (a) 混合音のスペクトログラム例，(b)$\mathcal{T} = 10$ のときの残響除去済み分離音のスペクトログラム，(c) 元の音源信号のスペクトログラム，(d) 音響イベントを表す $\lambda^{(j)}$ の推定結果．

Fig. 2 の (a)，(b) に $\mathcal{T}$ の値を変えたときの SDR，SIR の改善量をそれぞれ示す．畳み込み混合による残響除去を行うことで，音源分離性能が向上していることが分かる．Fig. 3 に (a) 混合音のスペクトログラム例，(b)$\mathcal{T} = 10$ のときの残響除去済み分離音のスペクトログラム，(c) 元の音源信号のスペクトログラム，(d) 音響イベントを表す $\lambda^{(j)}$ の推定結果を示す．$\lambda^{(j)}$ は黒いほどその時刻にその状態にあるらしいことを表し，$j = 2$ が有音状態に対応しているので，音響イベントがおおむね正しく検出されていることが見て取れる．

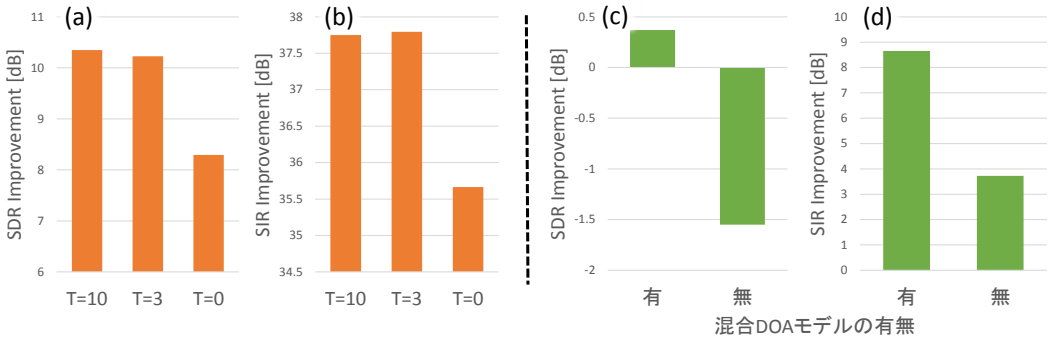


Fig. 2　$\mathcal{T}$ の値を変えたときの (a)SDR の改善量，(b)SIR の改善量．また，混合 DOA モデルの有り/無しを変えたときの (c)SDR の改善量，(d)SIR の改善量．

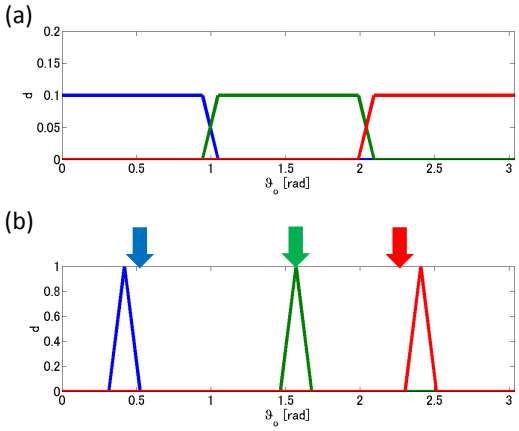


Fig. 4　音源ごとに色分けされた，各音源の到来方向を表すパラメータ $d$ の (a) 初期値と (b) 推定結果．

### 4.3　音源分離，残響除去，到来方向推定性能評価実験

提案法の音源分離，残響除去，到来方向推定性能の評価のために，混合 DOA モデルのあり/なしを変えて，残響下での教師なし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，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．ウィッシャート分布のハイパーパラメータ $\gamma$ は 3 とした．基底パワースペクトルの状態数 $Q = 15$，$\epsilon = 10^{-3}$，到来方向の分割数 $O = 30$，$\mathcal{T} = 3$ とした．$\boldsymbol{W}$ の初期値は乱数とし，$\boldsymbol{\rho}^{(q)}$ は一様とした．$d$ の初期値は Fig. 4(a) のように設定した．また $\theta_i$ の取りうる範囲を $\vartheta_{1+10(i-1)}$ から $\vartheta_{10i}$ に限定するように $\boldsymbol{\rho}^{(o)}$ の値を設定し，各音源ごとに推定する到来方向が重ならないようにした．パラメータ推論アルゴリズムは 50 回反復した．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 $\mathcal{T} > 0$ の場合に，$\boldsymbol{W}$ と $\boldsymbol{C}_{i,k,1:\mathcal{T}}$ を両方とも更新すると数値誤差によりうまく数値計算できないことがあった．これは，$\boldsymbol{W}$ と $\boldsymbol{C}_{i,k,1:\mathcal{T}}$ は各周波数ごとに求められるため，その積を観測信号に対してフィッティングする場合，それぞれのスケールに関しては任意性が生じてしまい，$\boldsymbol{C}_{i,k,1:\mathcal{T}}$ が非常に大きな値となってしまったからであると考えられる．そこで，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．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 を用いた．分離処理前の SDR，SIR はそれぞれ-4.77, -0.97[dB] であった．

Fig. 2 の (c)，(d) に混合 DOA モデル有/無のときの SDR，SIR の改善量を示す．混合 DOA モデルの導入により，より高い分離性能を示していることが分かる．Fig. 4(b) に，各音源が各到来方向にどのくらいいるらしいかを表すパラメータ $d$ の推定結果を音源ごとに色分けして示す．矢印は正解の到来方向である．到来方向がおおむね正しく推定されていることが分かる．

## 5　おわりに

本稿では，多チャンネル階乗隠れマルコフモデルによる音源分離，音響イベント検出，残響除去，到来方向推定の性能を，実験により評価した．音源分離性能の評価実験では，提案法は [6] の手法を上回る音源分離性能を示した．また音響イベント検出，残響除去，到来方向推定などを音源分離と統合的に行うことにより，提案法における音源分離性能が向上することを確認した．

**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] T. Higuchi, *et al.*, *Interspeech 2014*, pp. 850–854, 2014.
[2] T. Higuchi and H. Kameoka, "Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model," *MLSP 2014*.
[3] 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].
[4] D. D. Lee, and H. S. Seung, *Nature*, vol. 401, pp.788–791, 1999.
[5] P. Smaragdis, and J. C. Brown, *WASPAA 2003*, pp. 177–180, Oct. 2003.
[6] H. Sawada *et al.*, *ICASSP 2012*, pp. 261–264, 2012.
[7] 亀岡 他, 音講論 (春), 1–1–19, pp.713-716, Mar. 2012.
[8] S. Nakamura *et al.*, *LREC 2000*, pp. 965–968, 2000.
[9] E. Vincent *et al.*, *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1462–1469, 2006.
[10] A. Kurematsu *et al.*, *Speech Communication*, pp. 357–363, 1990.