4R-8

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

村瀬 昌満† 山本 俊一‡ Jean-Marc Valin* 駒谷 和範‡ 尾形 哲也‡ 奥乃 博‡

† 京都大学 工学部情報学科 ‡ 京都大学大学院 情報学研究科 知能情報学専攻 * Université de Sherbrooke

1. はじめに

ロボットが人間と社会の中で共生するためには，複数話者の同時発話，移動話者，発話中の割り込みなどの様々な形態の音声インタラクションに対応できる必要がある．

一般に，実環境では話者やロボットが同じ方向を向いて立ち止まっているわけではなく，ロボットが動くと，相対的に話者が移動することもあるので，移動話者への対応は不可欠である．さらに，多くの音源分離手法では音源定位情報をパラメータとしているので，移動話者に対する音源定位は重要な機能である．しかし，静止音源については，音源定位，音源分離に関する多数の研究があるが，移動音源に対する音源定位，音源分離の研究は始まったばかりである．

浅野ら [1] はパーティクルフィルタによる移動話者の問題に取り組んでいる．しかし，パーティクルフィルタを用いるため実時間性に欠けるという欠点がある．また，中臺ら [2] はカルマンフィルタを用いた視聴覚統合による移動話者追跡に取り組んでいる．しかし，音声認識に利用できる程度の精度は得られていない．

本稿では履歴長の異なる複数のカルマンフィルタ (KF) を用いた移動話者の実時間追跡システムを開発する．また，追跡する上で問題となる複数の移動話者追跡問題を取り上げる．

2. システム構成

2.1 処理の流れ

本システムでは以下の流れで複数話者の追跡を行う．

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 [3] 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 により話者の位置を推定する．
3. 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，各ステップについて詳しく述べる．

2.2 各時刻における領域分割による音源定位

Valin らによって提案された steered beamformer は，8 本の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である．本研究ではこれを用いる．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 個の頂点について遅延和を計算し，そのパワーを計算する．計算されたパワーが大きい方向に音源があると推定する．

この手法により静止音源や話者が一人の場合には高精度での音源定位が可能となっている．しかし，この手法をそのまま複数移動話者の追跡に用いた場合，話者が交差や接近したときに追跡に失敗するという問題が生じる．

Tracking of Moving Talkers using Time Series Features of Speech based on Kalman Filter: *Masamitsu Murase, Shunichi Yamamoto (Kyoto Univ.), Jean-Marc Valin (Sherbrooke Univ.), Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

2.3 カルマンフィルタ

steered beamformer を移動音源に対して適用した際に生じる音源の移動による誤差を低減させ，話者の現在の位置の推定するために，履歴長の異なる KF を用いる．

KF では，時刻 k における観測値ベクトル $\boldsymbol{y}_k$ とシステムの内部状態ベクトル $\boldsymbol{x}_k$ が次のように線形に遷移するシステムに，ノイズが重畳される場合を想定している．

$$\boldsymbol{x}_{k+1} = F\boldsymbol{x}_k + G\boldsymbol{w}_k \quad (1)$$

$$\boldsymbol{y}_k = H\boldsymbol{x}_k + \boldsymbol{v}_k \quad (2)$$

ここで，行列 F, H はそれぞれ内部状態を更新する行列と内部状態を観測値に写像する行列であり，$\boldsymbol{w}_k, \boldsymbol{v}_k$ はそれぞれプロセスノイズと観測ノイズである．

本研究では，次のようにモデル化を行う．

用いる特徴量は，話者のアジマス角 θ，エレベーション角 ϕ，周波数毎の強度の一次元ベクトル $\boldsymbol{f}$ とし，時刻 k における特徴量ベクトル $\boldsymbol{p}_k$ を

$$\boldsymbol{p}_k = (\theta_k, \phi_k, \boldsymbol{f}_k) \quad (3)$$

とする．音響的特徴として分離音声の F0 を用いることが考えられるが，これは音源分離システムの性能に大きく依存するため，今回は用いていない．このとき，時刻 k における内部ベクトル $\boldsymbol{x}_k$ を過去 l フレームにおける履歴とし，次のように定義する．

$$\boldsymbol{x}_k = (\boldsymbol{p}_k, \boldsymbol{p}_{k-1}, \boldsymbol{p}_{k-2}, \ldots, \boldsymbol{p}_{k-l}) \quad (4)$$

また，$\boldsymbol{p}_{k+1}$ は $\boldsymbol{p}_k, \boldsymbol{p}_{k-l}$ を用いて次のような線形の状態遷移を仮定する．

$$\boldsymbol{p}_{k+1} = \boldsymbol{p}_k + (\boldsymbol{p}_k - \boldsymbol{p}_{k-l})/l \quad (5)$$

このとき，F, G, H は次のように表される．

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

$$G = (I\ 0\ \cdots\ 0)^T \quad (7)$$

$$H = (I\ 0\ \cdots\ 0) \quad (8)$$

ただし，ここで I は 3×3 の単位行列である．

2.4 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ので話者の非線形な動きや音声のような非線形な周波数変化，さらに発話の停止や話者数の変化などにより，精度の低下が生じる．具体的には，話者の動きが一定である場合，履歴が長いフィルタで予測する方が，よりノイズが除去された軌跡を得ることができる．これに対して，話者の動きの変動が激しい場合，履歴の短いフィルタでなければ，観測値と予測値の間に大きな差が生まれてしまい，正確な追跡が困難となる．

図 1: SIG2 とマイクロフォンアレイ (矢印)

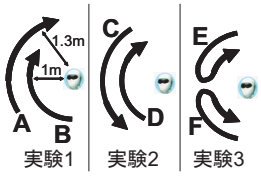

図 2: 話者の動き

長さの異なる履歴を持つ複数の KF を用いて同時に予測を行うことにより，より正確な追跡を実現する．具体的には，KF の数を N，時刻 t における観測値を $p(t)$，それに対するフィルタ l の予測値を $\hat{K}_l(t)$ とするとき，各話者の位置予測を以下のように行う．

1. 時刻 t において，次のように予測を行う．
   (a) 以下の式で表される l 番目の KF K_l を用いて，予測値 $\hat{K}_l(t)$ を得る．つまり，前回の時刻における観測値と予測値の誤差が最も小さいフィルタを選択する．

$$l = \arg\min_{i=1,\ldots,N} \left|\left| \hat{K}_i(t-1) - p(t-1) \right|\right| \quad (9)$$

   (b) 音源定位によって得られた話者の位置情報の観測値の中で，予測値 $\hat{K}_l(t)$ との差が閾値 δ 以下のものがあれば，それが，この話者の時刻 t における位置であると推定し，定位結果に話者ラベルを付与する．
   (c) 予測値 $\hat{K}_l(t)$ との差が閾値 δ 以下のものがなければ，l 番目の KF を除き，再び 1a へ戻る．
2. 得られた観測値 $p(t)$ を用いて全ての KF を更新し，再び 1 へ戻る．

3. 話者の追跡実験

KF による予測の性能を評価するために，話者の追跡実験を行った．

3.1 実験条件

録音はヒューマノイドロボット SIG2 の外装に 8 本のマイクロフォンを取り付けて行った (図 1)．2 話者の動きは次の 3 パターンである (図 2)．

実験 1 2 話者が 60° 離れた状態から交差せず平行に移動

実験 2 2 話者が 140° 離れた状態から交差するように移動

実験 3 2 話者が 140° 離れた状態から一度近づき再び離れるように移動

具体的には 2 人がスピーカを持ち移動し，ATR 音素バランス文を発話させ，この音声をサンプリング周波数 48kHz で録音した．

3.2 解析方法

音声に対し 0.04 秒毎に領域分割による音源定位を実行した．この結果に対し，実験 1, 2 においては KF を用いずに前フレームでの話者位置をもとにラベル付けしたものと，KF による予測値をもとにラベル付けをしたものを比較した．また，実験 3 においては履歴数が 5, 10, 15 フレームである複数の KF を用いた場合と，履歴数が 10 フレームで固定の KF のみを用いた場合を比較した．

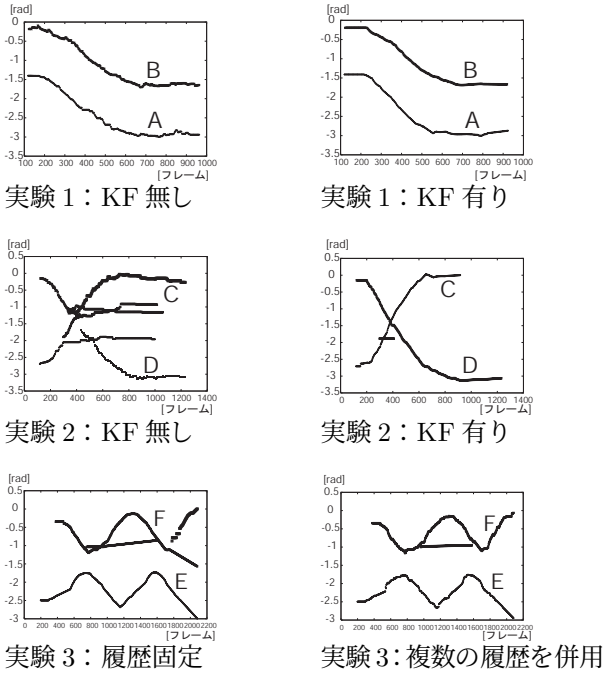

図 3: 実験結果

3.3 実験結果

それぞれの動きのパターンについて実験結果を図 3 に示す．図は，各フレームにおける音源の方向 (角度) を示している．1 フレームは 0.04 秒である．

3.4 考察

実験 1,2 の結果から，本研究での手法が音源定位誤差の軽減，話者が交差する際の追跡に効果的であることが示された．実験 2 では KF を用いない場合，2 話者が近付いて離れたように観測されているが，KF を用いることによって，追跡が成功していることが分かる．実験 2 の KF 無しの結果で 3 つ以上の音源が観測され続けているのは，話者の交差によって音源を見失い，その後も話者を探索し続けていることを示すものである．

また，実験 3 で履歴数が固定されている場合では，話者 E の 3 回目の方向転換のような急な方向転換に対応できていないのに対し，複数の履歴を用いた場合では追跡が可能となっていることが分かる．実験 3 においてはノイズのために 3 つ以上の音源が観測されているが，これは視覚情報などを統合や，分離音からの feedback を特徴量として利用することにより取り除くことができると考えられる．

4. おわりに

音声時系列の特徴を用い，複数の KF を用いることで複数話者が交差するように移動する場合にも，それぞれの話者の定位がより正確となり，適切に追跡できることが確認された．今後，本手法の定量的評価を行い，視聴覚情報の統合についても検討する予定である．

なお，本研究の一部は，科研費，21 世紀 COE，SCAT 研究助成 の支援を受けた．

参考文献

[1] 浅野他: マイクロフォンアレイを用いた移動音源の追跡と分離について，人工知能学会 AI チャレンジ研究会，1-8,2004.
[2] Nakadai, K. et al.: Real-Time Auditory and Visual Multiple-Object Tracking for Robots. *IJCAI-01*,1425-1432.
[3] Valin,J.-M. et al.: Localization of Simultaneous Moving Sound Sources for Mobile Robot Using a Frequency-Domain Steered Beamformer Approach.*IEEE ICRA2004*, 1033-1038.