人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B102-6

音源定位手法 MUSIC のベイズ拡張

Bayesian Extension of MUSIC for Sound Source Localization

大塚 琢馬 †, 中臺 一博 ‡, 尾形 哲也 †, 奥乃 博 †

Takuma Otsuka†, Kazuhiro Nakadai‡, Tetsuya Ogata†, Hiroshi G. Okuno†

† 京都大学大学院情報学研究科, ‡(株) ホンダ・リサーチ・インスティチュート・ジャパン

†**Graduate school of Informatics, Kyoto University, ‡HONDA Research Institute Japan, Co., Ltd.**

†{ohtsuka, ogata, okuno}@kyoto-u.ac.jp, ‡nakadai@jp.honda-ri.com

Abstract

This paper presents a Bayesian extension of MUSIC-based sound source localization (SSL) method. SSL is important for the separation of simultaneous speech signals as well as for auditory scene analysis by mobile robots. One of the drawbacks of existing SSL methods is the necessity of careful parameter tunings, e.g., the sound source detection threshold depending on the reverberation time and the number of sources. Our contribution consists of (1) automatic parameter estimation in the variational Bayesian framework and (2) tracking of sound sources with reliability. Experimental results demonstrate our method robustly tracks multiple sound sources in a reverberant environment with RT20 = 840 (ms).

1 はじめに

音響情報は人間の知覚の重要な位置を占める. 例えば, 人は足音を聞くことで目に頼ることなく誰かが近づいてきている, あるいは遠ざかっているといった状況を理解することができる. ロボットや計算機による周囲の音響情報の理解, つまり, 「音環境理解」の実現は, 聴覚障碍者の補助や, 人間の音に対する気づきを向上することができると期待される [Kubota et al., 2008].

音源定位はマイクロフォンアレイを用いた同時発話混合音声の分離 [Nakadai et al., 2010], 遠隔ロボットのオペレータへの音源方向提示 [Mizumoto et al., 2011], 移動ロボットによる音源検出と位置推定 [Sasaki et al., 2010]など, 音環境理解にとって重要な要素技術である. 図 1 に示すような, 複数音源, ロボットの移動, 音源移動など, 動的に音環境が変化する状況においても, 手間のかかるパラメータ設定をしなくてもロボットが頑健に各音源を定位, 追跡することが望まれる.

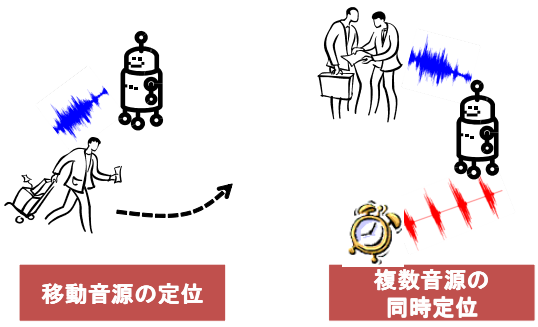

Figure 1: 動的環境下での音源定位

マイクロフォンアレイを用いた音源定位法はビームフォーミングに基づく手法 [Doclo et al., 2001] と, MUltiple SIgnal Classification (MUSIC) に基づく手法 [Schmidt, 1986; Asano et al., 2001; Danès et al., 2010] がロボットによく応用される. 我々は次の理由より, MUSIC 法を利用する. (1) MUSIC の方が雑音に頑健である, (2) 音源数がマイクロフォン数未満という条件下では, 比較的安定して複数音源の定位が可能である.

通常の MUSIC 法では, 音源が到来しているかどうかを MUSIC スペクトルと呼ばれる音源到来評価関数に対して閾値を設定して判定する. 多くの場合, 適切な閾値は環境中の音源数や残響時間などに依存するため, 状況に応じて最適な閾値の設定が重要である. MUSIC 法を用いた場合の環境中の音源数推定問題は, 赤池情報量規準の利用 [Danès et al., 2010] や, サポートベクターマシンの適用 [Yamamoto et al., 2006] によってこれまで取り組まれてきた. しかし, これらの手法で音源数が推定できたとしても, 適切な音源検出閾値の設定問題は依然として残っている. この問題に対する典型的な対策としては, マイクロフォンアレイを設

置した環境で録音した音響信号から計算した MUSIC スペクトルを見ながら手動で閾値を設定するという方法であった.

本稿では, MUSIC 法による音源定位のベイズ拡張を行い, 従来法で必要とされた閾値に相当する情報を自動的に学習することを試みる. これにより, 閾値設定の手間を省くと共に, 試行錯誤により設定された閾値の精度と同等以上の定位精度を実現する. 本手法は次の 2 つのステップから成る. (1) マイクロフォンアレイが置かれた環境で録音した数十秒程度の音響信号から, 音源存在閾値に相当するパラメータを学習する. 学習には変分ベイズ隠れマルコフモデル (VB-HMM) [Beal, 2003] に基づくパラメータ推定アルゴリズムを用いる. (2) VB-HMM により学習したパラメータを用いた複数音源の逐次的定位を行う. 逐次定位では, 観測モデルが VB-HMM より複雑になるため, パーティクルフィルタ [Arulampalam et al., 2002] を用いる.

2 MUSIC 法を用いる音源定位

まず本稿が扱う問題を述べ, MUSIC スペクトルの算出法を説明する. 本稿での水平面上の音源到来方向推定問題を, 図 2 に示した. 今回用いたマイクロフォンアレイは, マイクロフォンがロボットに円状に 8 本配置されており, 水平面上に 5° 刻みの解像度での定位を行う. 以下に本稿で扱う問題設定を示す.

入力 M チャネルの音響信号と, 各周波数ビンごとに D 方向からの伝達関数,
出力 N 個の音源到来方向,
仮定 同時に検出可能な最大音源数 N_{max} はマイクロフォンの数未満 ($N \leq N_{max} < M$).

水平面一周を 5° 刻みに定位するので, $D = 72$ である.

次に, MUSIC スペクトルの算出法について簡単に述べる. より詳細は文献 [Schmidt, 1986; Danès et al., 2010] などに記述されている. MUSIC 法は時間周波数領域[1] において適用される.

$\mathbf{x}_{\tau,\omega} \in \mathbb{C}^M$ を M チャネル音響信号の時間フレーム τ, 周波数ビン ω における複素振幅ベクトルとする. 各周波数ビン ω, ΔT (sec) 間隔の時刻 t に対して, (1) 入力信号の自己相関行列 $\mathbf{R}_{t,\omega}$ の計算, (2) $\mathbf{R}_{t,\omega}$ の固有値分解, (3) 固有ベクトルと伝達関数から MUSIC スペクトルの計算を行う.

(1) 入力信号の自己相関行列は時間 ΔT で観測したサンプル値の相関として計算する.

$$\mathbf{R}_{t,\omega} = \frac{1}{\tau(t) - \tau(t-\Delta T)} \sum_{\tau=\tau(t-\Delta T)}^{\tau(t)} \mathbf{x}_{\tau,\omega}\mathbf{x}_{\tau,\omega}^H, \tag{1}$$

ただし, $(\cdot)^H$ はエルミート転置, $\tau(t)$ は時刻 t に対応する時間フレームを表す. 入力ベクトル $\mathbf{x}_{\tau,\omega}$ の M 個の要素は

[1] 我々の実装では, サンプリング周波数 16000 (Hz) で, 窓長 512 (pt), シフト幅 160 (pt) の短時間フーリエ変換を行っている.

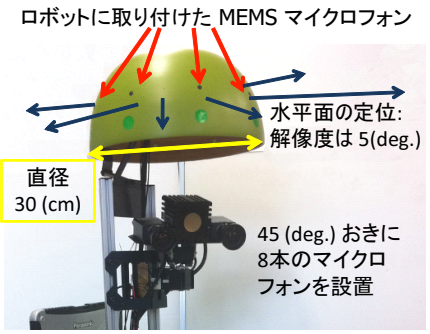

Figure 2: 使用したロボット Kappa. 本稿では水平面上の定位を扱う (青矢印). 8 本のマイクロフォンがロボットのボウルに沿って付けられている (赤矢印).

各チャネルに対応する.

(2) $\mathbf{R}_{t,\omega}$ を次のように固有値分解する.

$$\mathbf{R}_{t,\omega} = \mathbf{E}_{t,\omega}^H \mathbf{Q}_{t,\omega} \mathbf{E}_{t,\omega}, \tag{2}$$

ここで, $\mathbf{E}_{t,\omega}$ は固有ベクトル, $\mathbf{Q}_{t,\omega}$ は固有値から成る対角行列である. $\mathbf{E}_{t,\omega} = [\mathbf{e}_{t,\omega}^1 \ldots \mathbf{e}_{t,\omega}^M]$ と, $\mathbf{R}_{t,\omega}$ の M 個の固有ベクトルで表せ, $\mathbf{Q}_{t,\omega} = \mathrm{diag}(q_{t,\omega}^1 \ldots q_{t,\omega}^M)$ となる. ただし, 固有値 $q_{t,\omega}^m$ は降順に並べられているものとする.

入力信号に N 個の音源が含まれる場合, 固有値 $q_{t,\omega}^1$ から $q_{t,\omega}^N$ までが, 音源のエネルギーに対応する大きな値を持つ. それに対し, 残りの固有値 $q_{t,\omega}^{N+1}$ から $q_{t,\omega}^M$ まではマイクロフォンに伴う観測ノイズなどによる小さな値を取る. ここで重要なポイントは, $\mathbf{e}_{t,\omega}^{N+1}$ から $\mathbf{e}_{t,\omega}^M$ のノイズに対応する固有ベクトルは, 音源到来方向に対応する伝達関数ベクトルと直交するという点である [Schmidt, 1986].

(3) MUSIC スペクトルは以下のように計算する.

$$P_{t,d,\omega} = \frac{\|\mathbf{a}_{d,\omega}^H \mathbf{a}_{d,\omega}\|}{\sum_{m=N_{max}+1}^{M} \|\mathbf{a}_{d,\omega}^H \mathbf{e}_{t,\omega}^m\|}, \tag{3}$$

ただし, $\mathbf{a}_{d,\omega}$ は方向 d, 周波数ビン ω に対応する M 次元の伝達関数ベクトルである. これらの伝達関数はマイクロフォンアレイを用いて事前に測定したものである. 今, 観測されいる最大の音源数は N_{max} 個と仮定している. そのため, $\mathbf{e}_{t,\omega}^{N_{max}+1}$ から $\mathbf{e}_{t,\omega}^M$ までの固有ベクトルは, 常に音源到来方向 d に対応する伝達関数 $\mathbf{a}_{d,\omega}$ と直交する. 従って, 式 (3) の分母は音源到来方向の d に対しては 0 となる. つまり, MUSIC スペクトル $P_{t,d,\omega}$ は ∞ に発散する. ただし, 実際には, 壁からの反射音などの影響で MUSIC スペクトルは発散せず鋭いピークとして観測されることが多い.

周波数ビンごとの MUSIC スペクトルを合算する.

$$P'_{t,d} = \sum_{\omega=\omega_{min}}^{\omega_{max}} \sqrt{q_{t,\omega}^1} P_{t,d,\omega}, \tag{4}$$

ここで, $q_{t,\omega}^1$ は周波数ビン ω における最大固有値である. 我々の実装では, 音声信号を対象とするため, $\omega_{min} = 500$ (Hz), $\omega_{max} = 2800$ (Hz) とした.

従来法では各方向 d に対して, $P'_{t,d} > P_{thres}$ のように閾値

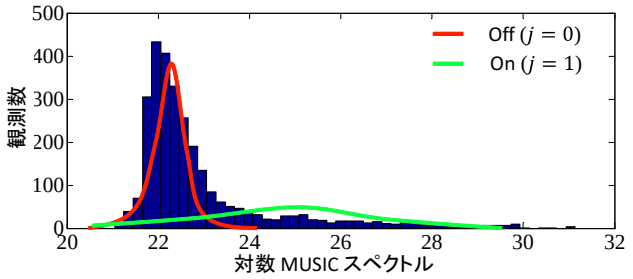

Figure 3: 対数 MUSIC スペクトルの分布. 青: 対数 MUSIC スペクトルのヒストグラム, 赤線: 音源がない場合のガウス分布; 緑線: 音源が存在する場合のガウス分布.

P_{thres} を用いて音源存在判定を行う. しかし, 適切な P_{thres} は残響時間や, 最大音源数 N_{max} に依存するため, 実験的に設定されることが多かった.

3 音源定位のベイズ拡張

MUSIC 法による音源定位のベイズ拡張アルゴリズムは次の 2 ステップから成る. (1) パラメータ学習: VB-HMM を用いて, 対象環境で録音した音響信号からパラメータの事後分布を計算する. (2) オンライン音源定位: パーティクルフィルタを用いて, 学習したパラメータの事後分布を元に複数音源の存在事後確率計算を行う. HMM では状態ベクトルとして D 次元の 2 値ベクトルを用い, 各次元の値が, その方向の音源が存在するか否かを示す. 音源存在閾値 P_{thres} に相当する情報が VB-HMM のパラメータの事後分布として自動的に学習される.

観測モデルはガウス混合モデル (GMM) を用いる. MUSIC スペクトルをガウス分布に従う観測値とみなし, 音源の有無に対応するガウス分布を利用する. ガウス分布を用いる理由は, 複数の周波数ビンの値を加算して対数をとった MUSIC スペクトルが近似的にガウス分布とみなせるためと, ガウス分布を用いることで計算が容易となるためである. 図 3 は対数スケールの MUSIC スペクトルである. 音源が存在しない (Off) のときのガウス分布は狭い MUSIC スペクトルの領域に形成され, 音源が存在する (On) ときの分布は値の広い領域を覆っている. VB-HMM の学習を通じて, 図 3 に示すようなガウス分布のパラメータである平均, 精度 (分散の逆数) の事後分布が計算される.

逐次的な音源定位には以下の 2 要件を満たす観測モデルを利用するためパーティクルフィルタを用いる. (1) 各時刻で同時に存在する音源数は高々 N_{max} 個. (2) $P'_{t,d}$ の極大点にしか音源は存在しない. 詳しい説明は 3.2 節に記す.

3.1 VB-HMM を用いたパラメータ学習

本手法は次の対数 MUSIC スペクトルを観測とする.

$$x_{t,d} = 10\log_{10} P'_{t,d}. \quad (5)$$

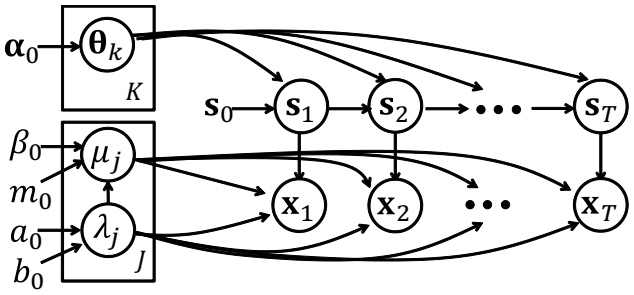

Figure 4: VB-HMM のグラフィカルモデル

$s_{t,d}$ を音源存在を表す 2 値変数し, $s_{t,d} = 1$ のときは時刻 t, 方向 d に音源が存在するものとする.

図 4 に VB-HMM の確率変数間の条件付き独立性を示すグラフィカルモデルを示す. 通常の HMM と VB-HMM との違いは, 状態遷移確率のパラメータ θ_k や, 観測確率のパラメータ μ, λ が固定値ではなく, 確率変数として扱われる点である. これらのパラメータの確率分布を学習し, オンライン音源定位時にはパラメータを積分消去することで, 最尤推定に基づく通常の HMM よりも学習初期値などに頑健な結果を得る.

3.1.1 観測モデル

VB-HMM で用いる観測モデルを以下に示す.

$$p(\mathbf{x}_t|\mathbf{s}_t,\mu,\lambda) = \prod_{d=1}^{D}\prod_{j=0}^{1}\mathcal{N}(x_{t,d}|\mu_j,\lambda_j^{-1})^{\delta_j(s_{t,d})}, \quad (6)$$

ただし, $\delta_y(x)$ は $x = y$ のとき $\delta_y(x) = 1$, さもなくば $\delta_y(x) = 0$ を表す. また, $\mathcal{N}(\cdot|\mu,\lambda^{-1})$ は, 平均 μ, 精度 λ の正規分布の確率密度関数を表す. パラメータ μ と λ には共役事前分布として, 正規-ガンマ分布を用いる.

$$p(\mu,\lambda|\beta_0,m_0,a_0,b_0) = \prod_{j=0}^{1}\mathcal{N}(\mu_j|m_0,(\beta_0\lambda_j)^{-1})\mathcal{G}(\lambda_j|a_0,b_0), \quad (7)$$

ただし, $\mathcal{G}(\cdot|a,b)$ は形状 a, 尺度 b のガンマ分布である.

3.1.2 状態遷移モデル

状態遷移モデルは基本的に, 各方向ビン d について, 前状態で音源がない場合 $s_{t,d} = 0$ と音源がある場合 $s_{t,d} = 1$ から, 次状態で音源が出現する, 継続する, 消滅するといった遷移を考える. 本稿ではさらに, 移動する音源についても考慮するために, 表 1 のように前状態の組み合わせから成る 4 つの場合を考える. すなわち, 前時刻の同方向ビン $s_{t-1,d}$ に音源が存在するかどうかと, 前時刻の隣接方向ビン $s_{t-1,d\pm1}$ のいずれかに音源が存在するかによって分類する. 例えば, θ_1 は前時刻に当該方向 d 及び隣接ビン $d\pm1$ に音源が存在しない状態から音源が出現する確率, θ_2 は, 前時刻に方向 d に音源が存在しないが, 隣接ビン $d\pm1$ には音源が存在したため, その音源が方向 d に移動してきて $s_{t,d} = 1$ となる確率を表す. 状態遷移確率は以下の通り.

Table 1: 隣接状態を考慮した状態遷移の場合分け

前状態 $s_{t-1,d}$	隣接前状態 $1-s_{t-1,d-1}s_{t-1,d+1}$	音源存在確率 $p(s_{t,d}=1\|s_{t-1,d-1:d+1})$
0 (off)	0	θ_1
0 (off)	1	θ_2
1 (on)	0	θ_3
1 (on)	1	θ_4

$$p(\mathbf{s}_t|\mathbf{s}_{t-1},\theta) \quad = \quad \prod_{d=1}^{D}\prod_{k=1}^{4}\prod_{j=0}^{1}\left(\theta_k^{s_{t,d}}(1-\theta_k)^{1-s_{t,d}}\right)^{f_k(\mathbf{s}_{t-1},d)} \tag{8}$$

ここで, $f_k(\mathbf{s}_{t-1},d)$ は表 1 に従って, 方向ビン d の周りの前状態の値 $s_{t-1,d-1}, s_{t-1,d}, s_{t-1,d+1}$ によって条件 k に合致するときに $f_k(\cdot,d)=1$ その他の場合は 0 を返す条件識別関数である. 初期状態としては, 音源は存在しない, すなわちすべての d に対して $s_{0,d}=0$ とする.

状態遷移パラメータである $\theta=[\theta_1,...,\theta_4]$ には, 式 (8) の共役事前分布としてベータ分布を用いる.

$$p(\theta|\alpha_0) \quad = \quad \prod_{k=1}^{4}\mathcal{B}(\theta_k|\alpha_{0,1},\alpha_{0,0}), \tag{9}$$

ただし, $\mathcal{B}(\cdot|c,d)$ はパラメータ c,d を持つベータ分布の確率密度関数である.

3.1.3 事後分布の推定

VB-HMM の学習は, 事後分布 $p(\mathbf{s}_{1:T},\theta,\mu,\lambda|\mathbf{x}_{1:T})$ を以下のように因数分解可能な分布に近似して推定する.

$$\begin{aligned} p(\mathbf{s}_{1:T},\theta,\mu,\lambda|\mathbf{x}_{1:T}) \quad &\approx \quad q(\mathbf{s}_{1:T},\theta,\mu,\lambda), \\ &= \quad q(\mathbf{s}_{1:T})q(\theta)q(\mu,\lambda), \end{aligned} \tag{10}$$

$(\cdot)_{1:T}$ は, 時刻 1 から T までの確率変数の集合を表す. 式 (10) で近似される分布は, 下記の観測変数 $\mathbf{x}_{1:T}$ の対数エビデンスの下限 $\mathcal{L}(q)$ を最大化するよう更新する [Beal, 2003; Bishop, 2006].

$$\log p(\mathbf{x}_{1:T}) = \log\int p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)d\mathbf{s}_{1:T}d\theta d\mu d\lambda \geq \mathcal{L}(q),$$

$$\begin{aligned} \mathcal{L}(q) =& \mathbb{E}_{\mathbf{s}_{1:T},\theta,\mu,\lambda}\left[\log p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)\right] \\ &-\mathbb{E}_{\mathbf{s}_{1:T},\theta,\mu,\lambda}\left[\log q(\mathbf{s}_{1:T})q(\theta)q(\mu,\lambda)\right]. \end{aligned} \tag{11}$$

ただし, $\mathbb{E}_{\mathbf{s}_{1:T},\theta,\mu,\lambda}[\cdot]$ は分布 $q(\mathbf{s}_{1:T})q(\theta)q(\mu,\lambda)$ に関する期待値である. 式 (11) が極大値に収束するまで, 各分布は以下のように交互に更新される.

$$\begin{aligned} \log q(\mathbf{s}_{1:T}) =& \mathbb{E}_{\theta,\mu,\lambda}\left[\log p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)\right] \\ \log q(\theta) =& \mathbb{E}_{\mathbf{s}_{1:T},\mu,\lambda}\left[\log p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)\right] \\ \log q(\mu,\lambda) =& \mathbb{E}_{\mathbf{s}_{1:T},\theta}\left[\log p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)\right] \end{aligned}$$

事前分布の共役性により, 事後分布は結局以下のように分布のパラメータを更新することと等価である. $q(\theta)=\prod_k q(\theta_k)$ はそれぞれの k に対し, 式 (12) に示すパラメータ $\hat{\alpha}_{k,1}, \hat{\alpha}_{k,0}$ を持つベータ分布となり, $q(\mu,\lambda)=\prod_j q(\mu_j,\lambda_j)$ は, 式 (13), (14) のように, パラメータ $\hat{\beta}_j,\hat{m}_j,\hat{a}_j,\hat{b}_j$ を持つ正規ガンマ分布となる.

$$\hat{\alpha}_{k,j} = \alpha_{0,j} + \sum_{t,d}\langle s_{t,d,j}f_k(\mathbf{s}_{t-1},d)\rangle, \tag{12}$$

$$\hat{\beta}_j = \beta_0 + w_j, \hat{m}_j = (\beta_0 m_0 + w_j\bar{x}_j)/(\beta_0+w_j), \tag{13}$$

$$\hat{a}_j = a_0 + \frac{w_j}{2}, \hat{b}_j = b_0 + \frac{w_j S_j^2}{2} + \frac{\beta_0 w_j(\bar{x}_j-m_0)^2}{2(\beta_0+w_j)}, \tag{14}$$

ただし, 変数 $s_{t,d,j}$ は, $s_{t,d}=0$ のとき, $s_{t,d,0}=1$, また, $s_{t,d}=1$ のとき, $s_{t,d,1}=1$ となる変数である. 式 (13), (14) に用いられる正規分布の十分統計量は

$w_j=\sum_{t,d}\langle s_{t,d,j}\rangle$, $\bar{x}_j=\frac{\sum_{t,d}\langle s_{t,d,j}\rangle x_{t,d}}{w_j}$, $S_j^2=\frac{\sum_{t,d}\langle s_{t,d,j}\rangle(x_{t,d}-\bar{x}_j)^2}{w_j}$.

と定義する. また, $\langle\cdot\rangle$ は式 (10) の分布による期待値演算子である. $q(\mathbf{s}_{1:T})$ に対応する, 各時刻の状態変数と状態遷移の期待値 $\langle s_{t,d,j}\rangle$, $\langle s_{t,d,j}f_k(\mathbf{s}_{t-1},d)\rangle$ は次のように計算する.

$$\langle s_{t,d,j}\rangle \propto \alpha(s_{t,d,j})\beta(s_{t,d,j}), \tag{15}$$

$$\langle s_{t,d,j}f_k(\mathbf{s}_{t-1},d)\rangle \propto \tilde{\alpha}(s_{t-1,d,k})\tilde{p}(s_{t,d}|\mathbf{s}_{t-1})\tilde{p}(x_{t,d}|s_{t,d})\beta(s_{t,d,j}), \tag{16}$$

ただし, $\alpha(s_{t,d,j})$ と $\beta(s_{t,d,j})$ はそれぞれ前向き・後ろ向き再帰式により計算される.

$$\alpha(s_{t,d,j}) \propto \sum_{k=1}^{4}\tilde{\alpha}(s_{t-1,d,k})\tilde{p}(s_{t,d}|\mathbf{s}_{t-1})\tilde{p}(x_{t,d}|s_{t,d}), \tag{17}$$

$$\beta(s_{t,d,j}) = \sum_{j'=0}^{1}\beta(s_{t+1,d,j'})\tilde{p}(s_{t+1,d,j'}|s_{t,d,j})\tilde{p}(x_{t,d}|s_{t,d}). \tag{18}$$

式 (16) 遷移, 観測確率の幾何平均は次の通り.

$$\tilde{p}(s_{t,d}=j|\mathbf{s}_{t-1}) \propto \quad \prod_{k=1}^{4}\exp\left\{\psi(\hat{\alpha}_{k,j})-\psi(\hat{\alpha}_{k,0}+\hat{\alpha}_{k,1})\right\}^{f_k(\mathbf{s}_{t-1},d)}, \tag{19}$$

[2]

$$\tilde{p}(x_{t,d}|s_{t,d}) \propto \prod_j \exp\left\{\frac{\psi(\hat{a}_j)-\log\hat{b}_j-1/\hat{\beta}_j}{2} - \frac{a_j(x_{t,d}-\hat{m}_j)^2}{2\hat{b}_j}\right\}^{s_{t,d,j}} \tag{20}$$

式 (15), (16) はともに, 添字 j,k を動かしたとき総和が 1 になるように正規化されている. $\tilde{\alpha}(s_{t-1,d,k})$ は, 状態遷移の条件 k に関する前向き確率である. 本節で示されたパラメータ更新式 (12)–(16) が収束するまで計算される. 初期値としては, $\langle s_{t,d,j}\rangle$ と $\langle s_{t,d,j}f_k(\mathbf{s}_{t-1},d)\rangle$ の値を, 観測変数 $x_{t,d}$ の値を m_0 の値を閾値として処理することで, 0 ないし 1 を与えることで行う.

3.2 パーティクルフィルタによるオンライン音源定位

本節ではパーティクルフィルタ [Arulampalam et al., 2002] を用いた, オンライン音源定位手法を述べる. オンライン推定では, 式 (12)–(14) で求めたパラメータの事後分布を利用する. パーティクルフィルタの推定対象は, MUSIC スペクトルの時系列データが与えられたときの, 各方向ビンにおける音源存在事後確率である. この分布を P 個のパーティクルを用いて以下のように近似計算する.

$$p(\mathbf{s}_t|\mathbf{x}_{1:t}) \quad \approx \quad w_p\mathbf{s}_t^p, \tag{21}$$

[2] $\psi(\cdot)$ はディガンマ関数.

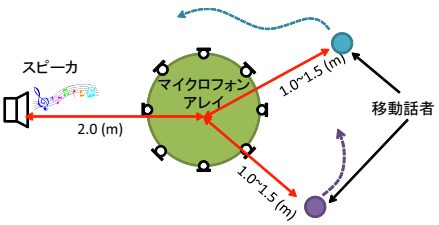

Figure 5: 実験条件: マイクロフォンアレイの周囲を動く移動話者と固定音源

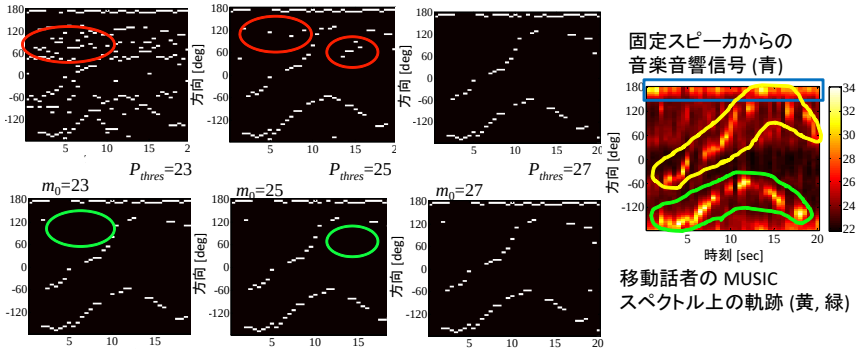

Figure 6: 音源定位結果: 白が音源が存在する方向, 時間ビン. 上図: 固定閾値 P_{thres} による定位結果. 下図: 初期値 m_0 を変えた場合の本手法による定位. 右図: 観測された対数 MUSIC スペクトル. 音楽音響信号が 180 [deg] 付近に存在し, 2 人の話者が移動している.

ただし, w_p はパーティクル p の重み, $\mathbf{s}_t^p$ は状態ベクトルの値である. これらの w_p と $\mathbf{s}_t^p$ は次のように得る.

(1) 提案分布から $\mathbf{s}_t^p$ をサンプルする.

$$\mathbf{s}_t^p \sim q(\mathbf{s}_t|\mathbf{x}_t, m, a, b), \tag{22}$$

$$q(\mathbf{s}_t^p|\mathbf{x}_t, \hat{m}, \hat{a}, \hat{b}) \propto \prod_d \prod_{j=0}^{1} C(x_{t,d})^{s^p_{t,d,1}} \exp(-\Delta^2_{d,j}/2)^{s^p_{t,d,j}}, \tag{23}$$

ただし, $x_{t,d}$ が極大値を取る d のとき, $C(x_{t,d}) = 1$ でその他の場合は $C(x_{t,d}) = 0$ となる. この項は, 時間 t の中で, $x_{t,d}$ の極大方向 d にのみ音源が存在する, つまり $s_{t,d} = 1$ となるよう導入されている. 提案分布の重みにはマハラノビス距離 $\Delta^2_{d,j} = (x_{t,d} - \hat{m}_j)^2 \hat{a}_j / \hat{b}_j$ を用いる.

(2) 各パーティクル p について, 重み w_p を算出.

$$w_p \propto \frac{\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p)\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p)}{q(\mathbf{s}_t^p|\mathbf{x}_t, \hat{m}, \hat{a}, \hat{b})}, \tag{24}$$

$$\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p) = \prod_d C(x_{t,d})^{s^p_{t,d,1}} \int p(\mathbf{x}_t|\mathbf{s}_t^p, \mu, \lambda) q(\mu, \lambda) d\mu d\lambda, \tag{25}$$

$$\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p) = \int p(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p, \theta) q(\theta) d\theta. \tag{26}$$

式 (25),(26) にある状態遷移, 観測確率は, VB-HMM で計算された式 (6),(8) の事後分布で積分消去することで計算できる. これにより, VB-HMM で学習されたパラメータの曖昧性を考慮したオンライン定位を行う. なお, 式 (25) の $C(x_{t,d})^{s^p_{t,d,1}}$ の項は, 式 (23) と同様に, $x_{t,d}$ の極大方向 d にのみ音源の存在を許す項である. 分布の共役性を用いると, この積分計算は次のように解析的に求まる.

$$\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p) = \prod_d C(x_{t,d})^{s^p_{t,d,1}} \mathcal{S}t(x_{t,d}|\hat{m}_j, \frac{\hat{\beta}_j \hat{a}_j}{(1+\hat{\beta}_j)\hat{b}_j}, 2\hat{a}_j)^{s^p_{t,d,j}}, \tag{27}$$

$$\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p) = \prod_d \prod_k \left(\hat{\alpha}_{k,s_{t,d}} / (\hat{\alpha}_{k,0} + \hat{\alpha}_{k,1})\right)^{f_k(\mathbf{s}^p_{t-1}, d)}, \tag{28}$$

ただし, $\mathcal{S}t(\cdot|m, \lambda, \nu)$ は平均 m, 精度 λ, 自由度 ν の Student t-分布である. さらに, 最大の音源数を N_{max} に抑えるため, 状態ベクトル $\mathbf{s}_t^p$ に存在する音源数が N_{max} を超える場合には観測確率は 0 とする.

全パーティクルの重み計算後, 各パーティクルの重み w_p は $\sum_{p=1}^P w_p = 1$ となるよう正規化する. この手順に従い, 式 (21) の音源存在の事後分布を計算する. 我々の実装手法では, 各ステップごとにパーティクルが持つ重みに比例してリサンプリング処理が行われる.

4 評価実験

評価実験では, VB-HMM によるパラメータ分布推定とパーティクルフィルタを用いたオンライン音源定位から成る本手法と, 従来の固定閾値を用いて音源定位する手法を比較する. オフラインでの VB-HMM での学習は, 1 人の話者がマイクロフォンの周囲を発話しながら動く音響信号で行った. オンラインの音源定位実験に使用した音源の配置を図 5 に示す. マイクロフォンアレイの周囲を移動する 2 話者と, 固定されたスピーカから音楽が再生されている. オフライン, オンラインで用いられた信号の長さはともに 20 (sec) である. パラメータの設定は次の通り. 観測信号の自己相関行列を計算する窓幅 $\Delta T = 500$ (msec), $N_{max} = 3$, $\alpha_0 = [1, 1]$, $\beta_0 = 1$, $a_0 = 1$, $b_0 = 500$. パーティクル数は $P = 500$ とした. 実験で使用した室内の残響時間は $RT_{20} = 840$ (msec) であった.

図 6 にオンライン音源定位の結果を示す. 従来法の閾値は $P_{thres} = 23, 25, 27$ に設定されており, 本手法の初期値は $m_0 = 23, 25, 27$ に設定されている. パーティクルフィルタの定位結果の図では, 事後分布の音源存在確率が 0.95 以上のビンを音源が存在するとして白く表示している. 従来法においては, 閾値を低く設定した場合は図 6 の赤枠で示すように音源の誤検出が頻発する. 対して, 本手法では緑枠で示すように, 学習の初期値に対して頑健に妥当な音源定位結果を示している. また, 本手法において音源存在確率の閾値を 0.95–1.00 まで動かして結果を検証したが, こ

れらの値を閾値に対しても頑健に同様の結果を示すことを確認した. この結果から, 本手法におけるオフライン学習, オンライン定位の枠組みが, 自動的に音源定位に適したパラメータに収束することが確認できる. さらに, 今回の実験条件から, 本手法は学習時に 1 音源しか用いなくても, 複数音源に対して安定したオンライン定位が可能であることが確認された.

4.1 議論と今後の課題

実験を通じて, 本手法は学習初期値や, 学習時とオンライン推定時の音源数ミスマッチに頑健であることを示した. しかし, 本手法には次の制約が存在する. (1) 音源ごとの軌跡は直接は推定されない. (2) 音声のポーズ等に応じて定位が途切れる. 混合音に含まれる各音源の定位結果を元に音源分離を行うシステムでは (例 [Nakadai et al., 2010]), 安定した音源分離のために音源ごとの軌跡, ポーズ等を接続した定位が重要である.

(1) 本稿で示した状態空間モデルでは, 各時間フレームで音源が存在する方向ビンを推定する. 音源ごとのトラッキング結果が必要な場合, 連続する時間フレームで近い定位結果をスムージングするといった後処理や, あるいは, 複数音源の移動を状態遷移モデルに組み込む必要がある.

(2) 音声では文の終わり等にポーズがしばしば入り, 対応する時間フレームの MUSIC スペクトルの値は減少する. 図 6 で示された本手法による定位結果でも, 0°, 6 (sec) 付近の話者の定位が途切れている様子が示されている. この問題も, 後処理で途切れた軌跡をつなげるといった方法や, 音のポーズを明示的に状態モデルに取り込むといった手法の改良による対処が考えられる.

5 まとめ

本稿では MUSIC 法に基づく音源定位法のベイズ拡張を述べた. 本手法は, (1) VB-HMM によるパラメータの自動学習, (2) パーティクルフィルタを用いたオンライン音源定位から成る. 評価実験では, $RT_{20} = 840$ (msec) の残響環境下で, 1 音源の音響信号の学習に対し, 3 音源同時音源定位を実現した. 今後の展開としては, 実際に移動ロボットに本手法を適用して, ロボット位置と環境中に存在する音源位置の推定を通した音環境理解システムの構築などが挙げられる.

謝辞: 本研究の一部は科研費特別研究員奨励金/基盤 (S), JST-ANR BINAAHR, GCOE の支援を受けた.

参考文献

[Arulampalam et al., 2002] M. Arulampalam et al.: A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking, *IEEE Trans. on Signal Proc.*, Vol. 50, No. 2, pp. 174–189, 2002.

[Asano et al., 2001] F. Asano et al.: Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition, *Proc. of Eurospeech*, pp. 1013–1016, 2001.

[Beal, 2003] M. J. Beal: Variational Algorithms for Approximate Bayesian Inference, *Ph.D. thesis*, Gatsby Computational Neuroscience U., Univ. Colledge London, 2003.

[Bishop, 2006] C. M. Bishop: Chapter 10, Approximate Inference, *Pattern Recognition and Machine Learning*, Springer, 2006.

[Danès et al., 2010] P. Danès and J. Bonnal: Information-Theoretic Detection of Broadband Sources in a Coherent Beamspace MUSIC Scheme, *Proc. of IROS*, pp. 1976–1981, 2010.

[Doclo et al., 2001] S. Doclo and M. Moonen: GSVD-based optimal filtering for multi-microphone speech enhancement, *Microphone arrays*, pp. 111–132, Springer, 2001.

[Kubota et al., 2008] Y. Kubota et al.: Design and Implementation of 3D Auditory Scene Visualizer towards Auditory Awareness with Face Tracking, *Proc. of IEEE Int'l Symboposium on Multimedia (ISM-2008)*, pp. 468–476, 2008.

[Mizumoto et al., 2011] T. Mizumoto et al.: Design and Implementation of Selectable Sound Separation on a Texai Telepresence System using HARK, *Proc. of ICRA*, pp. 2130–2137, 2011.

[Nakadai et al., 2010] K. Nakadai et al.: Design and Implementation of Robot Audition System "HARK", *Advanced Robotics*, Vol. 24, No. 5–6, pp. 739–761, 2010.

[Sasaki et al., 2010] Y. Sasaki et al.: Map-Generation and Identification of Multiple Sound Sources from Robot in Motion, *Proc. of IROS*, pp. 437–443, 2010.

[Schmidt, 1986] R. O. Schmidt: Multiple Emitter Location and Signal Parameter Estimation, *IEEE Trans. on Antennas and Propagation*, Vol. 34, No. 3, pp. 276–280, 1986.

[Yamamoto et al., 2006] K. Yamamoto et al.: Detection of Overlapping Speech in Meeting using Support Vector Machines and Support Vector Regression, *IEICE Trans. Fundamentals*, Vol. E89-A, No. 8, pp. 2158–2165, 2006.