複数のカルマンフィルタを用いた複数移動話者追跡と精度評価

○村瀬昌満† 山本俊一† Jean-Marc Valin‡ 中臺一博††
山田健太郎†† 駒谷和範† 尾形哲也† 奥乃博†
†京都大学大学院情報学研究科 ‡Université de Sherbrooke
††(株) ホンダ・リサーチ・インスティチュート・ジャパン

Multiple Moving Speakers Tracking based on Multiple Kalman Filters and Accuracy Evaluation

*Masamitsu Murase†, Shunichi Yamamoto†, Jean-Marc Valin‡, Kazuhiro Nakadai††,
Kentaro Yamada††, Kazunori Komatani†, Tetsuya Ogata†, Hiroshi G. Okuno†
†Graduate School of Informatics, Kyoto University ‡Université de Sherbrooke
††Honda Research Institute Japan Co., Ltd.

Abstract— This paper presents multiple moving speaker tracking using an 8ch microphone array system installed on a mobile robot. We use a set of Kalman filters with different history lengths in order to track non-linear movements of multiple speakers. For quantitative evaluation of the tracking, motion references of sound sources and a mobile robot, called SIG2, were measured accurately by ultrasonic 3D tag sensors. As a result, we showed that the system tracked three simultaneous sound sources even when SIG2 moved.

Key Words: moving speaker, Kalman filter, robot audition, microphone array

1. はじめに

実世界におけるロボットとのインタラクションにおいて，話者やロボットが静止しているとは限らない．さらに，多くの音源分離手法では音源の位置情報を必要としている．そのため，複数移動話者の追跡はヒューマノイドロボットの頑健なインタラクションを実現する上で重要な機能である．

我々はミッシングフィーチャー理論を利用した同時発話認識システムを開発した [1]．これは静止話者に対しては高い認識率を示すが，移動話者に対する場合，追跡の失敗により精度が低下する．これは，静止話者に対する音源定位手法をそのまま移動話者に適用したためであると言える．移動話者の追跡を行う場合，各時刻における定位結果を各話者ごとに接続する必要がある．そのため，静止話者の場合と異なり，音源が交差したのか接近後離れたのかといった曖昧性が生じることになる．

中臺ら [2] は，視聴覚統合による実時間複数話者追跡システムを開発した．これは，二本のマイクロフォンのみによる低い定位精度を，視覚を用いることにより補うというものである．しかし，水平方向の定位しかできないため，本稿で想定する音源分離や認識 [1] のパラメータとして利用することはできない．

麻生ら [3] は，パーティクルフィルタによる話者追跡システムを開発した．これは，パーティクルフィルタを用いて視聴覚を統合し，8 本のマイクロフォンを用いることによって，高精度の追跡結果が得られている．しかし，移動話者追跡の問題である，音源の交差，接近の曖昧性に関しては言及していない．

我々は，カルマンフィルタ (KF) [6] による音声の音響的特徴を用いた複数話者追跡手法をヒューマノイドロボット SIG2(図 1) に実装し，複数移動話者を正確に追跡できることを確認した [4]．本稿では，ロボットが移動する場合であっても複数話者を正確に追跡できることを確認した．また，超音波 3 次元タグを用いることでロボットと話者の正確な位置を測定し，追跡精度の定量的な評価を行った．

2. 複数移動話者追跡システムの構成

我々が開発したシステムでは以下の流れに従って処理を行う．

1. マイクロフォンアレイを用いた領域分割による音源定位により，各時刻における音源定位を行う．
2. 定位結果をもとに，複数の KF により一つ前の時刻までの軌跡から現在の話者の位置情報を推定する．
3. KF による推定結果をもとに，同一話者と推定される定位結果に同じラベルを付与する．

以下で，各処理について詳しく述べる．

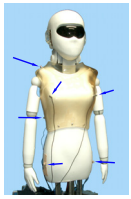


Fig. 1: SIG2& 8-channel microphones

2·1 各時刻における領域分割による音源定位

各時刻における音源定位手法として，steered beamformer [5] を用いた．これは，8 本のマイクロフォンを用いた音源定位であり，全方向のビームフォーマを計算しその出力が最大となる方向を探すというものである．具体的には，マイクロフォンアレイの周囲の空間を 5,120 個の三角形に分割し，その 2,562 個の各頂点

について遅延和を計算し，そのパワーを計算する．この手法によって，静止音源や一話者に対する高精度な音源定位が可能となっている．

しかし，この手法によって得られるのは各時刻における音源の定位結果のみであり，得られた時系列データに対して同一音源をグループ化するといったことはしない．そのため，この手法をそのまま複数移動話者の追跡に用いた場合，それぞれの話者ごとに追跡するのは困難であると言える．特に，話者が交差したのか接近後離れたのかを判断するのは難しいと言える．

2·2 カルマンフィルタを用いた話者追跡

KF では，線形に遷移するシステムにノイズが畳み込まれる場合を仮定している．これは以下の様に表される．

$$x_{k+1} = Fx_k + Gw_k \tag{1}$$

$$y_k = Hx_k + v_k \tag{2}$$

ここで，x_k は時刻 k におけるシステムの内部状態を表すベクトルであり，y_k は時刻 k における観測値を表すベクトルである．また，F は内部状態の遷移を表す行列であり，H は内部状態を観測値に写像する行列である．w_k, v_k はそれぞれプロセスノイズと観測ノイズである．本実験では，分散比 σ_w/σ_v として実験的に 0.01 を用いた．

本論文では，以下のようにモデル化を行った．時刻 k における特徴ベクトル p_k を次のように定める．

$$p_k = (\theta_k, \phi_k) \tag{3}$$

ここで，θ_k は時刻 k における話者のアジマス角，ϕ_k は時刻 k における話者のエレベーション角である．これらを用いて，時刻 k における内部状態 x_k を次のように定義する．

$$x_k = (p_k, p_{k-1}, p_{k-2}, \ldots, p_{k-l})^T \tag{4}$$

これは，過去 l フレームにおける特徴ベクトルの履歴である．

また，p_{k+1} は p_k, p_{k-l} から次のように近似できると仮定する．

$$p_{k+1} = p_k + \frac{p_k - p_{k-l}}{l} \tag{5}$$

よって，行列 F, G, H は次のように定義される．

$$F = \begin{pmatrix} (1+1/l)I & 0 & \cdots & 0 & (-1/l)I \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{pmatrix} \tag{6}$$

$$G = \begin{pmatrix} I & 0 & \cdots & 0 \end{pmatrix}^T \tag{7}$$

$$H = \begin{pmatrix} I & 0 & \cdots & 0 \end{pmatrix} \tag{8}$$

ここで，I は 2×2 の単位行列である．

2·3 履歴長の異なる複数の KF による予測

KF では線形状態遷移を仮定している．しかし，この仮定は実世界では成り立たないことが多く，精度の低下が生じる．具体的には，話者の非線形な動き，発話の割り込み，話者数の変化などによる影響が考えられる．我々は，異なる履歴長を持つ複数の KF による予測値を用いることにより，より正確な追跡を実現する．

話者の移動速度が一定である場合，線形な状態遷移であると見なせる期間が長いため，履歴長の長い KF の方が，より正確に話者の位置を予測することが可能となる．一方，話者の移動速度の変化が激しい場合，線形な状態遷移であると見なせる期間が短くなるため，履歴長の短い KF の方が，より正確に話者の位置を予測することができる．

我々は，複数の異なる履歴長を持つ KF を用い，適切な履歴長を持つ KF による予測値を選択して用いることによって，様々な話者の移動の移動の追跡を可能とする．具体的には，次のように行う．複数の異なる履歴長を持つ KF で次の話者の状態を並行に予測する．そして，前フレームでの予測誤差が最も小さかった KF による予測値を，現在の話者の推定位置とする．このような手法を用いることで，話者が一定の速度で移動する場合や，話者の移動速度の変化が激しい場合の両方に対応することができる．

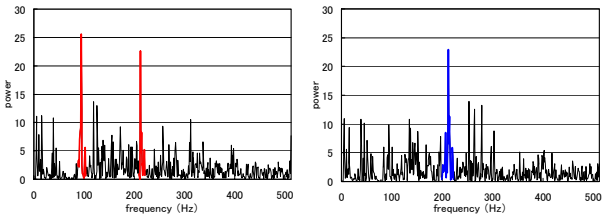
ここで，$p(t)$ を時刻 t における話者の位置の観測値，$\hat{K}_l(t)$ を時刻 t における KF l の予測値であるとする．また，KF の数を N とする．このとき，予測アルゴリズムは以下のように表される．

1. 時刻が KF の履歴長に満たないならば，今回の観測値のうちで前回の観測値と近いものを，その話者の現在の位置であるとする．また，その値を用いて KF を更新する．
2. そうでない場合，時刻 t における予測は次のように行われる．
   (1) 時刻 $t-1$ における誤差 $\|\hat{K}_l(t-1) - p(t-1)\|$ が最も小さい KF K_l を選択する．
   $$l = \underset{i=0,\cdots,N-1}{\arg\min} \left\| \hat{K}_i(t-1) - p(t-1) \right\| \tag{9}$$
   (2) 現在の複数の話者の位置を音源定位システムから得る．その中で，予測値 $\hat{K}_l(t)$ との差が閾値 δ よりも小さいものがあれば，それが現在の話者の位置であるとして選択される．
   (3) もし，予測値 $\hat{K}_l(t)$ との差が閾値 δ よりも小さいものが無ければ，l 番目の KF K_l を除いてステップ 2(1) へ戻る．
3. 得られた観測値 $p(t)$ を用いて，すべての KF を更新する．
4. 話者の真の状態を $p(t)$ を用いて推定し，ステップ 2 へ戻る．

本実験では，三つの KF を用い，それぞれの履歴長は 3, 5, 10 フレーム (120, 200, 400 [ms]) とした．

3. 音声の音響的特徴の使用による追跡精度の向上

複数話者の追跡における，交差，接近の曖昧性を解消するために，各話者の分離音声の音響的特徴を利用する．ここでは，音響的特徴として分離音声のパワースペクトルに注目する．2·3 節のアルゴリズムのステップ



(a) Speaker1　　(b) Speaker2

Fig.2 二話者のパワースペクトル

2(2) において，観測された話者方向における分離音声のパワースペクトルが，過去の話者のパワースペクトルと似ているものを同一の話者であるとして選択する．

分離音声のパワースペクトルは，遅延和を用いて計算される．各話者のきれいな分離音声が得られるならば，その分離音声の基本周波数を用いるなどのより精度の高い方法が考えられる．しかし，他の音源分離システムを用いた場合，その分離精度が追跡精度に影響する．本稿では提案手法の追跡精度のみによる追跡精度の向上を評価するために，遅延和による単純な方法を用いた．

話者の方向として，時刻 t において方向 d_s におけるスペクトル D_{d_s} は，遅延和を用いて以下のように求められる．

$$D_{d_s} = \sum_{i=0}^{M-1} x_i(t - \tau_{d_s,i}) \tag{10}$$

ここで，M はマイクロフォンの数であり，$\tau_{d_s,i}$ は方向 d_s におけるマイクロフォン i の遅延である．また，$x_i(t)$ は時刻 t におけるマイクロフォン i の入力信号である．さらに，次の式に従って高速フーリエ変換 (FFT) を行う．

$$\begin{aligned}\mathcal{F}[D_{d_s}] &= \mathcal{F}\left[\sum_{i=0}^{M-1} x_i(t - \tau_{d_s,i})\right] \\ &= \sum_{i=0}^{M-1} \exp\left(\frac{-2\pi i k \tau_{d_s,i}}{L} X_i(k)\right)\end{aligned} \tag{11}$$

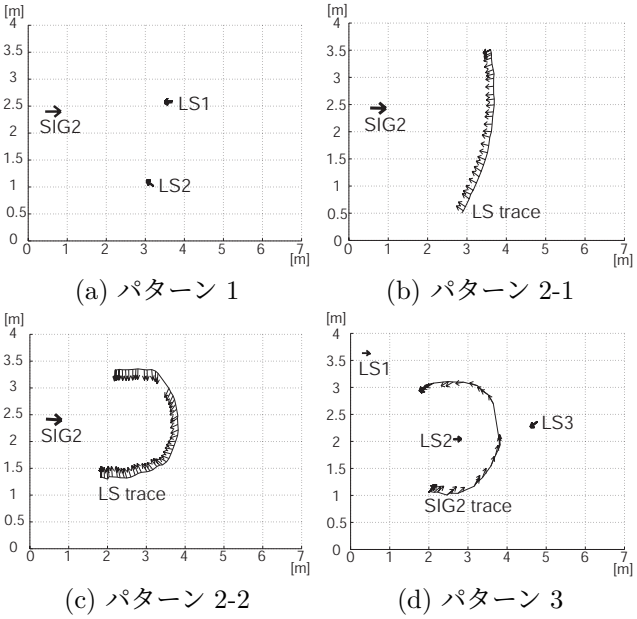
ここで，$\mathcal{F}[g]$ は，関数 g に対して窓幅 L の FFT を適用した結果であり，$X_i(k)$ は $x_i(t)$ に FFT を適用した結果である．

また，スペクトル間の類似度としてコサイン類似度を用いる．時刻 t において S 人の話者が観測されたとき，得られたそれぞれの話者の方向を $d_0, d_1, \ldots, d_{S-1}$ とする．このとき，時刻 $t-1$ において方向が d であった話者の現在の方向は，次の式によって求められる．

$$s = \underset{i=0,\cdots,S-1}{\arg\max}\ \mathcal{F}[D_{d_i}] \cdot \mathcal{F}[D_d]_{t-1} \tag{12}$$

ここで，$\mathcal{F}[D_d]_{t-1}$ は，時刻 $t-1$ において方向が d であった話者のスペクトルを表している．

例として，二人の話者が 15 度まで近づいたときの，それぞれの方向におけるパワースペクトルを図 2 に示す．録音はサンプリング周波数 48kHz で行い，FFT における窓幅は 1,024 フレームとした．基本周波数の異なる話者の場合，異なるパワースペクトルが得られることが分かる．



(a) パターン 1　　(b) パターン 2-1

(c) パターン 2-2　　(d) パターン 3

Fig.3 SIG2 とスピーカの位置
(“LS” は “loudspeaker” を示す)

4. 評価実験

本システムを評価するために，ヒューマノイドロボット SIG2(図 1) に 8 本のマイクロフォンを装着し，移動話者の追跡実験を行った．また，物体の正確な位置を測定するために，産業技術総合研究所で開発された超音波 3 次元タグ [7] を，実験室に設置した．我々の実験環境における測定誤差は，部屋の中央で約 5cm，壁際で約 10cm となっている．また，超音波 3 次元タグのサンプリング周波数は 20Hz である．実験室の壁のうち 1 面 (図 3 における y=0 の面) はガラス壁となっており，壁に近づくにつれ反響が大きくなっている．また，3 人の話者の発話内容として，ATR 音素バランス文から 3 種類の異なる文章を選んだ．

本実験室では，同時に一つの物体のみ正確な位置を測定できるため，次の三つのパターンにおいて実験を行った．

パターン 1 SIG2 と 1 つのスピーカを用い，両方を固定した．両者の間隔は 3m とした．また，スピーカは以下の場所に固定した．

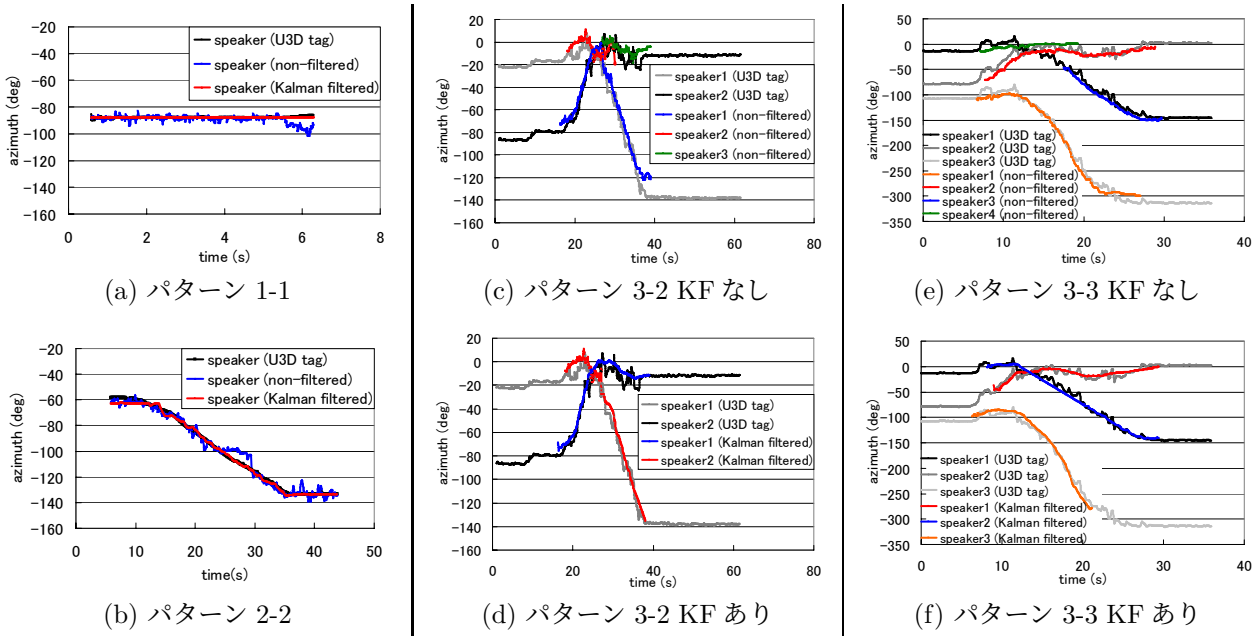
1. SIG2 の正面 (図 3(a) の LS1)
2. SIG2 の正面から右 30 度 (図 3(a) の LS2)

パターン 2 SIG2 と 1 つのスピーカを用い，SIG2 は固定し，スピーカを以下のように移動させた．

1. SIG2 を中心とした半径 3m の円周上を移動 (図 3(b))
2. SIG2 の周囲を非線形に移動 (図 3(c))

パターン 3 SIG2 といくつかのスピーカを用い，SIG2 を図 3(d) のように移動させた．スピーカの数は以下のようにした．

1. 1 つのスピーカ (図 3(d) の LS1)
2. 2 つのスピーカ (図 3(d) の LS1, LS2)
3. 3 つのスピーカ (図 3(d) の LS1, LS2, LS3)



(a) パターン 1-1　(c) パターン 3-2 KF なし　(e) パターン 3-3 KF なし

(b) パターン 2-2　(d) パターン 3-2 KF あり　(f) パターン 3-3 KF あり

Fig.4 追跡結果 (“U3D tag” は “ultrasonic 3D tag”(超音波 3 次元タグ) を示す．)

4·1　解析

8 本のマイクロフォンアレイを用い，サンプリング周波数 48kHz で録音を行い，音源定位を行った．そして，以下の手法について追跡精度の比較を行った．

ベースライン 音源定位結果が近いものを同一の話者であるとしてラベル付けを行う．

提案手法 複数の KF による予測値をもとに話者のラベル付けを行う．

超音波 3 次元タグによって測定される SIG2，スピーカの真の位置と，それぞれの手法によって得られた話者の位置の平均二乗誤差によって，評価を行った．

4·2　結果

スピーカの追跡結果を図 4 に，それぞれの手法によって得られた平均二乗誤差の結果を表 1 に示す．

図 4(b) より，提案手法では途切れることなく追跡が成功しているのに対し，ベースラインでは二つの軌跡に分割されていることが分かる．

また，パターン 3 において SIG2 では話者が交差するように観測される．そのため，交差したのか，接近後離れたのかという曖昧性が生じる．図 4(c), (d) より，ベースラインでは話者は交差していないと誤って追跡されているが，提案手法では正しく複数の話者が追跡されていることが分かる．また，このためベースラインにおけるパターン 3-2, 3-3 の平均二乗誤差は求められていない．

表 1 より，提案手法が，ロボットが動く場合にも話者が動く場合にも，定位誤差を減らすことに有効であることが示されている．

5.　結論

複数移動話者の追跡では，以下の問題が生じる．

- 同一話者への同一話者ラベルの付与
- 話者が交差したのか，接近後離れたのかといった追跡における曖昧性

Table 1 平均二乗誤差

パターン	ベースライン	提案手法
1-1	29 (deg^2)	0.34 (deg^2)
1-2	35 (deg^2)	1.0 (deg^2)
2-1	35 (deg^2)	2.1 (deg^2)
2-2	22 (deg^2)	3.6 (deg^2)
3-1	53 (deg^2)	25 (deg^2)
3-2	—	26 (deg^2)
3-3	—	28 (deg^2)

“—” は追跡の失敗を示す．

我々は，異なる履歴長を持つ複数の KF を用い，適切に KF の予測値を選択する手法を提案した．実験の結果，本手法が上記の問題を解決するのに有効な手法であることが示された．また，音声の音響的特徴を用いることで，話者が交差する場合であっても正確に追跡できることが示された．

今後，視覚情報との統合や音源分離システムの精度向上，音声認識システムへの応用なども検討していく予定である．

参考文献

[1] S. Yamamoto, *et al.*, “Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory,” *Proc. IEEE ICRA-2005.*

[2] K. Nakadai, *et al.*, “Real-time auditory and visual multiple-object tracking for robots,” *Proc. IJCAI-2001.*

[3] H. Asoh, *et al.*, “Tracking Human Speech Events using a Particle Filter,” *Proc. IEEE ICASSP-2005*, 1153–1156.

[4] 村瀬, *et al.*, “カルマンフィルタによる音声の時系列特徴を用いた複数移動話者の追跡,” 情報処理学会第 67 回全国大会講演論文集, 2-385–2-386.

[5] J.-M. Valin, *et al.*, “Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach,” *Proc. IEEE ICRA-2004.*

[6] E. Kalman, “A new approach to linear filtering and prediction problems,” *Trans. ASME – J. of Basic Engineering*, vol.82, 35–45, 1960.

[7] Y. Nishida, *et al.*, “3D Ultrasonic Tagging System for Observing Human Activity”, *Proc. IROS 2003.*