社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-9

ベイズモデルによるマイクロホンアレイ処理の移動ロボットへの応用

Bayesian Microphone Processing and Its Application to Mobile Robot Audition

大塚 琢馬 †, 石黒 勝彦 ‡, 澤田 宏 ‡, 奥乃 博 †

Takuma Otsuka†, Katsuhiko Ishiguro‡, Hiroshi Sawada‡, Hiroshi G. Okuno†

† 京都大学大学院情報学研究科, ‡ 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

†**Graduate school of Informatics, Kyoto University,** ‡ NTT Communication Science Laboratories, NTT Corporation

Abstract

This paper addresses a simultaneous sound source localization and separation method using a microphone array and its application to an auditory sensing by a mobile robot. The auditory function embedded in the mobile robot should deal with (1) the uncertainties in the environment such as the unknown reverberation and source number, (2) time-varying sound source locations observed by the robot, and (3) the motor noise caused by the robot motion. Our Bayesian formulation is employed to efficiently cope with the uncertainties. Sound source separation experiments in indoor and outdoor environments confirm encouraging results.

1 はじめに

ロボットが環境中を移動しながら，自身に備え付けられたセンサを用いてロボットがいる環境から情報を抽出することは，ロボットによる自律的に環境の探索，あるいは，遠隔のロボット操作者のナビゲーションにとって重要である．従来の移動ロボットによるセンシング技術は，カメラからの視覚情報に基づく自己位置推定と環境地図作成 (SLAM; Simultaneous Localization and Mapping) [Thrun *et al.*, 2004; Se *et al.*, 2005] などを中心に発達してきた．これらの視覚情報処理に加えて，ロボットが聴覚情報を扱えるようになると，次のような機能強化が期待できる．(1) 物体のオクルージョンに対する頑健性の獲得，(2) ものの変化の知覚，(3) 音声コミュニケーションの実現．例えば，(1) 視覚のみに頼ると壁の向こうの情報は取得出来ないが，音を聴くことで壁に遮られた場所の知覚を試みることが出来る．(2) 物体が動いたり状況が変化する場合には音を伴うことが多い．例として，Figure 1 上のように，グラスが机から落ちた場合は「ガシャン」と音がする．音環境理解機能を持つロボットであれば，このような出来事に気づきやすくなることが期待できる．(3) もちろん，人間の音声を指令として受け取るなどのコミュニケーションチャネルへの寄与も考えられる．

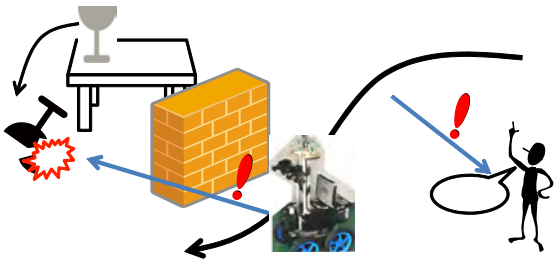

Figure 1: 移動ロボットにおける聴覚機能

ロボットが存在する環境中に複数の音源が同時に存在することがある．そのため，ロボットには複数音源の混合観測音から，個々の音源に分解する「聴き分け」機能が必須である．このような複数音源を扱うため，複数のマイクロホンを利用するマイクロホンアレイ処理 [Benesty *et al.*,] がよく用いられる．本稿では，マイクロホンアレイを用いた混合観測音から，各音源のある方向を推定する音源定位と，各音声信号を抽出する音源分離法を示し，移動ロボットの適用について述べる．マイクロホンアレイ処理に関する最も重要な課題の 1 つは，観測音中に含まれる未知要因を対処するという点である．マイクロホンアレイ処理の性能を左右する環境中の未知要因としては，観測音に含まれる音源数や，音源とマイク間の相対位置や周囲の壁などに依存する残響などが挙げられる．これらの未知要因に対して，本手法では音源数や残響に関する仮定がなく，入力データである観測音からこれらの情報も柔軟に扱うことの出来るベイズモデルに基づく音源定位・分離法を示す．

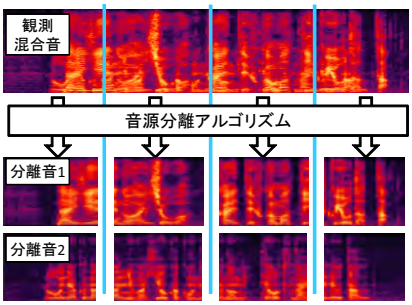

Figure 2: 時分割処理: 各区間内の空間特性は定常と仮定

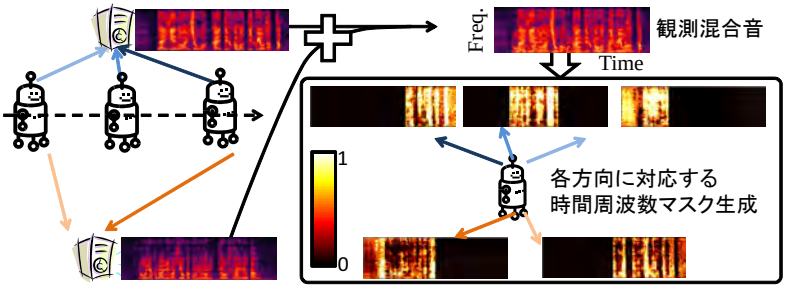

Figure 3: 方向ごとの時間周波数マスク推定: 音源が通過した方向に対応するマスクを生成

2 移動ロボットの聴覚機能における諸課題

ロボットが移動しながら混合音を観測し，個々の音源を分離する際には，音環境の未知要因の存在に加えて，2 つの大きな課題が存在する．

1. ロボットに搭載されたマイクロホンアレイと音源との間の相対的な位置関係 (以下，空間特性と呼ぶ) が時間変化する点と，
2. ロボットの移動などに伴って，車輪を動かすモータ音や地面の段差を乗り越える音が観測音に混入するため，目的となる外部音の signal-to-noise ratio (SNR) が劣化する点である．

マイクロホンアレイは空間フィルタリング機能を持つ．つまり，マイクロホンアレイによる音源分離は原則として，異なる方向から到来する音源の分離が可能である．従って，分離対象の音源は空間的にスパースに存在することを仮定する．また，マイクロホンアレイによる音源定位のためには，音源到来方向と各マイクへの波面到達時間差などの対応付けが必要なため，各マイクロホンの正確な配置や，ステアリングベクトルや伝達関数と呼ばれる波面到達時間差やマイク間の観測音振幅比などの事前情報が必要となる．さらに，多くのマイクロホンアレイによる音源定位や音源分離は定常な空間特性を仮定しているため，空間特性が時変である移動ロボット問題への適用には工夫が必要である．

ロボット聴覚ソフトウェア HARK [Nakadai *et al.*, 2010] は，Figure 2 のように，観測音を時分割し，各区間内では定常な空間特性を仮定した上で分離処理を行う．より具体的には，0.5 [s] 程度の固定窓幅で音源定位を行い，定常方向に各音源が定位された区間ごとに分離処理を行う．このシステムは実時間での音源定位・分離を実現するが，定位が失敗すると分離性能が大きく劣化するほか，精確な音源定位のためには音源数を与える必要や，残響時間など環境に依存したパラメータを設定する必要がある．また，音源分離も環境に依存する伝達関数を事前知識として要すること，環境中の音源数がマイク数未満である必要があるなど，環境依存性が課題として残されている．また，ロボット動作などに伴う自己発生音に対しては，マイクロホンアレイに対する自己発生音源の方向を指定することで，自己発生音を抑圧した信号の抽出を行う．この方法は，自己発生音源の位置がマイクロホンアレイに対して相対的に変わらない場合に有効である．

本稿で示す手法も定常な空間特性を仮定した手法であるが，Figure 3 のように時間変化する空間特性に対応する．Figure 3 左側のように，移動しながら音源を観測すると，ロボットから見た音源方向は時間変化する．このように観測した混合音に対して，十分大きなクラスタ数を用いて本手法による分離を行うと，Figure 3 右側のように，観測音の中で音源が通過した方向に応じた時間周波数マスク (TF マスク) が自動的に生成される．従って，Figure 2 のように定常空間特性を仮定できる十分小さな窓幅などを設定する必要なく，観測音中の各音源の相対的な移動に適した時間幅での音源分離が望める．このように，本手法は環境依存の要素が極力覗かれている点が利点であるが，クラスタリングの反復計算に由来する計算時間の大きさが欠点として挙げられる．

3 統一的音源定位・分離ベイズモデル

本手法は音源定位・分離問題を時間周波数領域でのクラスタリング問題として扱う．音源分離は各時間周波数点の信号ベクトルのクラスタリング問題とし，音源定位は各クラスとステアリングベクトルとの対応付け問題とする．パーミュテーション解決を含む分離処理には Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] を利用する．基本的なアイデアは，各時間フレームにおいて特定の音源が多く出現するようモデル化することで，各音源のマスクが周波数ビンをまたいで時間的に同期させるという仕組みである [Otsuka *et al.*, 2012]．通常の LDA は有限混合モデルであるため，本稿では音源数が無制限の HDP [Teh *et al.*, 2006] へと拡張する．

本節で用いる記号を Table 1 にまとめる．Figure 4 は確率変数の条件付き独立性を図示したグラフィカルモデルである．二重円で示された $\mathbf{x}_{tf}$ は観測変数を表し，円で示

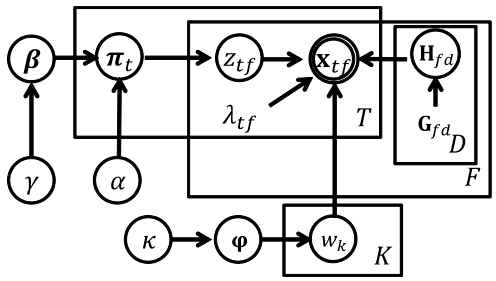

Figure 4: グラフィカルモデル

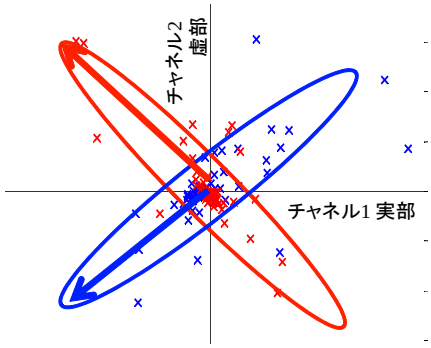

Figure 5: 周波数ビン 3200 Hz での観測信号散布図

Table 1: 記号表

記号	意味
t	時間フレーム ($1 \leq t \leq T$)
f	周波数ビン ($1 \leq f \leq F$)
k	クラスインデックス ($1 \leq k \leq K$)
d	方向インデックス ($1 \leq D \leq D$)
M	マイク数
$\mathbf{x}_{tf}$	観測信号 M 次元列ベクトル
z_{tf}	時間フレーム t 周波数ビン f のクラス
π_t	時間フレーム t でのクラス割合
w_k	クラス k の方向
φ	全クラスに対する方向割合
λ_{tf}	$\mathbf{x}_{tf}$ の逆数スケール
$\mathbf{H}_{fd}$	周波数ビン f 方向 d の精度行列
n_{tk}	時間フレーム t のクラス k サンプル数
n_{fk}, n_{fd}	周波数ビン f のクラス k，方向 d サンプル数
c_d	方向 d に割り当てられたクラス数

された記号は潜在確率変数，囲いのない変数は定数を表す．3.1 節では多チャンネル観測信号とステアリングベクトルの関係を示し，3.2 節に CGS による推論，3.3 節では定位，分離結果の出力法を説明した後，3.4 節で推論の初期化方法を述べる．変数集合は下付添字を略しチルダを付けて表す (例: $\tilde{\mathbf{x}} = \{\mathbf{x}_{tf} | 1 \leq t \leq T, 1 \leq f \leq F\}$)．次節以降で詳細は述べるが，$\tilde{\mathbf{z}}$ の推論が TF マスク推定による分離処理に相当し，$\tilde{\mathbf{w}}$ が定位に相当する．

3.1 多チャンネル複素信号の生成モデル

本節では Figure 4 に示された生成プロセスを説明する．時間周波数領域の多チャンネル信号観測モデルとして covariance mdoel [Duong *et al.*, 2010] を採用する．このモデルでは，各時間周波数点は平均ゼロ，スケール時変共分散の多変量複素正規分布に従うとする．Figure 5 に青と赤で示される 2 つの音源を 2 チャンネルで観測した信号の散布図を示す．これらの点は次のように生成されたと仮定する．時間 t，周波数ビン f の時間周波数点で優勢な信号は方向 d から到来しているとき，観測信号は $\mathbf{x}_{tf} = s_{tf}\mathbf{q}_{fd}$ として表されるとする．但し，s_{tf} はその点における音源成分とし，$\mathbf{q}_{fd}$ は方向 d に対応するステアリングベクトルである．ベクトル $\mathbf{x}_{tf}$ は M 次元ベクトルで，各成分は対応するマイクでの観測に対応する．このベクトルの共分散は $\mathbb{E}[\mathbf{x}_{\approx\upsilon}\mathbf{x}_{\approx\upsilon}^{\mathbb{H}}] = \mathbb{E}[|\sim_{\approx\upsilon}|^{\nvDash}\mathbf{q}_{\upsilon}\mathbf{q}_{\upsilon}^{\mathbb{H}}]$ と書ける．ただし，$\cdot^H$ はエルミート転置を表す．

Figure 5 に楕円で示された各音源の共分散行列は値の大きな固有値を持つ．これに対応する固有ベクトル (図中矢印) は，その音源がある方向に対応するステアリングベクトルと同一方向である．すなわち，各時間周波数点の属する楕円の推定は音源分離に対応し，クラスタの共分散の固有ベクトルを調べることは音源定位に対応する．

各時間周波数点の共分散は時変スケール $|s_{tf}|^2$ と固定の方向行列 $\mathbf{q}_{fd}\mathbf{q}_{fd}^H$ へと分解出来る．行列部分が固定されることが，音源位置が一定であるという仮定に対応する．ここで分布の共役性を取り入れるために共分散行列の逆数である精度行列 $\lambda_{tf}\mathbf{H}_{fd}$ を導入する．λ_{tf} は時変スケールであり，$\mathbf{H}_{fd} \approx (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \varepsilon\mathbf{I})^{-1}$ とする．$\mathbf{I}$ は単位行列である．λ_{tf} は [Otsuka *et al.*, 2012] では確率変数として推論対象であったが，本手法は $\lambda_{tf} = |\mathbf{x}_{tf}|^{-1}$ として値を固定する．この結果，$\mathbf{H}_{fd}$ を積分消去し，効率的な周辺化推論が可能となる．複素正規分布に従う尤度関数は次のように表される．

$$\mathbf{x}_{tf} | z_{tf}, \tilde{\mathbf{w}}, \lambda_{tf}, \tilde{\mathbf{H}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_{tf} | \mathbf{0}, (\lambda_{tf}\mathbf{H}_{fw_{z_{tf}}})^{-1}), \quad (1)$$

ここで，z_{tf} と w_k はそれぞれ $\mathbf{x}_{tf}$ のクラス，クラス k の方向を表す．従って，$w_{z_{tf}}$ は $\mathbf{x}_{tf}$ が属する方向となる．平均 μ，精度行列 Λ の複素正規分布 $\mathcal{N}_{\mathbb{C}}(\mathbf{x}|\mu, \Lambda^{-1})$ の確率密度関数の定義は $\frac{|\Lambda|}{\pi^M}\exp\{-(\mathbf{x}-\mu)^H\Lambda(\mathbf{x}-\mu)\}$ である [van den Bos, 1995]．$|\Lambda|$ は行列 Λ の行列式である．方向行列 $\mathbf{H}_{fd}$ は共役事前分布である複素ウィシャート分布 [Conradsen *et al.*, 2003] に従う．

$$\mathbf{H}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd} | \nu_{fd}, \mathbf{G}_{fd}), \quad (2)$$

複素ウィシャート分布 $\mathcal{W}_{\mathbb{C}}(\mathbf{H}|\nu, \mathbf{G})$ の確率密度関数は $\frac{|\mathbf{H}|^{\nu-M}\exp\{-\mathrm{tr}(\mathbf{H}\mathbf{G}^{-1})\}}{|\mathbf{G}|^{\nu}\pi^{M(M-1)/2}\prod_{i=0}^{M-1}\Gamma(\nu-i)}$ と定義される．ここで，$\mathrm{tr}(\mathbf{A})$ は行列 $\mathbf{A}$ の跡，$\Gamma(x)$ はガンマ関数である．ハイパーパラメータは $\nu_{fd} = M$，$\mathbf{G}_{fd} = (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \varepsilon\mathbf{I}_M)^{-1}$ と定める．$\mathbf{G}_{fd}$ は所与のステアリングベクトル $\mathbf{q}_{fd}^H\mathbf{q}_{fd} = 1$ と正規化して利用する．ε は逆行列演算のために導入し，0.01 を用いた．

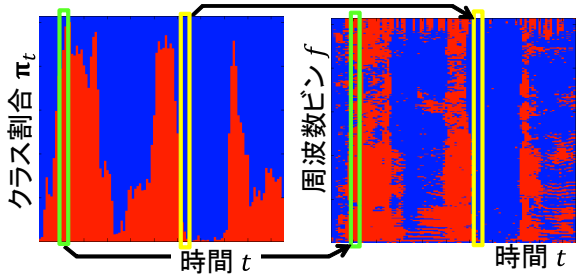

Figure 6: 左: クラス割合，右: 2 音源の TF マスク.

LDA の無限混合拡張である HDP [Teh *et al.*, 2006] を z_{tf} の生成過程として利用する．このモデルはパーミュテーション解決のために導入する [Otsuka *et al.*, 2012]．まず無限個のクラスの大域的な割合 β が生成され，時間ごとのクラス割合 π_t が β を元に生成される．最後に，各時間周波数点において z_{tf} が π_t に基づいて生成される．Figure 6 に示すように，各音源が優勢になる様子は周波数ビンをまたいで時間的に同期する様子が観察される．このことから，π_t を導入することで各クラスが時間ごとに同期する振る舞いをもたせることでパーミュテーション解決に寄与できると期待できる．HDP の棒折り過程による生成は次のように表される [Teh *et al.*, 2006].

$$\beta|\gamma \sim \mathrm{GEM}(\gamma), \quad \pi_t|\alpha,\beta \sim \mathrm{DP}(\alpha,\beta), \quad z_{tf}|\pi_t \sim \pi_t, \quad (3)$$

ただし，GEM(γ) は集中度 γ の Griffiths-Engen-McCloskey 分布，DP(α,β) は集中度 α，基底測度 β のディリクレ過程である．集中度の事前分布としてはガンマ分布を $\gamma \sim \mathcal{G}(\gamma|a_\gamma,b_\gamma)$，$\alpha \sim \mathcal{G}(\alpha|a_\alpha,b_\alpha)$ として用いる．ハイパーパラメータは $a_\gamma = 0.05, b_\gamma = 5, a_\alpha = 0.01, b_\alpha = 1$ とした．

方向変数 w_k は音源定位だけでなく，パーミュテーション解決にも寄与する．なぜなら，全周波数ビンに渡って生成されたクラスを同一方向から到来する信号として選別するためである．この潜在変数はディリクレ分布から生成される方向割合 φ に従う．

$$\varphi|\kappa \sim \mathcal{D}(\varphi|\frac{\kappa}{D}\mathbf{1}_D), \qquad w_k|\varphi \sim \varphi, \qquad (4)$$

ただし，$\mathbf{1}_D$ は要素がすべて 1 の D 次元ベクトルである．$\mathcal{D}(\cdot|\alpha)$ はパラメータ α に従うディリクレ分布を表す．本モデルはマイクロホンアレイの空間解像度は有限であるため，有限の方向数 D を扱う．κ に対してもガンマ分布 $\mathcal{G}(\kappa|a_\kappa,b_\kappa)$，$a_\kappa = 1, b_\kappa = 1$ を事前分布として利用する．

3.2 周辺化ギブズサンプラ (CGS) による推論

音源分離・定位問題には $\tilde{\mathbf{z}}$ と $\tilde{\mathbf{w}}$ の推論が鍵となる．これらの潜在変数は $\pi,\varphi,\tilde{\mathbf{H}}$ を積分消去した次の CGS

$$p(z_{tf}=k|\tilde{\mathbf{x}},\vartheta\backslash z_{tf}) \propto (\alpha\beta_k + n_{tk}^{\neg tf})$$
$$\frac{\Gamma(\hat{\nu}_{fw_k}^{\neg tf}+1)}{\Gamma(\hat{\nu}_{fw_k}^{\neg tf}-M+1)} \frac{\left|\mathrm{inv}(\hat{\mathbf{G}}_{\mathrm{fw_k}}^{\neg \mathrm{tf}})\right|^{\hat{\nu}_{fw_k}^{\neg tf}}}{\left|\mathrm{inv}(\hat{\mathbf{G}}_{\mathrm{fw_k}}^{\neg \mathrm{tf}}) + \lambda_{\mathrm{tf}}\mathbf{x}_{\mathrm{tf}}\mathbf{x}_{\mathrm{tf}}^{\mathrm{H}}\right|^{\hat{\nu}_{fw_k}^{\neg tf}+1}}, \qquad (5)$$

$$p(w_k=d|\tilde{\mathbf{x}},\vartheta\backslash w_k) \propto (\frac{\kappa}{D} + c_d^{\neg k})$$
$$\prod_f \left\{ \prod_{i=0}^{M-1} \frac{\Gamma(\hat{\nu}_{fd}^{\neg k}+n_{fk}-i)}{\Gamma(\hat{\nu}_{fd}^{\neg k}-i)} \frac{\left|\mathrm{inv}(\hat{\mathbf{G}}_{\mathrm{fd}}^{\neg \mathrm{k}})\right|^{\hat{\nu}_{fd}^{\neg k}}}{\left|\mathrm{inv}(\hat{\mathbf{G}}_{\mathrm{fd}}^{\neg \mathrm{k}}) + \sum_{\mathrm{t:z_{tf}=k}} \lambda_{\mathrm{tf}}\mathbf{x}_{\mathrm{tf}}\mathbf{x}_{\mathrm{tf}}^{\mathrm{H}}\right|^{\hat{\nu}_{fd}^{\neg k}+n_{fk}}} \right\}, \qquad (6)$$

によって確率的に生成する．ただし，$\vartheta\backslash z$ は z を除く全ての潜在変数の集合で，上付添字 $\neg tf$ と $\neg k$ は点 tf やクラス k を除いた統計量を表す．また，inv$(\cdot)$ は逆行列である．$\hat{\nu}_{fd}, \hat{\mathbf{G}}_{fd}$ は十分統計量を用いて次のように与えられる．

$$\hat{\nu}_{fd} = \nu_{fd} + n_{fd}, \qquad \hat{\mathbf{G}}_{fd}^{-1} = \mathbf{G}_{fd}^{-1} + \sum_{t:w_{z_{tf}}=d} \lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^H, \qquad (7)$$

ここで，$\sum_{t:w_{z_{tf}}=d}\cdot$ は周波数ビン f で方向 d に割り当てられた時間周波数点に関する和である．

K を推論時に生成されているクラス数とする．z_{tf} が未生成のクラス $K+1$ を取る確率を式 (5) 中で考慮するため，β は $K+1$ 次元として操作する [Teh *et al.*, 2006]．$z_{tf} = K+1$ の確率計算のため，$w_{K+1} = d$ を一時的に $\frac{\kappa/D+c_d}{\kappa+\sum_d c_d}$ に従い生成する．もし $z_{tf} = K+1$ が選択された場合，$K \leftarrow K+1$ とし，β の次元も $\beta_K \leftarrow b\beta_K$，$\beta_{K+1} \leftarrow (1-b)\beta_K$ として増やす．ただし，b はベータ分布 $\mathcal{B}(1,\gamma)$ から生成する．

その他のパラメータ $\alpha,\beta,\gamma,\kappa$ の更新は [Teh *et al.*, 2006; Escobar and West, 1995] の詳述されるように更新する．これらの変数は補助変数を導入して確率的に生成される．

3.3 音源定位・分離結果出力

ξ_{tf}^d を推論時にサンプルされた z_{tf}, w_k のうち，$w_{z_{tf}} = d$ である割合，同様に，η_k^d をサンプルされた w_k の中で $w_k = d$ である割合とする．音源分離は，同一方向から到来するクラスを統合した TF マスクを用いる．方向 d から到来する多チャンネル信号を $\hat{\mathbf{x}}_{tf}^d$ とすると，$\hat{\mathbf{x}}_{tf}^d = \xi_{tf}^d\mathbf{x}_{tf}$ に従って分離が可能となる．また，各方向の事後重み $P_d = \sum_{tf}\xi_{tf}^d$ を計算することでどの方向に音源が存在するか推定することが出来る．もし混合音から N 個の音源を抽出したい場合，P_d の大きなものから N 方向を選び，音源を抽出することで，音源の定位と分離が達成される．

3.4 推論初期化

推論の初期化は [Otsuka *et al.*, 2012] に似た方法で行う．推論開始時のクラス数を K とする．まず，w_k を K 個の重複のない領域から一様分布に従って生成する．$p(w_k = d) = \mathcal{U}(\{d|\frac{k-1}{K}D \le d < \frac{k}{K}D\})$，$\mathcal{U}(A)$ は集合 A 上の一様分布である．次に，z_{tf} を初期化された w_k とステアリングベクトルから生成された $\mathbf{G}_{fd}$ を用いて生成する，$p(z_{tf}=k) \propto \exp\left(-\mathbf{x}_{tf}^H\mathbf{G}_{fw_k}\mathbf{x}_{tf}\right)$.

4 実験結果

本節では，分離実験に用いた混合音の収録条件と音源分離結果を示す．Figure 7 に収録環境における音源配置とロボットの移動軌跡，および，ロボットに搭載されたマイク

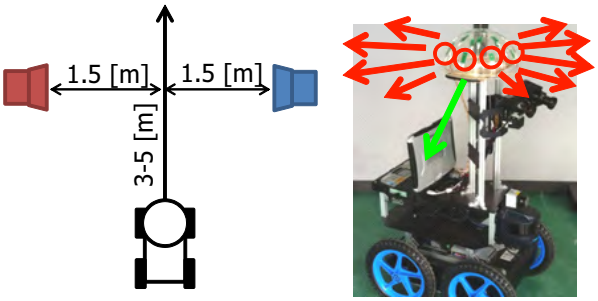

Figure 7: 音源配置とロボットの軌道 (左) とロボットに搭載されたマイクロホンアレイ (右)

ロホンアレイを示す．Figure 7 左のように，実験では 2 音源の間を直線的に移動しながら収録した．その際，片方の音源は常にロボットの左側，もう片方の音源はロボットの右側となるように動いた．これは，Figure 3 のように，様々な方向から分離された複数の分離音を同一の音源にまとめる処理を簡単化するためである．収録にはロボット上部に備え付けられた 8 チャネルマイクロホンアレイを用い，2 種類の環境で行った．1 つは屋外で，残響時間 $RT_{60} = 150$ [ms]，もう 1 つは屋内で，$RT_{60} = 800$ [ms] 以上の環境である．収録音は音源の種類に対する頑健性を示すため，Figure 7 中，青で示された右側スピーカからはピアノやギターから成る音楽音響信号を，赤で示された左側スピーカからは人間の音声や，鈴虫，カエルの鳴き声などを再生した．録音データの長さはおよそ 10 [s] であった．いずれの環境においても，多少の凹凸はあるがおおむね平坦な床面での走行を行い，ロボット本体の揺れを極力抑えた．

音源定位処理での方向の候補数 D の決定はマイクロホンアレイの持つ空間解像度などを考慮して行う．例えば，水平面上を 5° の解像度で定位を行う場合は，$D = \frac{360}{5} = 72$ と設定する．本実験では，ロボットの移動時に発生する車輪音を抑圧するために，上記の水平面上 72 方向 (Figure 7 右の赤矢印) のステアリングベクトルに加えて，ロボットの荷台方向 (Figure 7 右の緑矢印) のステアリングベクトルを用いて $D = 73$ とした．これにより，モータノイズなどはロボットの荷台方向として定位される音源に分離されることが期待できる．

5　実験結果

Figure 8，9 に，屋内，屋外それぞれの環境の観測音，分離音，再生された原音のスペクトログラムを示す．混合音と分離音に示された緑の枠は，その時間区間でロボットが移動していたことを示す．音源分離結果は Figure 3 のように同一音源でも様々な方向に分割された結果が得られるが，ロボットからみて左右どちらの方向に定位されているかに基いて各方向の音源を復元した．

ロボットの荷台方向に定位された車輪分離音について Figure 8，9 を比較すると，屋外環境については走行時以外に抽出された音はあまりないが，屋内環境については走行時以外も音声などが含まれている．さらに，屋内環境での車輪分離音の低周波領域では，右側分離音に含まれるべき成分を多く含んでいる．このように，残響の多い環境においては，直接音のみを対処しようとする音源分離手法の性能は特に低周波領域において劣化する．

Figure 8 での屋外環境での左右分離音は，左側音源の前半の虫の鳴き声は右側分離音や車輪分離音に埋もれてしまったが，ロボット移動中でもある程度音源分離が達成されている．虫の鳴き声の分離が特に困難な一因としては，この音源が比較的狭い周波数帯域のみにエネルギが集中しているため，他音源がこの領域でエネルギが大きかった場合に時間周波数マスクの推定が影響を受けるためと考えられる．一方，Figure 9 に示された左右分離音については，特にロボットが移動中の右側分離信号の抽出精度が劣化している．分離精度低下の要因としては残響の他，観測音中に含まれる右側の音楽音響信号の割合が少ない (SNR，signal to noise ratio が低い) ことが挙げられる．

以上のように，本手法には残響，狭帯域音，低 SNR 音などに伴う分離性能の劣化という限界はあるものの，(1) 異なる残響環境に対してパラメータなどの手動設定なしに，(2) ロボット自身の移動に伴って空間特性が時間変化する観測信号および，(3) 自己発生音の抑圧，を扱うことが可能であることが示された．

6　考察と今後の課題

本稿では，移動ロボットが備えるべき聴覚機能について，マイクロホンアレイが持つ空間フィルタリング機能の中で最も基本的な，音源定位・分離問題を扱った．移動ロボットを用いた音源分離実験を通じて，マイクロホンアレイを用いることで空間特性が事変である混合音の分離や，車輪音などの自己発生音の抑圧がある程度対処可能であることを示した．ただし，残響の大きな環境における分離性能低下が確認されたため，残響抑圧を取り入れた音源分離 [Yoshioka *et al.*, 2011] など，マイクロホンアレイの空間フィルタリング機能をさらに活用することが今後の課題の 1 つである．また，本手法は HARK [Nakadai *et al.*, 2010] などの環境に対するチューニングが必要であるが，短いターンアラウンド時間で高速処理可能な手法と違い，音環境の違いには頑健ながら処理に時間を要する手法である．したがって，ロボットなどへの応用のためには，これら 2 つの手法を状況に応じて効果的に使い分ける枠組みなども必要となる．

今回の音源分離実験では，分離対象の音源はロボットの左右に分かれるという仮定のもとで分離音の復元を行った．ロボットがより一般的な軌道で移動する場合はこのような

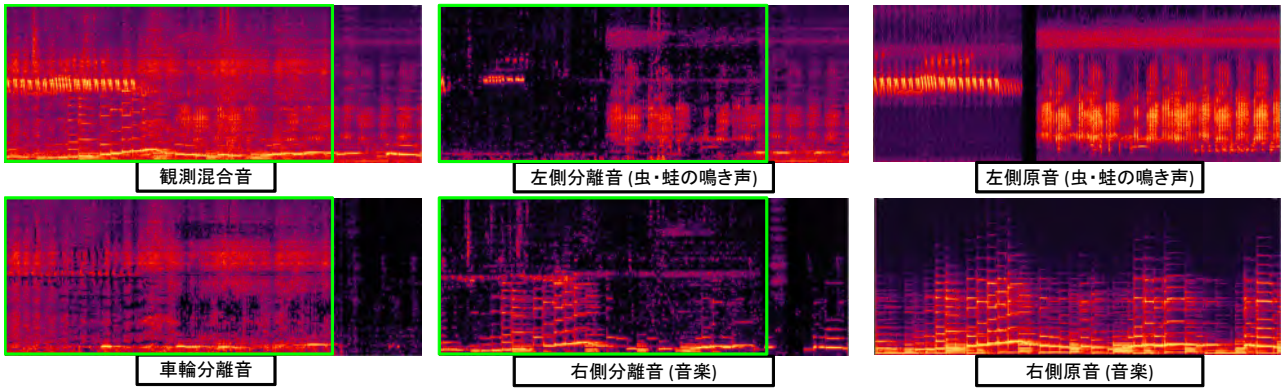

Figure 8: 屋外混合音分離結果: 残響時間 $RT_{60} = 150$ [ms]

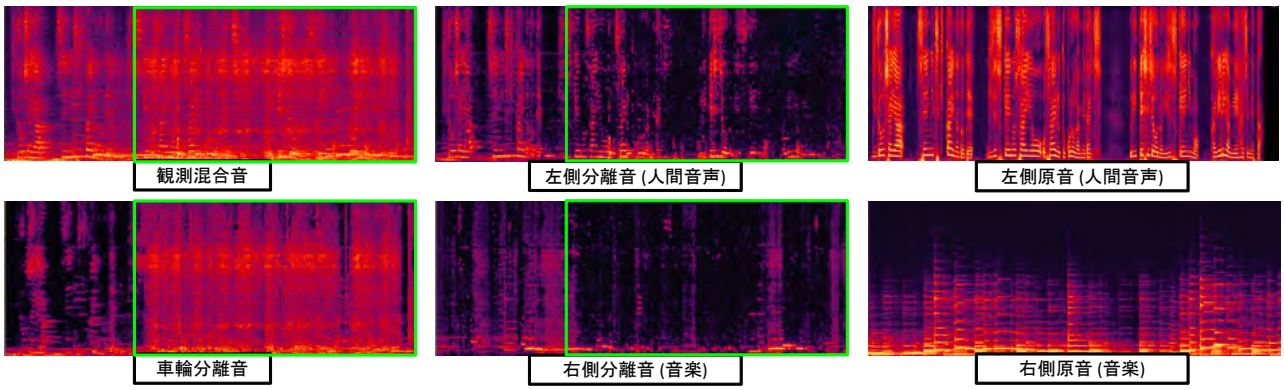

Figure 9: 屋内混合音分離結果: 残響時間 $RT_{60} = 800$ [ms]

方法は用いることが出来ないため，音源定位結果の時間連続性を考慮したトラッキング [Otsuka *et al.*, 2011] や，分離音の持つ音色などの特徴による同一音源の識別 [Sasaki *et al.*, 2009] を通じて，同一音源から発せられた分離音を集約する必要がある．状況をさらに一般化し，音源そのもののが移動する場合や，音が断続的に発せられ一時的に消失しうる場合では，これらの手法や視覚情報など異なるモダリティの統合など，マイクロホンアレイの枠組みを越えた手法が必要となる．

本稿でのロボット自己発生音の対処は，音源のマイクロホンアレイに対する相対位置は不変であることを仮定して行った．ヒューマノイドロボットなど，複雑な動作を行うロボットの自己発生音では音源位置の定常性の仮定が成り立たないこともありうる．解決策としては，自己発生音源の近くにマイクやセンサを設置し，マイクロホンアレイ処理に組み込んで抑圧する手法 [Sawada *et al.*, 2010] や，ロボットを動かすモータ指令値から自己発生音を予測し，抑圧する手法 [Ince *et al.*, 2011] などが挙げられる．

さらなる今後の展望としては，今回取り扱った音源定位・分離という汎用的ながら低次な問題から発展させ，分離結果を用いた複雑なタスクをこなすロボット (たとえば音を頼りにしたレスキューロボットや警備ロボット) などが考えられる．これらの高度なタスクを手がけるロボットを実現する要素技術の取捨選択や研究の加速には，データセットの整備も重要な今後の課題として挙げられる．

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

参考文献

[Benesty *et al.*,] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. Springer Topics in Signal Processing.

[Blei *et al.*, 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Conradsen *et al.*, 2003] K. Conradsen, A. A. Nielsen, J. Schou, and H. Skiriver. A Test Statistic in the Complex Wishart Distribution and Its Application to Change Detection in Polarimetric SAR Data. *IEEE Trans. on Geoscience and Remote Sensing*, 41(1):4–19, 2003.

[Duong *et al.*, 2010] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model. *IEEE Trans. on ASLP*, 18(7):1830–1840, 2010.

[Escobar and West, 1995] M. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[Ince *et al.*, 2011] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai. Assessment of General Applicability of Ego Noise Estimation. In *Proc. of International Conference on Robotics and Automation*, pages 3517–3522, 2011.

[Nakadai *et al.*, 2010] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and Implementation of Robot Audition System "HARK". *Advanced Robotics*, 24(5–6):739–761, 2010.

[Otsuka *et al.*, 2011] T. Otsuka, K. Nakadai, T. Ogata, and H. G. Okuno. Bayesian Extension of MUSIC for Sound Source Localization and Tracking. In *Proc. of International Conference on Spoken Language Processing*, pages 3109–3112, 2011.

[Otsuka *et al.*, 2012] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno. Bayesian Unification of Sound Source Localization and Separation with Permutation Resolution. In *Proc. of AAAI Conf. on Artificial Intelligence*, pages 2038–2045, 2012.

[Sasaki *et al.*, 2009] Y. Sasaki, M. Kaneyoshi, S. Kagami, H. Mizoguchi, and T. Enomoto. Daily Sound Recognition Using Pitch-Cluster-Maps for Mobile Robot Audition. In *Proc. of International Conference on Intelligent Robots and Systems*, pages 2724–2729, 2009.

[Sawada *et al.*, 2010]
H. Sawada, J. Even, H. Saruwatari, K. Shikano, and T. Takatani. Improvement of Speech Recognition Performance for Spoken-Oriented Robot Dialog System using End-fire Array. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 970–975, 2010.

[Se *et al.*, 2005] S. Se, D. G. Lowe, and J. J. Little. Vision-Based Global Localization and Mapping for Mobile Robots. *IEEE Trans. on Robotics*, 21(3):364–375, 2005.

[Teh *et al.*, 2006] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[Thrun *et al.*, 2004] S. Thrun, M. Montemerlo, D. Koller, B. Wegbreit, J. Nieto, and E. Nebot. FastSLAM: An Efficient Solution to the Simultaneous Localization and Mapping Problem with Unknown Data Association. *Journal of Machine Learning Research*, 2004.

[van den Bos, 1995] A. van den Bos. The Multivariate Complex Normal Distribution–A Generalization. *IEEE Trans. on Information Theory*, 41(2):537–539, 1995.

[Yoshioka *et al.*, 2011] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno. Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization. *IEEE Trans. on ASLP*, 19(1):69–84, 2011.