音響伝達特性を用いた単一マイクロホンによる話者の頭部方向の推定*

☆高島遼一，滝口哲也，有木康雄 (神戸大)

1 はじめに

人と人のコミュニケーション，あるいはロボットとのコミュニケーションにおいて，話者の頭部方向は聞き手にとって重要な手がかりの一つであり，我々は話し手の頭部の向きから「誰が話しているのか」だけでなく,「誰に向かって話しているのか」という情報まで得ることができる．この「誰が誰に向かって話しているのか」という情報は，特に複数のユーザーが会話をしている状況において有効であり，会議システム，ロボット対話，雑談とシステム要求の判別など，様々なタスクにおいて利用することができると期待される．

これまでに，マイクロホンアレーを用いて音源方向や位置を推定する研究が多くなされている．MUSIC (MUltiple SIgnal Classification) や CSP (Cross-power Spectrum Phase) といった手法では，マイクロホンアレーで収録される観測信号間の位相差を用いて音源方向や位置を推定している [1, 2, 3, 4]．また，バイノーラル信号を用いて，両耳間の音圧差や時間差から音源方向を推定する手法についても研究されている [5, 6]．

一方，話者の頭部方向の推定へ関心が向けられ出したのは比較的近年のことであり，いくつかの手法が提案されている [7, 8, 9, 10]．これらの手法は複数組のマイクロホンアレーからなるネットワークを用いており，従来の音源位置推定のアルゴリズムを拡張することで話者の頭部方向を推定している．文献 [7] で提案されている手法は，従来の音源位置推定法の一つである SRP-PHAT (Steered Response Power with the PHAse Transform) をベースとした手法であり，従来の SRP-PHAT の目的関数を，話者の頭部方向に依存する重み係数によって重み付けを行うことにより，話者の位置推定問題から頭部方向推定問題へ拡張している．文献 [8] では話者の頭部の方向ごとに変化する観測信号の音圧のパターンに着目しており，文献 [9] では SRP-PHAT を用いた手法と音圧パターンを用いた手法の両方を提案し組み合わせることでさらなる精度の向上を示している．また，文献 [10] では各マイクロホンアレーから算出された音源方向推定結果を用いて作成されるヒストグラムから，話者の頭部方向を推定する手法を提案している．

しかしながらこれらの手法は複数のマイクロホンアレーを，ユーザーを囲むようにして部屋の壁などに設置する必要があり，システムが大規模になってしまうという欠点がある．我々はこれまで，観測された音声信号の音響伝達特性が，発話された位置によって異なるという点に着目して，位置毎に発話された音声から音響伝達特性を推定し，それらを識別することにより単一マイクロホンで音源位置を推定する方法を提案してきた [11]．この手法では，ある位置から発話された音声からその音響伝達特性を，特定話者の音素 HMM を用いて推定し，推定された音響伝達特性を位置毎に学習する．その後，ある位置から発話された評価音声についても同様に音響伝達特性を推定し，それを識別することで音源の位置を推定する．

本稿では観測信号の音響伝達特性が，話者の位置だけではなく頭部方向にも依存することに着目し，音響伝達特性の識別による話者の頭部方向を推定する手法を提案する．以前に提案した音源位置推定手法では，話者の位置ごとの音響伝達特性を学習・識別していたのに対し，本稿における提案手法では，各音源位置とその位置における各頭部方向の音響伝達特性を学習・識別する．従来の頭部方向の推定法と異なり，本手法はあらかじめ音響伝達特性を学習しておく必要があるが，マイクの位置を任意の場所に設置することができるという利点がある．評価実験では実環境下において音源位置のみの推定，頭部方向のみの推定，音源位置及び頭部方向の推定の 3 つのタスクにおいて実験を行い，その有効性を示す．

2 音源位置と頭部方向の推定

2.1 提案手法の概要

本研究では音響伝達特性を用いて音源の位置と頭部方向を推定する．音響伝達特性は音源の位置や頭部方向によって異なる値を持つため，あらかじめこれを各音源位置とその位置における頭部方向毎に学習しておけば，評価音声に対してもその音響伝達特性を識別することで音源位置及び頭部方向を推定することができる．

提案手法の概要を Fig. 1 に示す．まず，位置と頭部方向の組み合わせ毎の音響伝達特性を学習するために，それぞれの位置 θ において頭部を各方向 ϕ へ向けた状態で発話された音声 $O_{train}^{(\phi,\theta)}$ を収録し，その音響伝達特性をクリーン音声の音素 HMM を用いて推定する．次に，位置と頭部方向毎に推定された音響伝達特性 $\hat{H}_{train}^{(\phi,\theta)}$ を SVM (Support Vector Machine) により学習する．そして評価したい音声 $O_{test}^{(\phi,\theta)}$ についても学習データと同様に，音素認識結果により得られるラベル情報を用いて音響伝達特性 $\hat{H}_{test}^{(\phi,\theta)}$ を推定し，それを SVM で識別することで，音源位置と頭部方向 $(\hat{\phi},\hat{\theta})$ を推定する．

2.2 音素 HMM による音響伝達特性の推定

本節では音素 HMM を用いて観測信号 O から音響伝達特性 H を推定する手法について述べる．ある場

*Single-channel head orientation estimation based on discrimination of acoustic transfer function. by Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

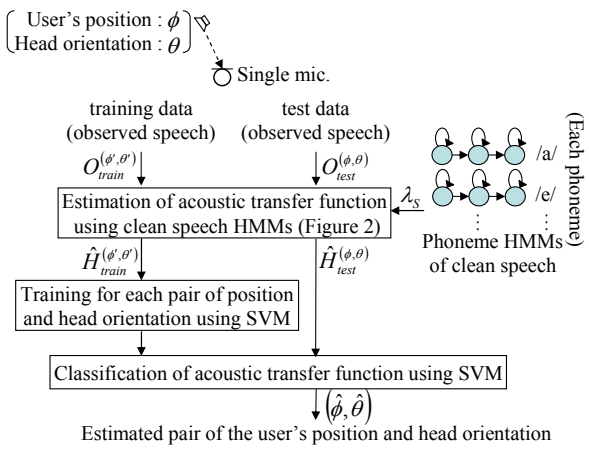

Fig. 1　提案手法の概要

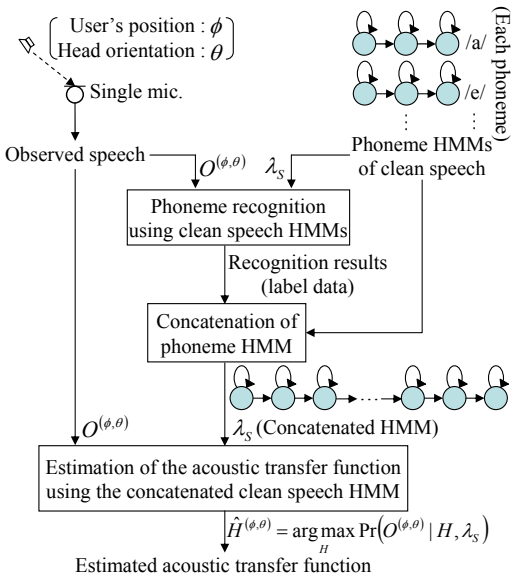

Fig. 2　音素 HMM による音響伝達特性の推定

所で発声されたクリーン音声 S は，音響伝達特性 H の影響を受けて観測される．このとき，フレーム n における観測信号 O のケプストラムは，

$$O_{cep}(d;n) \approx S_{cep}(d;n) + H_{cep}(d;n) \tag{1}$$

と近似される．d はケプストラムの次元を表す．ケプストラムは，音声情報を効率よく表現できるパラメータの一つであり，音声認識などでよく用いられていることから，本手法においてもケプストラムを特徴量として用いている．仮に S が既知であれば，音響伝達特性 H は

$$H_{cep}(d;n) \approx O_{cep}(d;n) - S_{cep}(d;n) \tag{2}$$

として求めることができるが，実際の環境では S が未知であるため，直接 H を求めることはできない．そこで，S の統計モデルをあらかじめ学習しておき，最尤推定法により O から H を推定する．

音響伝達特性の推定の流れを Fig. 2 に示す．あらかじめ特定話者のクリーン音声の MFCC を音素 HMM でモデル化しておく．HMM を用いて音響伝達特性を推定するためには，その音声信号の音素ラベルが必要であるため，まず学習した音素 HMM を用いて観測信号を音素認識する．そして出力された音素認識結果をラベルとして音素 HMM を連結し，連結された HMM を用いて観測信号から最尤推定法により音響伝達特性の MFCC を推定する．

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|\lambda_S, H) \tag{3}$$

λ_S はクリーン音声のモデルパラメータを表す．(3) 式の解は EM アルゴリズムによって推定される．その際，Q 関数は以下のように定義される．

$$\begin{aligned} & Q(\hat{H}|H) \\ &= E[\log \Pr(O,p,b_p,c_p|\hat{H},\lambda_S)|H,\lambda_S] \\ &= \textstyle\sum_p \sum_{b_p} \sum_{c_p} \frac{\Pr(O,p,b_p,c_p|H,\lambda_S)}{\Pr(O|H,\lambda_S)} \\ &\quad \cdot \log \Pr(O,p,b_p,c_p|\hat{H},\lambda_S) \end{aligned} \tag{4}$$

b_p と c_p はそれぞれ音素 p における HMM の状態，混合要素を表す．O，p，b，c の同時確率 $\Pr(O,p,b_p,c_p|\hat{H},\lambda_S)$ は以下のように展開される．

$$\begin{aligned} & \Pr(O,p,b_p,c_p|\hat{H},\lambda_S) \\ &= \textstyle\prod_n a_{b_p(n-1),b_p(n)} w_{b_p(n),c_p(n)} \\ &\quad \cdot \Pr(O(n)|p,b_p(n),c_p(n);\hat{H},\lambda_S) \end{aligned} \tag{5}$$

n，a，w はそれぞれフレーム番号，状態遷移確率，混合重みを表す．ここで，(1) 式より O は S と H の加算とみなされるため，O の事後確率をクリーン音声 HMM を用いて以下のように表すことができる．

$$\begin{aligned} & \Pr(O,p,b_p,c_p|\hat{H},\lambda_S) \\ &= \textstyle\prod_n a_{b(n-1),b(n)} w_{b(n),c(n)} \\ &\quad \cdot N(O(n);\mu^{(S)}_{p,j,k} + \hat{H}_{(n)}, \Sigma^{(S)}_{p,j,k}) \end{aligned} \tag{6}$$

$N(O;\mu,\Sigma)$ は多次元正規分布を表し，$\mu^{(S)}_{p,j,k}$，$\Sigma^{(S)}_{p,j,k}$ はそれぞれ S の状態 $b(n)=j$，混合要素 $c(n)=k$ における平均ベクトルと共分散行列 (対角行列) を表す．これらを用いて (4) 式を展開し，H に関わる項のみを取り出すと以下のようになる．

$$\begin{aligned} & Q(\hat{H}|H) \\ &= -\textstyle\sum_p \sum_j \sum_k \sum_n \gamma_{p,j,k}(n) \\ &\quad -\textstyle\sum_{d=1}^{D} \left\{ \frac{1}{2}\log(2\pi)^D \sigma^{(S)^2}_{p,j,k,d} \right. \\ &\quad \left. + \frac{(O(d;n)-\mu^{(S)}_{p,j,k,d}-\hat{H}(d;n))^2}{2\sigma^{(S)^2}_{p,j,k,d}} \right\} \end{aligned} \tag{7}$$

$$\gamma_{p,j,k}(n) = \Pr(O(n),p,j,k|\lambda_S) \tag{8}$$

D は次元数，$\mu^{(S)}_{p,j,k,d}$，$\sigma^{(S)^2}_{p,j,k,d}$ はそれぞれ平均ベクトルの d 次元目の値と，共分散行列の d 番目の対角要素の

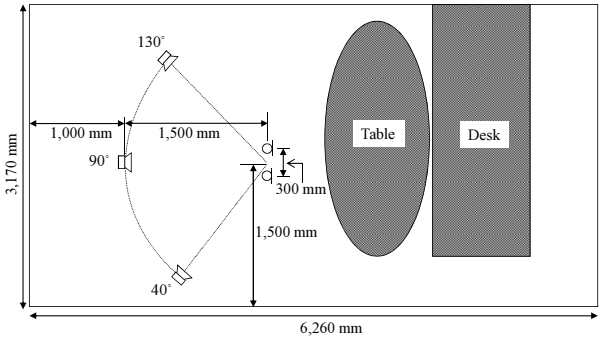

Fig. 3　実験環境

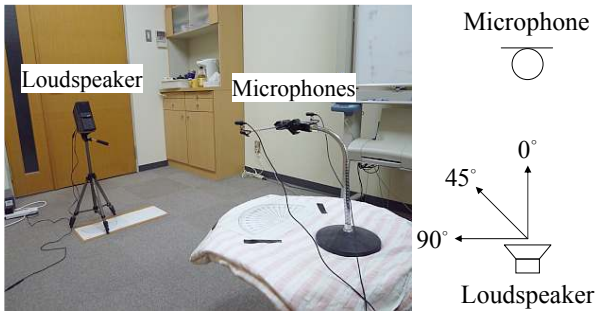

Fig. 4　収録環境のとスピーカーの回転方向

値を表す．(7) 式を最大にする H は，$\partial Q(\hat{H}|H)/\partial\hat{H} = 0$ を解くことで求められる．

$$\hat{H}(d;n) = \frac{\sum_p \sum_j \sum_k \gamma_{p,j,k}(n) \frac{O(d;n) - \mu^{(S)}_{p,j,k,d}}{\sigma^{(S)2}_{p,j,k,d}}}{\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k}(n)}{\sigma^{(S)2}_{p,j,k,d}}}. \quad (9)$$

(9) 式によって，音源位置と頭部方向毎に音響伝達特性を推定し，それらを SVM によって学習・識別することで，音源位置と頭部方向の推定を行う．

3　評価実験

3.1　実験環境

提案手法を評価するために，特定話者による実環境実験を行った．実験環境を Fig. 3 に，収録環境とスピーカーの回転の様子を Fig. 4 に示す．約 6.3 m × 3.2 m × 2.8 m (W × D × H) の部屋において，ある位置にスピーカーを設置し，スピーカーの向きを変えながら特定話者の音声を再生し，これを 2ch マイクロホンで収録した．ただし，提案手法では 2ch マイクロホンの内，一方のマイクロホンで収録された音声のみを用いた．部屋の残響時間は約 350 msec，マイクロホンとスピーカーの距離は約 1.5 m である．スピーカーは BOSE Mediamate II を，マイクロホンには指向性マイク (SONY ECM-66B) を使用した．音源位置の候補は 40°，90°，130° の 3 種類，スピーカーの回転方向 (頭部方向) は 0°，45°，90° の 3 種類で，これらの組み合わせは計 9 通り存在する．

音声データは ATR 研究用日本語音声データベースセット A より男性話者 1 名の単語音声を用い，サンプリング周波数 12 kHz，窓幅 32 msec，フレームシフト 8 msec の分析条件で MFCC 16 次元を特徴量として使用した．音響伝達特性の推定におけるクリーン音声の音素 HMM は，2,620 単語を用いて学習した．音素数は 54，各音素 HMM の状態数は 3，混合数は 32 である．音響伝達特性の学習には 50 単語を，評価には 166 単語を，組み合わせを変えて 4-fold のクロスバリデーションにより推定精度を算出した．なお，クリーン音声の学習，位置及び頭部方向の学習，評価に用いたデータはそれぞれ異なる発話内容の単語を使用している．SVM には SVM^{light} を，カーネル関数に RBF (Gaussian) カーネルを使用し，one-vs-rest 法によりマルチクラス識別を行った．

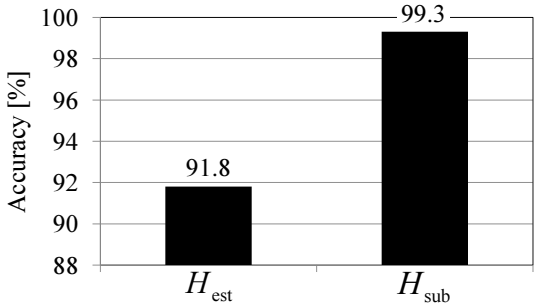

Fig. 5　各手法における音源位置推定精度 (3 クラス識別)

3.2　実験結果

本実験では二種類の手法について比較を行った．一方は本稿における提案手法で，クリーン音声 HMM を用いて観測信号から推定した音響伝達特性を識別する手法である．以降この手法を H_{est} と呼ぶ．もう一方は式 (2) に正解のクリーン音声 MFCC を代入することで得た音響伝達特性を識別する手法である．この手法では本来未知であるクリーン音声情報を与えているため，提案手法に比べてより真の値に近い音響伝達特性が得られている．以降この手法を H_{sub} と呼ぶ．

まず，それぞれの音源位置において，スピーカーの頭部方向を 0° に限定して，音源位置推定の精度評価を行った．各手法における音源位置推定精度を Fig. 5 に示す．提案手法で約 92 %，H_{sub} を用いた手法では約 99 % の認識率で音源位置 3 か所の分類が行えている．

次に，それぞれの場所にスピーカーの位置を固定し，スピーカーの頭部方向のみを変えることで，頭部方向の推定における精度の評価を行った．スピーカーの各頭部方向における推定精度を Fig. 6 に示す．提案手法では，スピーカーの頭部方向が 0° と 90° の場合において，84 % 以上の精度で頭部方向を推定することができているが，45° の場合，推定精度が著しく低下している．一方，真の音響伝達特性の値に近い H_{sub} を用いた場合では 45° においても約 90 % の認識率で推定が行えている．

最後に，音源の位置とスピーカーの回転方向の両

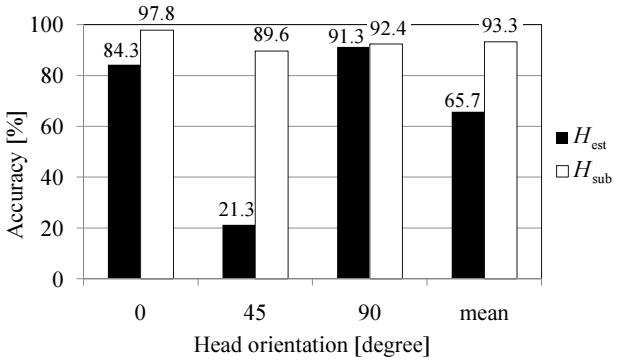

Fig. 6　スピーカーの各頭部方向における，頭部方向推定精度 (3 クラス識別)

Table 1　それぞれの位置 (pos.)，スピーカーの頭部方向 (ori.) における，音源位置と頭部方向の推定精度 (9 クラス識別)．上表は提案手法における推定精度，下表は H_{sub} における推定精度を表す．

pos. \ ori.	0 deg.	45 deg.	90 deg.	mean
40 deg.	44.3	15.7	68.2	42.7
90 deg.	83.7	29.8	84.9	66.2
130 deg.	76.8	50.8	87.5	71.7
average	68.3	32.1	80.2	60.2

pos. \ ori.	0 deg.	45 deg.	90 deg.	mean
40 deg.	96.8	84.6	86.6	89.4
90 deg.	97.6	86.4	91.3	91.8
130 deg.	97.0	94.4	98.2	96.5
average	97.1	88.5	92.0	92.6

方を変化させて，音源位置と頭部方向の同時推定精度を評価した．それぞれの位置，スピーカーの頭部方向における推定精度を Table 1 に示す．音源位置 3 箇所，頭部方向 3 方向の計 9 クラス識別において，提案手法では平均約 60 % ，H_{sub} を用いた場合で平均約 93 % の認識精度で識別が行えている．また，これまでの実験結果と同様に，提案手法ではスピーカーの頭部方向 45° において推定精度が低下することが分かる．

4　おわりに

本稿では，音響伝達特性が話者の位置や頭部方向によって異なる特性を持つ点に着目し，各音源位置・頭部方向において発話された音声信号から，その音響伝達特性をクリーン音声の音素 HMM を用いて推定し，推定された音響伝達特性を SVM により学習・識別することで，音源の位置と頭部方向をシングルチャネルで推定する手法について検討を行った．音源位置 3 箇所，頭部方向 3 方向の 9 クラス識別実験において，クリーン音声情報を与えて算出した音響伝達特性を用いた場合，9 割以上の認識率で推定が行えているのに対し，提案手法により推定した音響伝達特性を用いた場合では，頭部方向 45° の推定精度が著しく悪く，0° や 90° のような大きな頭部方向の変化しか識別できていないことが分かった．この理由として，提案手法では音響伝達特性を正確に推定しきれておらず，発話内容によって異なるクリーン音声成分がノイズとなったために，音響伝達特性の微小な変化を識別できなかったことが考えられる．そのため，より正確な音響伝達特性の推定が今後の課題として挙げられる．また，テスト環境の位置や頭部方向が少しずれた場合の推定精度や，未知の位置・頭部方向の推定についても今後検討を行う．

謝辞　本研究は日本学術振興会特別研究員奨励費 (23·2495) の助成を受けたものである．

参考文献

[1] D. Johnson and D. Dudgeon, “Array Signal Processing,” Prentice Hall, 1996.
[2] M. Omologo and P. Svaizer, “Acoustic event localization in noisy and reverberant environment using CSP analysis,” Proc. ICASSP96, pp. 921-924, 1996.
[3] F. Asano and H. Asoh and T. Matsui, “Sound source localization and separation in near field,” IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, E83-A, pp. 2286-2294, 2006.
[4] Y. Denda and T. Nishiura and Y. Yamashita, “Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation,” IEICE Trans. on Information and Systems, E89-D, pp. 1050-1057, 2000.
[5] F. Keyrouz and Y. Naous and K. Diepold, “A new method for binaural 3-D localization based on HRTFs,” Proc. ICASSP06, pp. V-341-V-344, 2006.
[6] M. Takimoto and T. Nishino and K. Takeda, “Estimation of a talker and listener’s positions in a car using binaural signals,” The Fourth Joint Meeting ASA and ASJ, pp. 3216, 2006.
[7] A. Brutti, M. Omologo, and P. Svaizer, “Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays,” Proc. Interspeech05, pp. 2337-2340, 2005.
[8] J. M. Sachar and H. F. Silverman, “A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array,” Proc. ICASSP04, vol. 4, pp. 65-68, 2004.
[9] C. Segura, A. Abad, J. Hernando and C. Nadeu, “Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR,” Proc. Interspeech08, pp. 1325-1328, 2008.
[10] M. Togami and Y. Kawaguchi, “Head orientation estimation of a speaker by utilizing kurtosis of a DOA histogram with restoration of distance effect,” Proc. ICASSP10, pp. 133-136, 2010.
[11] R. Takashima, T. Takiguchi, Y. Ariki, “HMM-based Separation of Acoustic Transfer Function for Single-channel Sound Source Localization,” ICASSP2010, pp. 2830-2833, 2010.