人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-058-08

雑踏環境における音源地図の生成

坂東 宜昭[1*]　升山 義紀[2,1]　佐々木 洋子[1]　大西 正輝[1]

[1] 産業技術総合研究所　[2] 東京都立大学

概要: 本稿では，非負値行列因子分解 (NMF) に基づく，雑踏音環境のマッピングについて述べる．移動ロボットを用いての音源の空間的な配置を推定する音環境マッピングは，知的システムが周辺環境を認識し適切な行動とるために不可欠である．従来の枠組みは，初段の音源定位に強く依存した構成となっており，残響や拡散性雑音の強い雑踏環境下では性能を発揮できなかった．そこで本研究では，音源定位の代わりに NMF を用いた音源分離に基づく音環境マッピングを提案する．本手法は，まずスペクトル特徴に基づき観測信号を分解したあと，個別の音源に対して位置推定するので，雑踏環境下でも安定してマッピングできる．数値シミュレーションにより提案法の有効性を評価した．

1　はじめに

自律移動ロボットが，環境中の音源の位置や分布を推定する音環境マッピングは，ロボットが周囲の音環境を空間的に認識し適切に応答するために不可欠な基盤技術である [1–6]．特に，展示会場や市街地など雑踏環境でも頑健に動作する枠組みを確立できれば，周囲の音響情景に合わせた適応的な接客や，都市規模の環境モニタリングなど様々な応用技術を実現できる．

従来の音環境マッピングの多くは，まずマイクロホンアレイを用いた音源定位 [7–10] により音源到来方向を推定し，異なる観測点での推定結果から三角測量することで，音源地図を生成していた [1–4]．この枠組では，時間フレームごとに動作する音源定位をオンライン処理として実装できるため，実時間で地図を生成できる．一方，最終出力である地図の推定性能が，初段である音源定位に強く依存しており，残響や拡散性雑音が強い雑踏環境では大きく性能が劣化する課題があった．

本研究では，収録した録音信号全体を直接モデル化する音環境マッピングを提案する．本手法の初段では，フレーム独立に処理していた音源定位の代わりに，録音信号全体を単一の生成モデルで音源分離する非負値行列因子分解 (NMF) [11, 12] を適用する．NMF で分解された個別の音源信号に対し，音源定位と三角測量によりその位置を推定する．初段で NMF を用いて観測信号を予め分解することで，アレイ信号処理では性能劣化しやすい残響や拡散性雑音が含まれる観測信号でも頑健な動作を実現する．また NMF は，観測信号全体を一挙に分解する枠組みであるため，同じ音源を複数回観測した場合など，長い録音信号全体の共起関係をモデル化でき，安定して複数の音源を弁別できる．数値混合音を用いた評価実験でその有効性を評価した．

*連絡先：国立研究開発法人 産業技術総合研究所
〒 135-0064 東京都江東区青海 2-4-7 産総研 臨海副都心センター 8F
E-mail: y.bando@aist.go.jp

2　NMF に基づく音環境マッピング

提案法では，移動ロボットを用いて収録した観測混合音 $\mathbf{y}_{\tau,ft} \in \mathbb{C}^M$ およびロボットの自己位置 $\mathbf{x}_\tau \in \mathbb{R}^2$ から，音源 $n = 1, \ldots, N$ の座標 $\mathbf{s}_n \in \mathbb{R}^2$ を推定する．ここで，$\tau = 1, \ldots, \mathcal{T}$, $f = 1, \ldots, F$ および $t = 1, \ldots, T$ は，それぞれミニバッチ，周波数ビンおよび時間フレームインデックスを表す．本手法では，観測信号を T フレームのミニバッチに分割し，バッチ τ 内ではロボットの移動を無視できると仮定する．以降で述べる通り，提案法は，1) NMF による観測信号分解，2) 基底クラスタリングによる音源パワースペクトル密度の推定，3) 三角測量による音源位置推定の 3 つの処理で構成される．

2.1　ガンマ過程非負値行列因子分解

提案法ではまず，スペクトル情報に基づき観測信号を分解するため，NMF を適用する．具体的には，多チャネル観測信号 $\mathbf{y}_{\tau,ft}$ の平均パワースペクトログラム $\bar{y}_{\tau,ft} \triangleq \frac{1}{M}\|\mathbf{y}_{\tau,ft}\|_2^2 \in \mathbb{R}_+$ を，K 本のスペクトル基底 $\mathbf{w}_k = [w_{k1}, \ldots, w_{kF} \in \mathbb{R}_+^F$ とそれらの時間アクティベーション $\mathbf{h}_{\tau,k} = [h_{\tau,k1}, \ldots, h_{\tau,kT}] \in \mathbb{R}_+^T$ の積へ分解する．このとき，基底数 K は，観測信号の複雑さに合わせて最適な値が変動するため，事前に決定することが難しい．そこで本手法では，基底数を事前に定めないガンマ過程 NMF [12] を用いて観測 $\bar{y}_{\tau,ft}$ を表現する．

$$\bar{y}_{\tau,ft} \sim \mathrm{Exp}\left(\sum_{k=1}^{K} g_{\tau,k} w_{kf} h_{\tau,kt}\right) \tag{1}$$

ここで，$g_{\tau,k} \in \mathbb{R}_+$ は，各基底の観測に対する寄与度合いを表すゲイン変数である．このゲイン変数にスパースな事前分布を仮定することで，十分大きい K を設定すれば，観測の複雑さに合わせて必要な本数の基底の

みを用いた分解が行われる．ゲイン変数 $g_{\tau,k}$ には，共役事前分布である以下のガンマ分布を仮定する．

$$g_{\tau,k} \sim \mathcal{G}\left(\frac{a_0}{K}, a_0\right) \tag{2}$$

ここで，$a_0 \in \mathbb{R}_+$ は $g_{\tau,k}$ のスパース度合いを制御するハイパーパラメータである．本モデルは，$K \to \infty$ で $g_{\tau,k}$ がガンマ過程に従う近似モデルとなっている．また，基底 w_{kf} とアクティベーション $h_{\tau,kt}$ には，それぞれ期待値が 1 となる以下のガンマ事前分布を仮定する．

$$w_{kf} \sim \mathcal{G}\left(a_1, a_1\right), \quad h_{\tau,kt} \sim \mathcal{G}\left(a_2, a_2\right) \tag{3}$$

ただし $a_1 \in \mathbb{R}_+$ と $a_2 \in \mathbb{R}_+$ は，それぞれ w_{kf} と $h_{\tau,kt}$ のスパース度合いを制御するハイパーパラメータである．

NMF の推論では，真の事後分布 $p(\mathbf{G}, \mathbf{W}, \mathbf{H} \mid \bar{\mathbf{Y}})$ を近似する変分事後分布 $q(\mathbf{G}, \mathbf{W}, \mathbf{H}) \triangleq q(\mathbf{G})q(\mathbf{W})q(\mathbf{H})$ を求める．この推論は，文献 [12] と同様に導出した以下の更新則を反復することで行われる．

$$q(g_{\tau,k}) \leftarrow \mathcal{GIG}\Bigg(\frac{a_0}{K}, a_0 + \sum_{f,t} \frac{\langle w_{kf} h_{\tau,kt} \rangle}{\omega_{\tau,tf}}, \sum_{f,t} \bar{y}_{\tau,ft} \phi^2_{\tau,tfk} \left\langle w^{-1}_{kf} h^{-1}_{\tau,kt} \right\rangle \Bigg) \tag{4}$$

$$q(w_{kf}) \leftarrow \mathcal{GIG}\Bigg(a_1, a_1 + \sum_{\tau,t} \frac{\langle g_{\tau,k} h_{\tau,kt} \rangle}{\omega_{\tau,tf}}, \sum_{\tau,t} \bar{y}_{\tau,ft} \phi^2_{\tau,tfk} \left\langle g^{-1}_{\tau,k} h^{-1}_{\tau,kt} \right\rangle \Bigg) \tag{5}$$

$$q(h_{\tau,kt}) \leftarrow \mathcal{GIG}\Bigg(a_2, a_2 + \sum_{f} \frac{\langle g_{\tau,k} w_{kf} \rangle}{\omega_{\tau,tf}}, \sum_{f} \bar{y}_{\tau,ft} \phi^2_{\tau,tfk} \left\langle g^{-1}_{\tau,k} w^{-1}_{kf} \right\rangle \Bigg) \tag{6}$$

ここで，$\mathcal{GIG}$ は一般化ガウス分布 [12] を表し，$\omega_{\tau,tf} \in \mathbb{R}_+$ と $\phi_{\tau,tfk} \in \mathbb{R}_+$ ($\sum_k \phi_{\tau,tfk} = 1$) は以下の値をとる補助変数である．

$$\omega_{\tau,tf} = \langle g_{\tau,k} w_{kf} h_{\tau,kt} \rangle \tag{7}$$

$$\phi_{\tau,tfk} \propto \left\langle \frac{1}{g_{\tau,k} w_{kf} h_{\tau,kt}} \right\rangle^{-1} \tag{8}$$

この更新則は，対数周辺尤度 $\log p(\bar{\mathbf{Y}})$ の変分下限を最大化するように導出されている．変分下限の最大化は，変分事後分布と真の事後分布との Kullback–Leibler ダイバージェンス $\mathcal{D}_{\mathrm{KL}}[q(\mathbf{G}, \mathbf{W}, \mathbf{H}) \mid p(\mathbf{G}, \mathbf{W}, \mathbf{H} \mid \bar{\mathbf{Y}})]$ の最小化に対応する [13]．

2.2 基底のクラスタリング

初段の NMF で得た K 個の基底は，次段のベイズ混合ガウスモデル (GMM) により，N 個の音源にクラスタリングされる．NMF は，観測信号を K 個の基底とアクティベーションの組に分解するが，これらは N 個の音源と 1 対 1 対応しない．本稿では，NMF の各基底が複数の音源を表現せず，1 つの音源のみと対応すると仮定し，基底を N 個のクラスタに分割する．具体的には，バッチごとの基底ゲインの期待値 $\langle g_{\tau,k} \rangle$ を特徴量としてベイズ GMM によりクラスタリングする．ベイズ GMM は，そのディリクレ事前分布の効果により，十分な音源数 N を準備しておけば，不要な音源クラスを自動的に縮退させることができる．

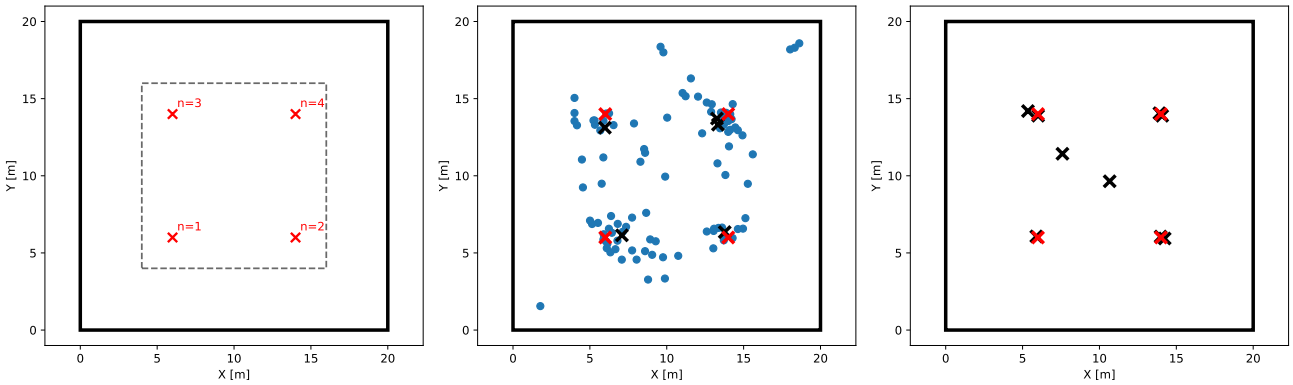
2.3 音源位置の推定

再終段では，ガンマ過程 NMF とベイズ GMM により得た N 個の音源パワースペクトル密度を用いて，多チャネル観測混合音 $\mathbf{y}_{\tau,ft}$ から音源位置 s_n を推定する．具体的にはまず，観測信号を Wiener フィルタリングすることで，音源像 $\hat{\mathbf{y}}_{\tau,nft} \in \mathbb{C}^M$ を得る．次に，遅延和ビームフォーマを用いて音源像の空間スペクトルを計算し，バッチ τ での音源 n の到来方向 $d_{\tau,n}$ を推定する．得られた音源到来方向 $d_{\tau,n}$ と自己位置 $\mathbf{x}_\tau$ を用いて三角測量し，音源位置 $\mathbf{s}_n$ を得る．

2.4 視覚情報に基づく枠組みとの関連

活発に研究されている隣接技術の 1 つである，単眼カメラによる環境マッピングについて概観しながら，音環境マッピングとの関連および提案法の位置づけについて議論する．視覚情報を用いた代表的なマッピング技術として，既知の自己位置から地図を推定する多視点ステレオ [14] と，自己位置と地図を同時推定する visual SLAM (simultaneous localization and mapping) [15, 16] や SfM (structure from motion) [17] がある．従来の音源定位に基づくマッピングは，音源定位により得た音源到来方向から地図上の音源位置を推定する点で，多視点ステレオに近い枠組みである．一方，本稿で扱う音環境マッピングでは，自己位置推定しないものの，音源位置だけでなく混合音から音源信号を推定する点で，マッピングのみを解く多視点ステレオより visual SLAM の方が類似点が多い．Visual SLAM は，画像上の疎な特徴点に対して幾何的な誤差を最小化する特徴点法 [15] と，画像間の輝度値の誤差を直接最小化する直接法 [16] に大別できる．提案法の観測モデルである式 (1) は，多チャネル録音信号 $\mathbf{y}_{\tau,ft}$ に対するゼロ平均多変量複素ガウス尤度に対応している[1]．観測信号に対する誤差 (尤度) を直接モデル化している点で，提案法は直接法に近い．一方提案法は，NMF，基底クラスタリングおよび三角測量の 3 つの処理をカスケード接続

[1] 完全な一致には，対数尤度に $1/M$ を乗じる必要がある．



(a) 音源位置 (赤) とロボット軌跡 (灰)　(b) 音源定位に基づくマッピング　(c) NMF に基づくマッピング (提案法)

図 1: (a) に実験設定を，(b) と (c) にマッピング結果を示す．灰色線はロボットの軌跡を，赤 × は正解の音源位置を表す．黒 × は推定された音源位置で，青点は HARK により検出された各音源の三角測量結果である．

して構築されており，visual SLAM や SfM のような全体最適化にはなっておらず，後段の処理へ誤差が蓄積されていくと考えられる．音源地図と音源分離の同時最適化は今後の課題とする．

3　評価実験

本節では，移動ロボットで観測した録音信号を模した，数値混合音を用いた評価実験について報告する．

3.1　実験設定

屋内を移動ロボットが巡回して収集した 10 分間の $M = 4$ チャネル混合音を生成して評価に用いた．この実験では，残響時間 (RT_{60}) 600 ms を持ち高さ 4 m で 20 m 四方の室内を想定し，4 個の音源を室内に配置した．各音源は，FSDKaggle2018 データセット [18] から選んだそれぞれ 20 秒以上の長さを持つクリップを繰り返し発している．これらのクリップには，犬の鳴き声やチャイム音など，非定常な信号を選んだ．ロボットは，図 1 のように，この室内を 2 分間で 1 週する速度で移動した．混合音は鏡像法 [19] を用いて生成した．また，観測信号には拡散性雑音として，WSJ0 [20] から選んだ英語音声によるバブルノイズを信号対雑音比 15 dB で重畳している．混合音は，サンプリング周波数 16 kHz として生成した．

音源定位に基づく音環境マッピングを実装し，提案法と比較した．この手法は，ロボット聴覚オープンソースソフトウェア HARK [21] が提供している MUSIC (multiple signal classification) 法 [9, 10] による音源定位と音源トラッキング法から音源到来方向軌跡を推定し，三角測量により音源位置を推定する．HARK による音源定位では，音源信号の無音区間やロボットが音源から定期的に遠ざかることを要因として，別の時刻に観測された同じ音源を別の音源として検出してしまう．そこで，得られた複数の音源位置をベイズ GMM でクラスタリングして，最終的な音源位置の出力とした．

表 1: 正解音源位置から最近傍の推定位置の平均誤差．

手法	誤差 [m]
音源定位に基づくマッピング	0.797
NMF に基づくマッピング (提案法)	0.104

3.2　実験結果

図 1-(b) と-(c) に推定結果を，表 1 に位置推定誤差を示す．音源定位 (MUSIC) に基づくマッピング (図 1-(b)) では，103 個の音源とその位置 (青点) が検出され，これらをクラスタリングして 5 点の音源位置が推定された．この枠組では，音源信号の無音区間により音源追跡に失敗し，短い音源として検出されているため，三角測量の精度が低く，真値から大きく離れた位置を持つ音源が多く検出されている．これらをクラスタリングした結果，位置推定誤差は 0.797 m に留まっている．提案法による結果 (図 1-(c)) では，9 個の音源位置が推定され，2 個を除き，4 つの目的音源から 1 m 以内に定位されている．位置推定誤差は 0.104 m であった．部屋の中心付近に，音源位置から大きく外れた音源が 2 つ検出されているが，これらは拡散性雑音や目的音源の残渣成分と考えられる．提案法は，録音信号全体を一挙に直接分解するため，より安定して音源位置を推定できている．一方で提案法においても，音源 $n = 3$ の正解位置に複数個の推定結果 (黒 ×) が重なっ

ているように，単一の音源を複数の音源として検出してしまっている．これは，同じ音源でも大きく異なるアクティベーションパターンを持つ基底を，単一の音源としてクラスタリングできていないためと考えられる．本問題は，NMF による音源分離と後段の音源位置推定を一挙に最適化する拡張により，互いの情報を相補的に利用することで解決できると考えられる．

4 おわりに

本稿では，NMF による信号分解を基盤とした音環境マッピングについて述べた．従来法の多くは，混合音から各音源の到来方向を推定する音源定位に強く依存した構成で，性能劣化の原因になっていた．そこで本研究では，初段で NMF による音源分離を適用することで，拡散性雑音や残響のある環境でも安定した動作を実現する．数値混合音を用いた評価実験では，HARK により構築した音源定位に基づくマッピングより高い精度で音源位置を推定できていることを確認した．今後は，音源分離と音源位置推定の同時最適化による音源クラスタリングの性能向上や，オンライン推論によるリアルタイムマッピングへの拡張などを進める．

謝辞 本研究の一部は，JST ACT-X 数理・情報のフロンティア JPMJAX200N の支援を受けた．

参考文献

[1] Y. Sasaki, S. Kagami, and H. Mizoguchi, “Multiple sound source mapping for a mobile robot by self-motion triangulation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 380–385.

[2] Y. Sasaki, R. Tanabe, and H. Takernura, “Online spatial sound perception using microphone array on mobile robot,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2478–2484.

[3] J. Even, Y. Morales, N. Kallakuri, J. Furrer, C. T. Ishi, and N. Hagita, “Mapping sound emitting structures in 3D,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 677–682.

[4] S. Michaud, S. Faucher, F. Grondin, J.-S. Lauzon, M. Labbé, D. Létourneau, F. Ferland, and F. Michaud, “3D localization of a sound source using mobile microphone arrays referenced by SLAM,” *arXiv preprint arXiv:2007.11079*, 2020.

[5] C. Schymura and D. Kolossa, “Potential-field-based active exploration for acoustic simultaneous localization and mapping,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 76–80.

[6] D. Su, T. Vidal-Calleja, and J. V. Miro, “Towards real-time 3D sound sources mapping with linear microphone arrays,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1662–1668.

[7] F. Grondin and J. Glass, “SVD-PHAT: A fast sound source localization method,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4140–4144.

[8] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone arrays*. Springer, 2001, pp. 157–180.

[9] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, “Intelligent sound source localization for dynamic environments,” in *IEEE/RSJ international conference on Intelligent Robots and Systems (IROS)*, 2009, pp. 664–669.

[10] R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[11] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[12] M. D. Hoffman, D. M. Blei, and P. R. Cook, “Bayesian nonparametric matrix factorization for recorded music,” in *International Conference on Machine Learning (ICML)*, 2010, pp. 1–8.

[13] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[14] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, “A comparison and evaluation of multi-view stereo reconstruction algorithms,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 519–528.

[15] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM,” *IEEE Transactions on Robotics*, pp. 1–17, 2021.

[16] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[17] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[18] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, “General-purpose tagging of Freesound audio with audioset labels: Task description, dataset, and baseline,” in *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2018, pp. 69–73.

[19] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[20] J. Garofalo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) complete,” *Linguistic Data Consortium, Philadelphia*, 2007.

[21] K. Nakadai, H. G. Okuno, and T. Mizumoto, “Development, deployment and applications of robot audition open source software HARK,” *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 16–25, 2017.