移動ロボットによる音環境理解に向けて

○大塚琢馬 [1] 石黒勝彦 [2] 澤田宏 [2] 奥乃博 [1]
[1] 京都大学大学院情報学研究科 [2] NTT コミュニケーション科学基礎研究所

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) [1] や，カメラと赤外線センサを組み合わせた自動車の自動運転技術 [2] 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，図 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．聴覚処理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．

実環境中に存在する音源の性質から，ロボットがこれらの音を扱うために必要な要素技術を考える．実環境中の音源には，空調設備のように継続的に音を発生するものや，割れるガラスやドアの開閉音など散発的に発生するものがある．ロボットは時には同時に発生するこれらの音を聞き分け，個々の音を取り扱う必要がある．図 1 下部に，実環境中における音環境理解に必要な要素技術を 3 つのカテゴリに分けて示す．

Detection 散発的音源の時間上の検出や，音源の到来方向推定 (音源定位) [3] による空間上の検出など．

Decomposition 観測した混合音を個々の音に分解する音源分離 [4]，直接音と反射音を分ける残響除去 [5].

Decoding 観測音から抽出した個々の音から情報を抽出する処理であり，ロボットのタスクに依存する．例としては，人間の発話に対する音声認識や，音源種類の同定などが挙げられる．

本稿では，上記の 3 カテゴリのうち，音環境理解ロボット全般に共通して重要との立場から，1. detection と 2. decomposition に焦点を当てる．これらのカテゴリに属する機能は主にマイクロホンアレイを用いる手法が有力である [3–6]．以下では特に，環境中に存在する複数の静止音源を移動ロボットによって観測して個々の音源を分離するタスクを扱い，マイクロホンアレイによって対処可能な範囲を示す．また，時々刻々ロボットが移動しながら音を観測した時に，同一の音源から発せられた音をまとめるための，各時刻の推定音源方向のトラッキングや，音源同定の必要性を議論する．

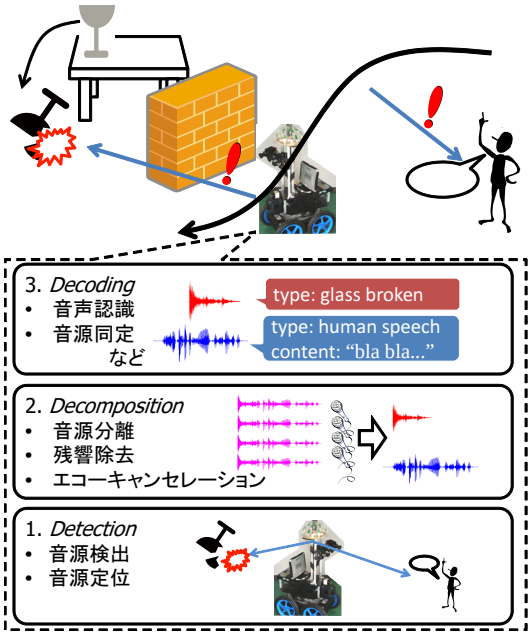

図 1 移動ロボットによる音環境理解と必要な要素技術

2. 移動ロボットによる音情報抽出の課題

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，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．

これらの課題に加えて，観測する音環境に課す仮定をどれほど緩和出来るかも考慮する必要がある．例えば，許容する残響時間，音源数は既知か未知か，スペクトル形状 (音色) が定常な音源か否かなどが挙げられる．これらの課題や手法の仮定を踏まえて，音源定位・分離法を示す．

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

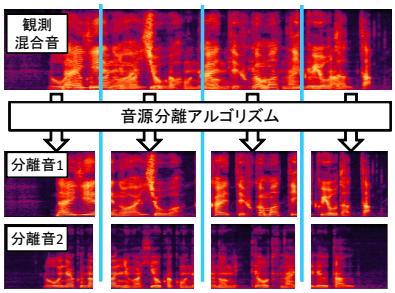

図 2 時分割処理: 各区間内の空間特性は定常と仮定

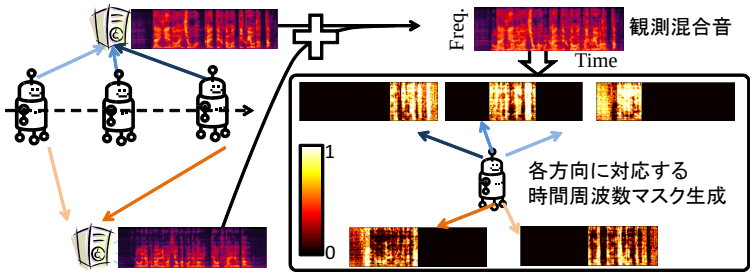

図 3 方向ごとの時間周波数マスク推定: 音源が通過した方向に対応するマスクを生成

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 [7] は，図 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．

佐々木らの移動ロボットによる音源定位・分離も，図 2 の時分割処理である [8]．本手法は，32 チャネルマイクロホンアレイによる delay-sum beamformer を用いて音源定位や分離を行う．本手法も，精度の良い定位を実現するためには環境依存のパラメータ設定が必要な他，比較的大規模なチャネル数のマイクロホンアレイを用いるため，同期サンプリングを行う装置などの可用性などの課題を残す．

これに対して Otsuka らは，環境依存のパラメータ設定せずとも安定して定位や分離を行う手法を開発している [4]．本手法は入力された観測音に含まれるそれぞれの音源方向に対して時間周波数マスクを推定することで音源の分離と定位を行う．より具体的には，音源分離を観測音の時間周波数領域上のクラスタリング問題として扱い，クラスタリング結果を時間周波数マスクとして生成する．また，音源定位は，生成された時間周波数マスクと事前に与えられたステアリングベクトルとの対応付けとして，音源分離と定位を同時に最適化する．本手法は，観測音に含まれる情報を効果的に利用するベイズ推定によるクラスタリング問題として定式化されており，様々な音源数，残響時間に対しても共通のパラメータを用いて安定した性能を示すことが可能である．

Otsuka らの手法も定常な空間特性を仮定した手法であるが，図 3 のように時間変化する空間特性に対応する．図 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，図 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が自動的に生成される．従って，図 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. 実験設定と手法概略

本節では，分離実験に用いた混合音の収録条件と用いる音源定位・分離法 [4] の概略を述べる．図 4 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ロホンアレイを示す．図 4 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，図 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，図 4 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った．

3·1 ベイズクラスタリングによる音源定位・分離

本手法は，観測した音響信号を短時間フーリエ変換を通じて時間周波数領域に変換した多チャンネルスペ

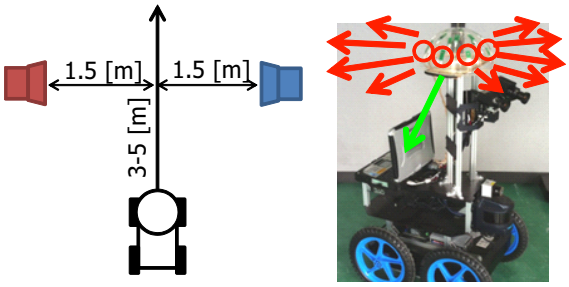

図 4 音源配置とロボットの軌道 (左) とロボットに搭載されたマイクロホンアレイ (右)

クトログラム $\mathbf{x}_{tf}$ に対して，音源分離と音源定位に対応する 2 種類のクラスタリングを行う．ここで，$\mathbf{x}_{tf}$ は時間フレーム t，周波数ビン f の M 次元複素ベクトルであり，M は観測に用いたマイク数である．音源分離のための時間周波数マスクは，各 $\mathbf{x}_{tf}$ がどの音源に属するかを示す潜在変数 z_{tf} の割り当て確率の推定を通じて行う．音源定位は，音源を示す添字を k とすると，$z_{tf}=k$ として音源 k に属する観測音がどの方向から到来しているかを示す潜在変数 w_k の割り当て確率の推定によって行う．つまり，各音源 k は離散的に定義された D 個の方向の中から $w_k=d$ のように選択される．

これらの潜在変数に関する確率計算は，

$$p(\mathbf{Z},\mathbf{W},\Theta|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Z},\mathbf{W},\Theta)p(\mathbf{Z}|\Theta)p(\mathbf{W}|\Theta)p(\Theta)$$

のように，観測信号 $\mathbf{X}$ が与えられたときの潜在変数の事後確率のベイズ推定として扱う．ただし，$\mathbf{X}$，$\mathbf{Z}$，$\mathbf{W}$ は，変数 $\mathbf{x}_{tf}$，z_{tf}，w_k を集合的に表し，Θ は上記説明からは省かれた，モデルに含まれるその他の潜在変数を表す．モデルの詳細や，事後確率の具体的な推定方法は文献 [4] に詳しい．

今回の実験では，両環境の観測混合音に対して環境に依存したパラメータの手動設定を行うことなく分離処理を行った．ベイズ推定を行う利点の 1 つは，音源マスク $\mathbf{Z}$ に事前分布 $p(\mathbf{Z}|\Theta)$ がモデルが過度に複雑になりすぎないように与えられているため，観測音に含まれる混合音の複雑さに従って適切に推定結果が導かれることが挙げられる．このことが本手法の，音環境中の音源数などに対する頑健さの由来の 1 つである．

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 (図 4 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 (図 4 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．

4. 実験結果

図 5，6 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．音源分離結果は図 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図 5，6 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．

図 5 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ことが挙げられる．一方，図 6 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．

5. 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 [5] 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 [7] 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 [3] や，分離音の持つ音色などの特徴による同一音源の識別 [9] 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，音源そのものが移動する場合や，音が断続的に発せられ一時的に消

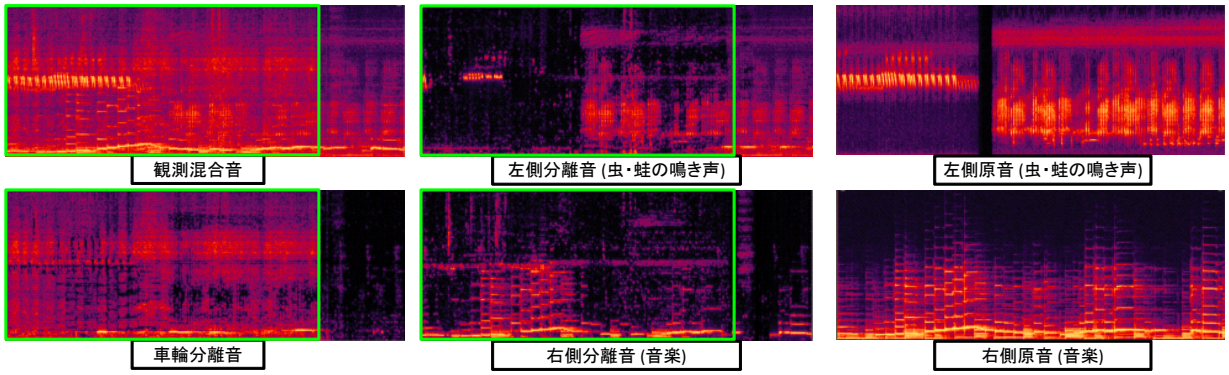

図 5 屋外混合音分離結果: 残響時間 $RT_{60} = 150$ [ms]

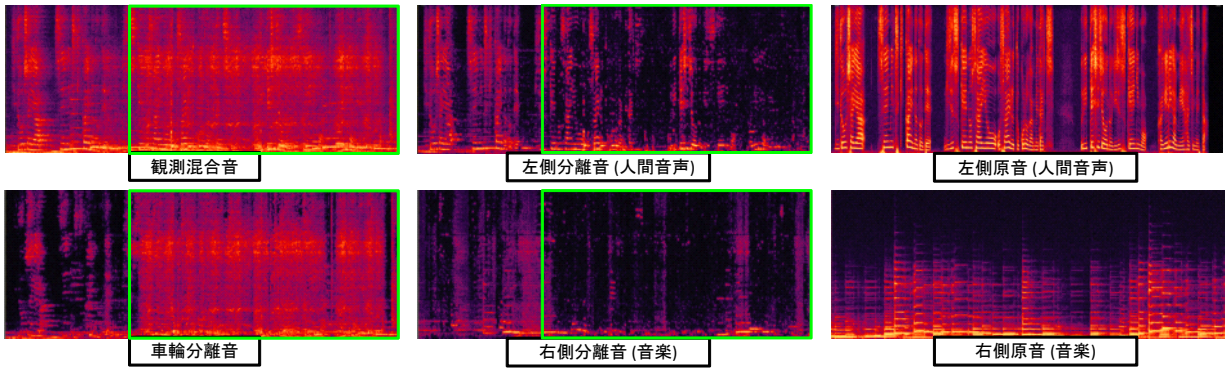

図 6 屋内混合音分離結果: 残響時間 $RT_{60} = 800$ [ms]

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 [10] 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 [11] 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数えることができる．

参考文献

[1] 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.

[2] S. Thrun. Toward Robotic Cars. *Communications of the ACM*, 53(4):99–106, 2010.

[3] T. Otsuka, K. Nakadai, T. Ogata, and H. G. Okuno. Bayesian Extension of MUSIC for Sound Source Localization and Tracking. In *Proc. of INTERSPEECH*, pages 3109–3112, 2011.

[4] 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*, 2012. to appear.

[5] 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.

[6] R. Takeda, K. Nakadai, T. Takahashi, T. Ogata, and H. G. Okuno. Efficient Blind Dereverberation and Echo Cancellation based on Independent Component Analysis for Actual Acoustic Signals. *Neural Computation*, 24(1):234–272, 2011.

[7] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, H. Yuji, and H. Tsujino. Design and Implementation of Robot Audition System “HARK”. *Advanced Robotics*, 24(5–6):739–761, 2010.

[8] 佐々木 洋子, 加賀美 聡, and 溝口 博. マイクアレイのメインローブモデルを用いた点音源検出手法. **ロボット学会論文誌**, 27(3):325–333, 2009.

[9] 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 IROS*, pages 2724–2729, 2009.

[10] 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 IROS*, pages 970–975, 2010.

[11] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai. Assessment of General Applicability of Ego Noise Estimation. In *Proc. of ICRA*, pages 3517–3522, 2011.