社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

確率的モデル化に基づく移動音源の劣決定ブラインド音源分離

樋口　卓哉† 　高宗　典弘† 　中村　友彦† 　亀岡　弘和†,††

† 東京大学情報理工学系研究科　〒 113–8654 東京都文京区本郷 7–3–1
†† 日本電信電話株式会社　〒 243–0198 神奈川県厚木市森の里若宮 3–1
E-mail: †{higuchi,takamune,nakamura,kameoka}@hil.t.u-tokyo.ac.jp

あらまし　移動音源を対象とした劣決定ブラインド音源分離の問題を扱う．実環境では，音源が移動することによって，音源からマイクまでの伝達特性が時間変化してしまうことがある．従って時変な伝達特性を仮定した音源分離手法が必要である．提案手法では，音声のスパース性と音源位置の連続性の仮定を基に，多チャンネル信号の生成モデルに各音源の到来方向を状態とした隠れマルコフモデルを組み込み，パラメータ推論を通して移動音源の追跡と分離を同時実現する手法を提案する．

キーワード　劣決定ブラインド音源分離，移動音源，DOA，隠れマルコフモデル，変分ベイズ法

Underdetermined Blind Separation of Moving Sources Based on Probabilistic Modeling

Takuya HIGUCHI†, Norihiro TAKAMUNE†, Tomohiko NAKAMURA†, and Hirokazu KAMEOKA†,††

† The University of Tokyo　Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8654 Japan
†† Nippon Telegraph and Telephone Corporation　Morinosatowakamiya 3–1, Atsugi-shi, Kanagawa, 243–0198 Japan
E-mail: †{higuchi,takamune,nakamura,kameoka}@hil.t.u-tokyo.ac.jp

Abstract　This paper deals with the problem of the underdetermined blind separation and tracking of moving sources. In practical situations, sound sources such as human speakers can move and so blind separation algorithms must be designed to track the temporal change of the impulse responses. We propose solving this problem through the posterior inference of the parameters in a generative model of an observed multichannel signal, formulated under the assumption of the sparsity of time-frequency components of speech and the continuity of speakers ' movements. Specifically, we describe a generative model of mixture signals by incorporating a generative model of a time-varying frequency array response for each source, described using a path-restricted hidden Markov model (HMM). Each hidden state of the present HMM represents the direction of arrival (DOA) of each source.

Key words　underdetermined blind signal separation, moving sources, DOA, hidden Markov model, Bayesian inference

1. はじめに

本研究の最終目標は，実環境においていつ・なにが・どこで鳴ったのかを解析するシステム (音響情景分析システム) の実現である．これはすなわち，マイクで観測された信号から，音源分離，到来方向推定，残響除去，音響イベント検出などを行うことを意味する．音響情景分析システムは，会議において複数の音声信号の混じった録音データから会議録を自動作成したり，聴覚障害者へのリアルタイム音響情景提示，あるいはロボットに周囲の音環境を認識する機能を備えさせる用途への応用が期待されている．本稿では音響情景分析システム実現のための一部として，移動音源を対象とした劣決定ブラインド音源分離の問題を扱う．ブラインド音源分離 (Blind Source Separation; BSS) とは，音源から観測信号までの伝達特性が未知である場

合に，複数の音源信号が混合した観測信号から，元の音源信号を分離する技術である．

BSS の問題では観測信号から音源信号とその混合過程を推定する必要があるため，いわゆる不良設定問題となり，そのままでは解を限定することができない．従って通常は音源やその混合過程に対して何らかの仮定を置き，その仮定により立てられる規準をもとに，音源信号や伝達特性などの未知変数を推定する最適化問題として定式化される．例えば，BSS において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，この場合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．

音声を対象とした劣決定 BSS では，音声の時間周波数成分のスパース性を利用したアプローチが有効であることが知られている [2]〜[9]．音声のスパース性とは，音声信号の時間周波数成分がほとんどの時間周波数点においてほぼ 0 となる性質である．この性質により，複数の音声が同時に発話された状況でも，各音声は時間周波数領域において互いにほとんど重なり合わないと仮定できる場合が多い．従って多チャンネルの観測信号が得られている場合では，チャネル間の位相や振幅の違い等を手掛かりとして各時間周波数点でどの音源が最も優勢らしいかを推定することにより，分離信号を得ることができる．

以上の音声のスパース性を観測信号の生成モデルに組み込むためには，観測信号の生成モデルを時間周波数領域で定式化する必要がある．通常，各マイクロフォンの観測信号は音源信号と室内インパルス応答の畳み込み混合で表されるが，音源からマイクロフォンまでのインパルス応答長に対して十分に長い時間窓をもつ時間周波数分解を用いると，畳み込み混合を近似的に瞬時混合で表すことができる．この観測信号モデルに基づく BSS は周波数領域 BSS と呼ばれ，時間領域の BSS に対して演算量の少ないアルゴリズムを実現できる点や，音声のスパース性を組み込める点など特徴がある一方で，周波数ごとに分離した信号を音源ごとにグルーピングするパーミュテーション整合と呼ぶ問題を解決する必要がある．

音声のスパース性を仮定した劣決定 BSS とパーミュテーション整合を一挙に解決するアプローチとして，到来方向 (Direction of Arrival; DOA) クラスタリングが提案されている [4-7]．DOA クラスタリングでは，時間周波数領域における音源のスパース性を仮定し，音の到来方向に応じて観測信号間に位相差や振幅の違いが生まれることを利用して，時間周波数マスクを設計し観測信号を分離する手法である．しかし音源が移動する場合においては，音源信号の混合過程が時間変化することにより，BSS における以上のアプローチをそのまま適用できなかった．

本研究の目的は，各音源の移動に伴い伝達特性が時間変化する場合にも，音源位置を追跡しながら適切に音源分離を行える手法を実現することである．我々は以前，音源到来方向を離散値の潜在変数と扱い，その混合モデルにより各音源のステアリングベクトルを確率モデル化し，観測信号の生成モデルに組み込むことでパラメータ推論を通してパーミュテーション整合と周波数領域 BSS を同時に行うアプローチを提案した [8] (なお，ほぼ同時期に大塚らによっても類似したアプローチが提案されている [9])．本稿ではこれを拡張し，時間変化する各音源のステアリングベクトルを，離散化された各角度を状態とする隠れマルコフモデル (Hidden Markov Model; HMM) により確率モデル化し，実環境では音源は短い時間に大きく到来方向を変化させないという仮定を，HMM の遷移確率を設計することにより観測信号の生成モデルに組み込む．そして，変分ベイズ法に基づくパラメータ推論を通してパーミュテーション整合，各移動音源の到来方向追跡，周波数領域 BSS を同時に行う手法を提案する．

2. 観測モデル

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，m 番目のマイクロフォンで観測される信号の時間周波数成分を $y_m(\omega_k, t_l)$，i 番目の音源信号の時間周波数成分を $s_i(\omega_k, t_l)$ とし，$\boldsymbol{y}(\omega_k, t_l) = (y_1(\omega_k, t_l), \ldots, y_M(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^M$, $\boldsymbol{s}(\omega_k, t_l) = (s_1(\omega_k, t_l), \ldots, y_I(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^I$ とする．ただし，$1 \leqq k \leqq K$, $1 \leqq l \leqq L$ は時間周波数領域においてそれぞれ周波数および時間に対応するインデックスである．先に述べた通り，時間周波数領域において観測信号 $\boldsymbol{y}(\omega_k, t_l)$ は近似的に

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

のように音源信号 $s_1, \ldots, s_I$ の瞬時混合の形で表すことができる．ここで，$\boldsymbol{a}_i(\omega_k)$ は音源 i のステアリング (方向) ベクトルを表し，これを並べた行列 $\boldsymbol{A}(\omega_k) = (\boldsymbol{a}_1(\omega_k), \ldots, \boldsymbol{a}_I(\omega_k)) \in \mathbb{C}^{M \times I}$ を混合行列と呼ぶ．$\boldsymbol{n}(\omega, t)$ は背景雑音やフレーム長を超える残響成分など，瞬時混合近似で表現できない成分である．ここで音声のスパース性を仮定し，各時間周波数点 (ω_k, t_l) においてアクティブである音源のインデックスを $z_{k,l} \in \{1, \ldots, I\}$ と表すと，式 (1) は

$$\boldsymbol{y}(\omega_k, t_l) = \boldsymbol{a}_{z_{k,l}}(\omega_k) s(\omega_k, t_l) + \boldsymbol{n}(\omega_k, t_l) \qquad (2)$$

のように書き直せる．この観測モデルおいては，各時間周波数点において $z_{k,l}$ 番目の音源以外の成分はすべて 0 と仮定されたことになる．従って各時間周波数点で音源成分を表す変数は $z_{k,l}$ のみで十分であり，このため上式では $s_i(\omega_k, t_l)$ において

音源のインデックス i を省いている．すなわち $s(\omega_k, t_l)$ は各時間周波数点においてアクティブないずれかの音源の成分を表す変数となる．以後紙面のスペースの節約のため，ω_k と t_l を下付き添え字 k, l で表記することにする．

3. 生成モデル

3.1 観測信号の生成プロセス

観測モデルをもとに，観測信号が生成されるプロセスを生成モデルにより確率的に記述する．

まず，雑音成分 $\boldsymbol{n}_{k,l}$ が，平均が 0，共分散が $\Sigma_k^{(n)}$ の複素正規分布に従うと仮定すると，もし $\boldsymbol{a}_{1:I,k} = \{\boldsymbol{a}_{1,k}, \ldots, \boldsymbol{a}_{I,k}\}$,$s_{k,l}$ および $z_{k,l}$ が既知であれば，式 (2) より $\boldsymbol{y}_{k,l}$ は

$$\boldsymbol{y}_{k,l}|\boldsymbol{a}_{1:I,k,l}, s_{k,l}, z_{k,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{z_{k,l},k} s_{k,l}, \Sigma_k^{(n)}) \tag{3}$$

により生成される．ここで，$z_{k,l}$ を離散値の潜在変数と見なせば，$\boldsymbol{y}_{k,l}$ の確率分布は混合正規分布となる [6], [7]．和泉らは，この確率モデルに基づき，Expectation-Maximization (EM) アルゴリズムにより最尤の時間周波数マスクを推定するアプローチを提案している [6]．

3.2 混合 DOA モデル [8]

本節ではまず音源位置が固定の場合を考え，次節で音源が移動する場合に拡張する．これまで各音源の伝達周波数特性 $\boldsymbol{a}_{i,k}$ を周波数インデックス k ごとに独立な変数であるかのように扱っていたが，もし各音源が単一方向から平面波として到来すると仮定できるならば，例えばマイクロフォン数が 2 の場合，伝達周波数特性 $\boldsymbol{a}_{i,k}$ は，到来方向 (Direction-of-Arrival; DOA)θ の関数として

$$\boldsymbol{h}(\theta, \omega) = \begin{bmatrix} 1 \\ e^{\mathrm{J}\omega B \cos\theta / C} \end{bmatrix} \tag{4}$$

として陽に表される．ただし，$0 \leqq \theta \leqq 2\pi$，$B$ をマイクロフォンの間隔 (m)，C を音速 (m/s) とする．実際には残響や時間周波数展開におけるの瞬時混合近似などにより，$\boldsymbol{a}_{i,k}$ は上記の理論式から逸脱することが予想される．そこで，音源 i の到来方向 θ_i が既知のとき，$\boldsymbol{a}_{i,k}$ は $\boldsymbol{h}(\theta_i, \omega_k)$ を平均とした複素正規分布より生成されると仮定する．

しかし当然ながら到来方向 θ_i は実際には観測することができないため，これを潜在変数見なすことにすると，$\boldsymbol{a}_{i,k}$ の生成モデルは DOA を潜在変数とした混合モデルとなる．

そこで次に，到来方向 θ_i が生成されるプロセスを確率的に記述する．まず，$\vartheta_1, \ldots, \vartheta_D$(すべて定数) からなる D 個の DOA 候補の集合を用意する．例えば 180 度を D 等分した角度 $\vartheta_d = (d-1)\pi/D, (d = 1, \ldots, D)$ の集合を考える．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と，音源 i の到来方向 θ_i が生成されるプロセスは以下のように記述できる．

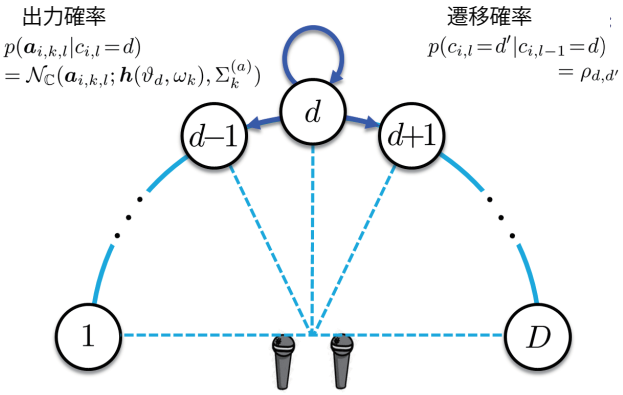

図 1　HMM によってモデル化された時変なステアリングベクトル

$$c_i|\boldsymbol{\rho}_i \sim \mathrm{Categorical}(c_i; \boldsymbol{\rho}_i) \tag{5}$$

$$\theta_i = \vartheta_{c_i} \tag{6}$$

ただし，$\boldsymbol{y} = (y_1, \ldots, y_D)$，$\sum_d y_d = 1$ とすると，$\mathrm{Categorical}(x; \boldsymbol{y}) \propto y_x$ である．また，$\boldsymbol{\rho}_i = (\rho_{i,1}, \ldots, \rho_{i,D})$ である．$c_i \in \{1, \ldots, D\}$ 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 $\rho_{i,1}, \ldots, \rho_{i,D}$) から生成されることを意味している．このプロセスにより各音源の DOA が決定され，伝達周波数特性 $\boldsymbol{a}_{i,k}$ は

$$\boldsymbol{a}_{i,k}|c_i \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k}; \boldsymbol{h}(\vartheta_{c_i}, \omega_k), \Sigma_k^{(a)}) \tag{7}$$

により生成される．これを 3.1 節の生成モデルに組み込み，生成モデル全体のパラメータ推論を行うことは，パーミュテーション整合，各音源の DOA 推定，周波数ごとの音源分離を協調的に行うことに相当する [8]．

3.3 DOA-HMM

音源が移動する場合，時刻ごとにステアリングベクトルが変化してしまうため，移動音源を扱えるようにするためには $\boldsymbol{a}_{i,k}$ を時刻 l に依存する変数 $\boldsymbol{a}_{i,k,l}$ に拡張する必要がある．このとき，式 (2) は

$$\boldsymbol{y}_{k,l} = \boldsymbol{a}_{z_{k,l},k,l} s_{k,l} + \boldsymbol{n}_{k,l} \tag{8}$$

と書き直せる．

ここで，3.2 節の自然な拡張として，各音源の DOA インデックス c_i を時刻 l に依存する変数 $c_{i,l}$ に拡張し，$c_{i,1}, \ldots, c_{i,L}$ を状態系列とした HMM によりステアリングベクトル系列 $\boldsymbol{a}_{i,k,1}, \ldots, \boldsymbol{a}_{i,k,L}$ を確率モデル化することを考える (図 1)．このとき，音源 i の時刻 l における DOA$\theta_{i,l}$ の生成プロセスは，

$$c_{i,l}|c_{i,l-1} \sim \mathrm{Categorical}(c_{i,l}; \boldsymbol{\rho}_{c_{i,l-1}}) \tag{9}$$

$$\theta_{i,l} = \vartheta_{c_{i,l}} \tag{10}$$

と表せる．$\boldsymbol{\rho}_d = (\rho_{d,1}, \ldots, \rho_{d,D})$ は状態 d から状態 $1, \ldots, D$

への遷移確率を表し，$\rho_{d,d'}$ を要素とする $D \times D$ 行列 $\boldsymbol{\rho} = (\rho_{d,d'})_{D\times D}$ を遷移行列という．このモデル化により，実際の移動音源は十分短い時間の間に大きく到来方向を変える可能性は低いという仮定を，隣接する状態への遷移確率を高めに設定することにより生成モデルに組み込むことができる．

以上のステアリングベクトル系列の確率モデルを 3.1 節のモデル (の時変版) に組み込み，次章で述べるパラメータ推論を通してパーミュテーション整合，移動音源の追従，周波数ごとの音源分離を同時に行うことが提案手法の要点である．

4. 変分推論アルゴリズム

観測信号 $\boldsymbol{Y} = \boldsymbol{y}_{1:K,1:L}$ が与えられたもとで，以上の生成モデルのパラメータ $\boldsymbol{A} = \boldsymbol{a}_{1:I,1:K,1:L}$, $\boldsymbol{S} = s_{1:K,1:L}$, $\boldsymbol{Z} = Z_{1:K,1:L}$, $\boldsymbol{C} = c_{1:K,1:L}$ の事後分布 $p(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{Y})$ を求めたい．この事後分布を解析的に得ることは難しいが，変分推論法に基づき近似分布を反復計算により得ることができる．以下では，ρ, $\Sigma_{1:K}^{(n)}, \Sigma_{1:K}^{(a)}$ は実験的に定める定数とする．

変分推論は事後分布 $p(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{Y})$ と，

$$\int \cdots \int q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}) \mathrm{d}\boldsymbol{A} \cdots \mathrm{d}\boldsymbol{C} = 1 \tag{11}$$

を満たす非負の変関数 $q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C})$ との間の Kullback-Leibler ダイバージェンス

$$\mathcal{F}[q] = \left\langle \log \frac{q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C})}{p(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{Y})} \right\rangle_{q(\boldsymbol{A},\boldsymbol{S},\boldsymbol{Z},\boldsymbol{C})} \tag{12}$$

を q に関して最小化することが目的となる．ただし $\langle f(x) \rangle_{q(x)}$ は $\int q(x) f(x) dx$ を表す．無論，$\mathcal{F}[q]$ は $p = q$ のとき最小となるが，q に関して

$$q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}) = q(\boldsymbol{A}) q(\boldsymbol{S}) q(\boldsymbol{Z}) q(\boldsymbol{C}) \tag{13}$$

となるような分布クラスを考え，$\mathcal{F}[q]$ を $q(\boldsymbol{A})$, $q(\boldsymbol{S})$, $q(\boldsymbol{Z})$, $q(\boldsymbol{C})$ について交互に最小化するステップを繰り返すことで，当該分布クラスの中で $p(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{Y})$ を最も良く近似する分布を得ようというのが変分推論法の基本的な考え方である．

導出は省略するが，各パラメータにおける最適な分布は以下の形で表記できる．

$$\hat{q}(\boldsymbol{X}) \propto \exp \boldsymbol{E}_{\Theta \setminus \boldsymbol{X}} [\log p(\boldsymbol{Y}, \Theta)], \tag{14}$$

ここで $\boldsymbol{X}$ はパラメータのひとつを表しており，$\boldsymbol{E}_{\Theta\setminus\boldsymbol{X}}[\log p(\boldsymbol{Y}, \Theta)]$ は観測と，$\boldsymbol{X}$ を除くパラメータ群の同時確率の期待値である．各パラメータの更新式は具体的に以下の形として求まる．

$$\hat{q}(\boldsymbol{A}) = \prod_{i,k,l} \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k,l}; \boldsymbol{m}_{i,k,l}, \Gamma_{i,k,l}) \tag{15}$$

$$\hat{q}(\boldsymbol{S}) = \prod_{k,l} \mathcal{N}_{\mathbb{C}}(s_{k,l}; \mu_{k,l}, \sigma_{k,l}) \tag{16}$$

$$\hat{q}(\boldsymbol{Z}) = \prod_{k,l} \hat{q}(z_{k,l}), \hat{q}(z_{k,l} = i) = \phi_{i,k,l} \tag{17}$$

ただし

$$\Gamma_{i,k,l}^{-1} = (\phi_{i,k,l}(|\mu_{k,l}|^2 + \sigma_{k,l}^2))\Sigma_k^{(n)-1} + \Sigma_k^{(a)-1}, \tag{18}$$

$$\begin{aligned}\boldsymbol{m}_{i,k,l} =& \Gamma_{i,k,l}(\Sigma_k^{(n)-1}\phi_{i,k,l}\mu_{k,l}^* \boldsymbol{y}_{k,l} \\ &+ \Sigma_k^{(a)-1} \sum_d \hat{q}(c_{i,l} = d) \boldsymbol{h}(\vartheta_d, \omega_k)),\end{aligned} \tag{19}$$

$$\frac{1}{\sigma_{k,l}^2} = \sum_i \phi_{i,k,l} \mathrm{tr}[(\boldsymbol{m}_{i,k,l}\boldsymbol{m}_{i,k,l}^{\mathsf{H}} + \Gamma_{i,k,l})\Sigma_k^{(n)-1}], \tag{20}$$

$$\mu_{k,l} = \sigma_{k,l}^2 (\sum_i \phi_{i,k,l} \boldsymbol{m}_{i,k,l}^{\mathsf{H}}) \Sigma_k^{(n)-1} \boldsymbol{y}_{k,l}, \tag{21}$$

$$\begin{aligned}\varphi_{i,k,l} =& \exp(2\mathrm{Re}[\mu_{k,l}\boldsymbol{y}_{k,l}^{\mathsf{H}}\Sigma_k^{(n)-1}\boldsymbol{m}_{i,k,l}] \\ &- (|\mu_{k,l}|^2 + \sigma_{k,l}^2)\mathrm{tr}[(\boldsymbol{m}_{i,k,l}\boldsymbol{m}_{i,k,l}^{\mathsf{H}} + \Gamma_{i,k,l})\Sigma_k^{(n)-1}]),\end{aligned} \tag{22}$$

$$\phi_{i,k,l} = \frac{\varphi_{i,k,l}}{\sum_i \varphi_{i,k,l}} \tag{23}$$

である．なお以上の更新則は，ステアリングベクトルが時変であるように拡張されている点を除いて [8] と同様である．また Forward-Backward アルゴリズムによって $\hat{q}(\boldsymbol{C})$ を求めることができる．具体的な更新式は以下の形となる．

$$\hat{q}(\boldsymbol{C}) = \prod_{i,l} \frac{\alpha(\vartheta_{c_{i,l}})\beta(\vartheta_{c_{i,l}})}{\sum_{\vartheta_{c_{i,l}}} \alpha(\vartheta_{c_{i,l}})\beta(\vartheta_{c_{i,l}})}, \tag{24}$$

ただし α と β はそれぞれ Forward 変数，Backward 変数であり，$\hat{q}(\boldsymbol{a}_{i,k,l}|\vartheta_d)$ を用いて以下のように書き表せる．

$$\alpha(\vartheta_{c_{i,l}}) = \hat{q}(\boldsymbol{a}_{i,k,l}|\vartheta_{c_{i,l}}) \sum_{\vartheta_{c_{i,l-1}}} \alpha(\vartheta_{c_{i,l-1}}) \rho_{c_{i,l-1},c_{i,l}}, \tag{25}$$

$$\beta(\vartheta_{c_{i,l}}) = \sum_{\vartheta_{c_{i,l+1}}} \beta(\vartheta_{c_{i,l+1}}) \hat{q}(\boldsymbol{a}_{i,k,l+1}|\vartheta_{c_{i,l+1}}) \rho_{c_{i,l},c_{i,l+1}} \tag{26}$$

ここで，変分法における最適な $\hat{q}(\boldsymbol{a}_{i,k,l}|\vartheta_d)$ は，具体的には以下のように書ける．

$$\begin{aligned}\hat{q}(\boldsymbol{a}_{i,k,l}|\vartheta_d) \propto& \exp \boldsymbol{E}_{\boldsymbol{a}_{i,k,l}}[\log p(\boldsymbol{a}_{i,k,l}|\vartheta_d)] \\ =& \exp(-\mathrm{tr}[(\boldsymbol{m}_{i,k,l}^{\mathsf{H}}\boldsymbol{m}_{i,k,l} + \Gamma_{i,k,l})\Sigma_k^{(a)-1}] \\ &+ 2\mathrm{Re}[\boldsymbol{h}(\vartheta_d, \omega_k)^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{m}_{i,k,l}] \\ &- \boldsymbol{h}(\vartheta_d, \omega_k)^{\mathsf{H}}\Sigma_k^{(a)-1}\boldsymbol{h}(\vartheta_d, \omega_k))\end{aligned} \tag{27}$$

以上の変分推論アルゴリズムによって推定された $s_{k,l}$ の平均値 $\mu_{k,l}$ に確率値 $\phi_{k.l}$ を乗じることで，音源 i の推定信号を得ることができる．

5. 複数移動音源の到来方向推定と音源分離実験

提案法の有効性を示すため，移動音源に対して音源分離と到来方向推定性能の検証を行った．移動音源として移動音源

データベース [10] の男性話者の音声信号 2 つを (移動音源 A, B)，固定音源として音声データベース [11] の女性話者の音声信号に室内インパルス応答を畳み込み加算したもの 1 つを用い，それらを人工的に混合したものを観測信号とした．残響時間は 0 ms である．移動音源を変えることで，10 通りの混合音声データセットを作成し，実験した．標本化周波数は 16 kHz とした．短時間フーリエ変換 (フレーム長は 64 ms，フレームシフトは 16 ms) により算出した．$\Sigma_k^{(n)}$ と $\Sigma_k^{(a)}$ はそれぞれ $\boldsymbol{I}$, $10^{1.5} \times \boldsymbol{I}$ とした．また角度の分割数は $D = 180$ とした．初期値に関しては，$\hat{q}(z_{1:K,1:L} = i)$ はすべての i に対して 1/3 とした．$\hat{q}(c_{1:3,1:L} = d)$ の初期値は，$\hat{q}(c_{1,1:L} = 46)$, $\hat{q}(c_{2,1:L} = 91)$, $\hat{q}(c_{3,1:L} = 136)$ をそれぞれ比較的大きな値に設定した．また本実験では，音源信号の推定値 $s(\omega_k, t_l)$ は $y_1(\omega_k, t_l)$ に固定して更新しなかった．比較的ノイズが小さい環境下においてはこれは理にかなっており，$s(\omega_k, t_l)$ が局所解に落ちてしまうことを防ぐ狙いがある．また高周波帯域でおこる空間的エイリアシングによって，$\hat{q}(\boldsymbol{C})$ が局所解に落ちることを防ぐために，まず低周波帯域のみで 4. 章の反復アルゴリズムを実行し，徐々に周波数帯域を広げながら反復する方法をとった．全体として 100 回反復アルゴリズムを実行した後，音源成分の推定値 $\mu_{k,n}$ に，音源 i が時間周波数点でどれだけアクティブらしいかを表す確率値 $\phi_{i,k,n}$ を乗じたものを，音源 i の推定時間周波数成分とした．音源分離性能の評価基準として，式 (28) により導出される Signal-to-Interference-Ratio (SIR) [12] を用いた．SIR の計算には，3 つの音源のうち一番短い長さの音源が終了するまでを用いた．

$$\mathrm{SIR}_i = 10 \log_{10} \frac{\sum_{k,l} \hat{s}_{i,k,l}}{\sum_{i' \neq i} \sum_{k,l} \hat{s}_{i',k,l}} [\mathrm{dB}] \qquad (28)$$

ただし $\hat{s}_{i,k,n}$ は音源 i の推定信号 $\phi_{i,k,n}\mu_{k,n}$ に含まれる音源 i の信号成分である．

また各時刻の到来角度の推定値には，推定された到来角度の確率分布から各時刻において最も確率値の高い角度を用いた．

さらに音源の移動を仮定しない従来法が，移動音源に対して良い分離性能をもたないことを示すため, [8] の手法を用いて同様の音源分離実験を行った場合の結果と提案法の結果を比較した．

表 1 に各音源ごとにデータセットと時刻で平均をとった SIR とその標準偏差 (Standard Deviation; SD) の値を示す．3 つの音源すべてにおいて，従来法では SIR が低く音源分離が行えていないのに対して，提案法では SIR が改善されているのがわかる．3 つの音源における SIR の平均値は，従来法で 0.15 dB, 提案法で 6.35 dB であった．

次に，到来方向推定の結果例を図 2 に示す．実際の到来方向と比べて，1 s 付近から音源同士の到来方向が重なり，かつ音

表 1　提案法と従来法における SIR の平均値と標準偏差

SIR(±SD) [dB]	移動音源 A	移動音源 B	固定音源
提案法	4.82(±3.94)	6.07(±3.24)	8.16(±1.50)
従来法	-1.10(±1.37)	-2.00(±1.70)	3.55(±0.78)

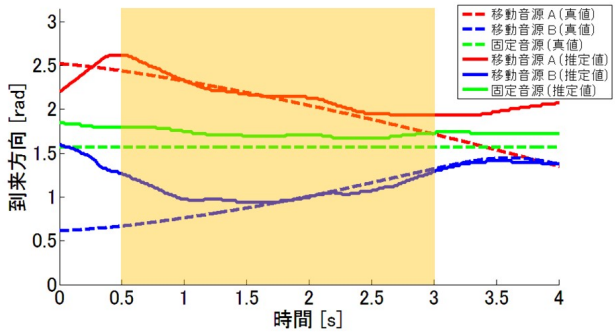

図 2　各音源における到来方向の真値と推定値の例．おおよそ 0.5 s から 3 s の間に発話が行われている．

声の終了する 3 s 付近までは，おおむね正しく推定されていることが分かる．最初の約 1 s の間で到来方向推定の精度が良くないのは，生成モデルに組み込まれた，到来方向が急に変化しにくいという仮定により，音声の入っていない初期時刻付近のデータに対して推定された到来方向から滑らかにつなぐように到来方向が推定されてしまうからであると考えられる．固定音源の推定方向にバイアスがのっているのは，理想的なステアリングベクトルと実際のステアリングベクトルとの誤差からくる推定誤差である可能性だけでなく, [10] のデータベース作成時のマイクロフォンの角度誤差である可能性も考えられる．

6. おわりに

本稿では，音源が移動することでステアリングベクトルが時間変化する場合においても安定して動作する BSS アルゴリズムの実現を目指した．音声の時間周波数成分のスパース性に基づく周波数領域の劣決定 BSS モデルをベイズ的に記述し，時変なステアリングベクトルを，離散化した到来角度を状態とする隠れマルコフモデルとして，混合 DOA モデルと組み合わせて表現し，短い時間において音源の到来角度が大きく変化する確率は小さいという仮定を遷移確率として観測信号の生成モデルに組み込んだ．これにより，変分ベイズ法に基づくパラメータ推論を通して，周波数ごと，時間ごとのパーミュテーション整合，時間変化する到来角度の推定，音源分離を一挙に行えることが，提案法の主要な特徴である．2 つの移動音源と 1 つの固定音源の混合音声に対する音源分離実験では，固定音源を仮定した従来法と比較して，提案法では SIR が平均で 6.20 dB 向上した．今後は，残響下 (瞬時混合が成り立たない場合) においても妥当な観測信号の生成モデルを設計することや，音響イベントを表すパラメータなどを生成モデルに組み込むことで，音

源分離，到来方向推定，残響除去，音響イベント検出などを統合的に行い，実環境音響情景分析の実現を目指す予定である．

7. 謝　　辞

本研究は JSPS 科研費 26730100 の助成を受けたものです．

文　　献

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] Ö. Yılmaz & S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, 52(7), pp. 1830–1847, 2004.

[3] Y. Mori , H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita, "Real-time implementation of two-stage blind source separation combining SIMO-ICA and binary masking," *Proc. 9th International Workshop on Acoustic Echo and Noise Control (IWAENC 2005)*, pp. 229–232, 2005.

[4] M. I. Mandel , D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Adv. Neural Info. Process. Syst.*, 2006, pp. 953–960.

[5] S. Araki *et al.*, , H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, 87(8), pp. 1833–1847, 2007.

[6] 和泉 洋介, 小野 順貴, 嵯峨山 茂樹, "EM アルゴリズムを用いた音声スパース性に基づく 2ch BSS, " 日本音響学会春季研究発表会講演集, pp.555–556, Mar. 2007.

[7] H. Sawada , S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2010.

[8] 亀岡 弘和, 佐藤 美沙, 小野 拓磨, 小野 順貴, 嵯峨山 茂樹, "ノンパラメトリックベイズアプローチによる劣決定スパース BSS, " 日本音響学会春季研究発表会講演集, pp.713–716, Mar. 2012.

[9] T. Otsuka , K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian unification of sound source localization and separation with permutation resolution," in *Proc. of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, pp. 2038–2045, 2012.

[10] S. Nakamura *et al.*, , K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. 2nd International Conference on Language Resources & Evaluation (LREC 2000)*, pp. 965–968, 2000.

[11] A. Kurematsu , K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, pp. 357–363, 1990.

[12] E. Vincent *et al.*, , R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1462–1469, 2006.