音響伝達特性を用いたシングルチャネル音源位置推定における未学習位置の推定*

☆高島遼一，滝口哲也，有木康雄 (神戸大)

1 はじめに

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) といった従来手法では，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].

しかしながら，マイクロホンアレーを用いた音声インターフェースは，システムが大規模になってしまうという欠点がある．そのため，小型な音声インターフェースが必要とされる環境では，単一マイクロホンで行える音声処理技術の需要が高まっており，近年では雑音抑圧や音源分離の分野においても，単一マイクロホンで処理できる手法が多く提案されてきている [2, 3].

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 [4]．この手法ではまず，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．

この手法により，単一マイクロホンでも音源の位置を識別することが可能となった．しかし，この手法は事前に想定される音源位置毎に音響伝達特性を学習させる必要があり，学習していない位置の推定が出来ないという問題があった．

そこで，本稿では限られた位置の音響伝達特性を用いて，音響伝達特性から位置への回帰モデルを学習することで，未学習位置の推定もその音響伝達特性と回帰モデルを用いて行うという手法について検討する．本稿では回帰モデルとしては重回帰分析を採用し，実環境下で収録したインパルス応答とそれを畳みこんだ音声を用いてその性能を評価した．

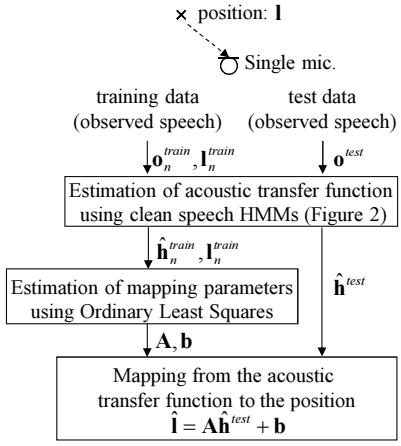

Fig. 1 提案手法の概要

2 音源位置の推定

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を推定している．音響伝達特性は音源の位置によって異なる値を持つため，これを用いて音源の位置を推定することができる．以前に提案していた手法では，あらかじめ位置毎に音響伝達特性を学習し，評価音声に対してその音響伝達特性を識別することで音源の位置を推定していた．しかしこの手法は伝達特性を学習した位置のみしか識別できないため，想定する位置の候補が増える度に，その位置の伝達特性を取得して学習する必要があった．

そこで本手法では，限られた位置の音響伝達特性を用いて，音響伝達特性から位置への回帰モデルを学習することで，学習データに含まれなかった位置についても，その回帰モデルを用いて推定を行う．

提案手法の概要を Fig. 1 に示す．まず，学習用の音響伝達特性を得るために，特定の位置 $\mathbf{l}^{train}$ で発話された音声 $\mathbf{o}^{train}$ 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，学習データから推定された音響伝達特性のケプストラム $\hat{\mathbf{h}}^{train}$ と位置のラベル $\mathbf{l}^{train}$ とのペアから，伝達特性から位置への変換パラメータ $\mathbf{A}$，$\mathbf{b}$ を重回帰分析により推定する．

そして評価したい音声 $\mathbf{o}^{test}$ についても学習データと同様にして音響伝達特性 $\hat{\mathbf{h}}^{test}$ を推定し，学習した変換パラメータを用いて音源位置 $\hat{\mathbf{l}}$ を推定する．

*Single-channel talker localization for unlearned position using the acoustic transfer function. by Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

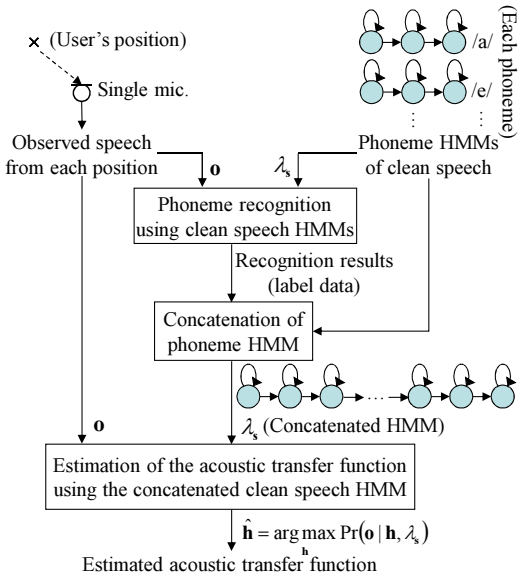

Fig. 2　音素 HMM による音響伝達特性の推定

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 $\mathbf{o}$ から音響伝達特性 $\mathbf{h}$ を推定する手法について述べる．ある場所で発声されたクリーン音声 $\mathbf{s}$ は，音響伝達特性 $\mathbf{h}$ の影響を受けて観測される．本研究では，フレーム n における観測信号 $\mathbf{o}_n$ のケプストラムを以下のような加算モデルで近似することにする．

$$\mathbf{o}_n \approx \mathbf{s}_n + \mathbf{h}_n \tag{1}$$

$\mathbf{o}$，$\mathbf{s}$，$\mathbf{h}$ はそれぞれ D 次元のベクトルである．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 $\mathbf{s}$ が既知であれば，音響伝達特性 $\mathbf{h}$ は

$$\mathbf{h}_n \approx \mathbf{o}_n - \mathbf{s}_n \tag{2}$$

として求めることができるが，実際の環境では $\mathbf{s}$ が未知であるため，直接 $\mathbf{h}$ を求めることはできない．そこで，$\mathbf{s}$ 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 $\mathbf{o}$ から $\mathbf{h}$ 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．

$$\hat{\mathbf{h}} = \underset{\mathbf{h}}{\operatorname{argmax}} \Pr(\mathbf{o}|\lambda_{\mathbf{s}}, \mathbf{h}) \tag{3}$$

$\lambda_{\mathbf{s}}$ 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．

$$\begin{aligned} & Q(\hat{\mathbf{h}}|\mathbf{h}) \\ &= E[\log \Pr(\mathbf{o}, p, b_p, c_p|\hat{\mathbf{h}}, \lambda_{\mathbf{s}})|\mathbf{h}, \lambda_{\mathbf{s}}] \\ &= \textstyle\sum_p \sum_{b_p} \sum_{c_p} \frac{\Pr(\mathbf{o}, p, b_p, c_p|\mathbf{h}, \lambda_{\mathbf{s}})}{\Pr(\mathbf{o}|\mathbf{h}, \lambda_{\mathbf{s}})} \\ &\quad \cdot \log \Pr(\mathbf{o}, p, b_p, c_p|\hat{\mathbf{h}}, \lambda_{\mathbf{s}}) \end{aligned} \tag{4}$$

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．$\mathbf{o}$，p，b，c の同時確率 $\Pr(\mathbf{o}, p, b_p, c_p|\hat{\mathbf{h}}, \lambda_{\mathbf{s}})$ は以下のように展開される．

$$\begin{aligned} & \Pr(\mathbf{o}, p, b_p, c_p|\hat{\mathbf{h}}, \lambda_{\mathbf{s}}) \\ &= \textstyle\prod_n a_{b_{p,n-1}, b_{p,n}} w_{b_{p,n}, c_{p,n}} \\ &\quad \cdot \Pr(\mathbf{o}_n|p, b_{p,n}, c_{p,n}; \hat{\mathbf{h}}, \lambda_{\mathbf{s}}) \end{aligned} \tag{5}$$

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 $\mathbf{o}$ は $\mathbf{s}$ と $\mathbf{h}$ の加算とみなされるため，$\mathbf{o}$ 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．

$$\begin{aligned} & \Pr(\mathbf{o}, p, b_p, c_p|\hat{\mathbf{h}}, \lambda_{\mathbf{s}}) \\ &= \textstyle\prod_n a_{b_{n-1}, b_n} w_{b_n, c_n} \\ &\quad \cdot N(\mathbf{o}_n; \mu^{(\mathbf{s})}_{p,j,k} + \hat{\mathbf{h}}_n, \Sigma^{(\mathbf{s})}_{p,j,k}) \end{aligned} \tag{6}$$

$N(\mathbf{o}; \mu, \Sigma)$ は多次元正規分布を表し，$\mu^{(\mathbf{s})}_{p,j,k}$，$\Sigma^{(\mathbf{s})}_{p,j,k}$ はそれぞれ $\mathbf{s}$ 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，$\mathbf{h}$ に関わる項のみを取り出すと以下のようになる．

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

$$\gamma_{p,j,k,n} = \Pr(\mathbf{o}_n, p, j, k|\lambda_{\mathbf{s}}) \tag{8}$$

D は次元数，$\mu^{(\mathbf{s})}_{p,j,k,d}$，${\sigma^{(\mathbf{s})}_{p,j,k,d}}^2$ 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 番目の対角要素の値を表す．(7) 式を最大にする $\mathbf{h}$ は，$\partial Q(\hat{\mathbf{h}}|\mathbf{h})/\partial\hat{\mathbf{h}} = 0$ を解くことで求められる．

$$\hat{h}_{d,n} = \frac{\sum_p \sum_j \sum_k \gamma_{p,j,k,n} \frac{o_{d,n} - \mu^{(\mathbf{s})}_{p,j,k,d}}{{\sigma^{(\mathbf{s})}_{p,j,k,d}}^2}}{\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k,n}}{{\sigma^{(\mathbf{s})}_{p,j,k,d}}^2}}. \tag{9}$$

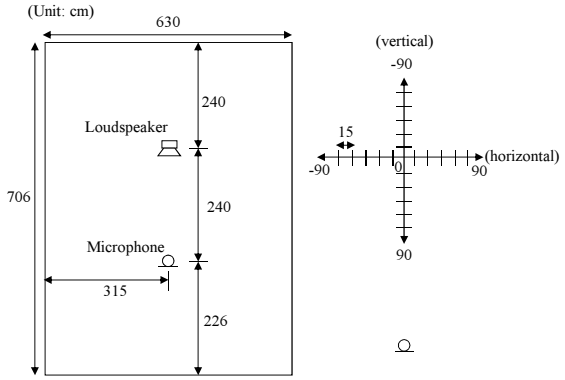

Fig. 3　インパルス応答の収録環境 (左) とスピーカーの位置 (右)

2.3　重回帰分析による音響伝達特性から位置へのマッピング

学習位置で得られた音響伝達特性と，その位置のラベルを用いて，音響伝達特性から位置への回帰モデルを学習する．

$$\mathbf{l} = \mathbf{A}\mathbf{h} + \mathbf{b} \tag{10}$$

ここで，位置ラベル $\mathbf{l}$ は例えば位置を 3 次元座標系で表現した場合は 3 次元のベクトルとなり，回帰パラメータ $\mathbf{A}, \mathbf{b}$ はそれぞれ行列，ベクトルとなる．しかし本稿における実験では，高さ・垂直方向を固定して水平方向のみ移動させるというように，他の 2 次元を固定して 1 次元のみを独立に評価しているため，この場合 $\mathbf{l}, \mathbf{b}$ はスカラー，$\mathbf{A}$ は行ベクトルとなる．

重回帰分析では回帰パラメータを最小二乗法により求める．

$$\min_{\mathbf{A},\mathbf{b}} \sum_n ||\mathbf{l}_n - \mathbf{A}\mathbf{h}_n - \mathbf{b}||^2 \tag{11}$$

これを解くと以下のように回帰パラメータが得られる．

$$\begin{aligned} \mathbf{W} &= \mathbf{L}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1} \\ \mathbf{W} &= [\mathbf{b} \;\; \mathbf{A}] \\ \mathbf{L} &= [\mathbf{l}_1 \cdots \mathbf{l}_N], \;\; \mathbf{H} = \begin{bmatrix} 1 & \cdots & 1 \\ \mathbf{h}_1 & \cdots & \mathbf{h}_N \end{bmatrix} \end{aligned} \tag{12}$$

3　評価実験

3.1　実験環境

実環境下で収録したインパルス応答と，それを畳みこんだ音声信号を用いて評価実験を行った．インパルス応答の収録環境と，スピーカーの位置を Fig. 3 に示す．スピーカーの位置は水平方向，垂直方向ともに-90cm から 90cm まで 15cm 間隔でスピーカーを移動させ，計 25 箇所でインパルス応答を TSP 法 [5] により測定した．

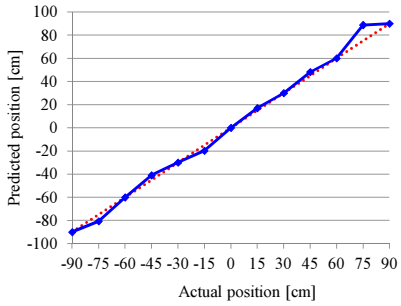

Fig. 4　インパルス応答による水平方向の回帰分析結果

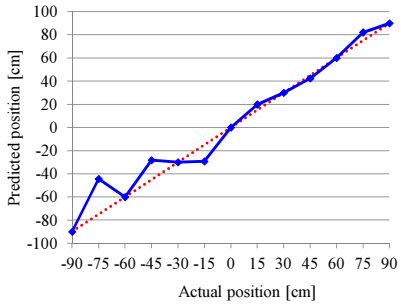

Fig. 5　インパルス応答による垂直方向の回帰分析結果

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，これに測定したインパルス応答を畳みこむことで，観測信号のシミュレーションデータを作成した．

3.2　インパルス応答を用いた予備実験

まず予備実験として，インパルス応答そのものを用いて，重回帰分析の性能を評価した．この実験では 2.2 節で述べた音響伝達特性の推定を行う必要がないため，インパルス応答の全体に矩形窓をかけて計算した 32 次元のケプストラムを特徴量として用いた．実験は水平方向のみ，垂直方向のみでそれぞれ独立に行っており，一方を評価するとき他方は 0cm に固定している．水平，垂直方向それぞれ，-90，-60，$\cdots$，60，90cm の 7 箇所のデータを用いて回帰パラメータを学習し，全 13 箇所のインパルス応答の位置を回帰パラメータにより推定する．

その際，0cm 以上と以下の二つのグループに分け，グループ毎に異なる回帰パラメータを学習することにした．そして評価の際はテストデータが 0cm 以上か 0cm 以下かは既知として，該当するグループの回帰パラメータを使用した．このようにした理由は，グループ分けを行った方が回帰分析の性能が高かったことと，以前の提案手法を用いれば 0cm 以上，以下のような大まかな識別は高精度で可能であることが分かっているためである．

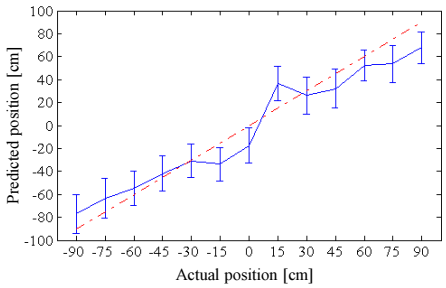

Fig. 6　音声信号から推定した音響伝達特性による水平方向の回帰分析結果

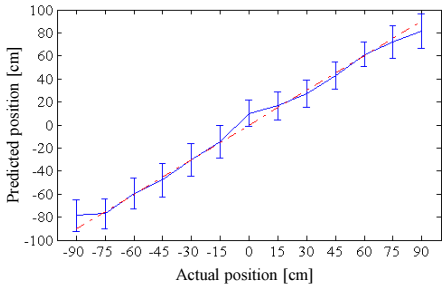

Fig. 7　音声信号から推定した音響伝達特性による垂直方向の回帰分析結果

水平方向，垂直方向それぞれの回帰分析結果を Fig. 4，Fig. 5 に示す．青の実線が推定値であり，これが実測値である赤の点線と近いほど回帰モデルが優れていることを表す．図より，垂直方向の 0cm 以下のグループのみ 15〜30cm の誤差が生じているが，それ以外についてはほぼ実際の位置通りに推定されていることが分かる．

3.3　音声信号を用いた実験

次にインパルス応答を畳みこんだ音声信号を用いて実験を行った．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．回帰パラメータの学習にはクリーン音声 HMM の学習データとは異なる 50 単語を位置毎に用いた．

予備実験と同様に 7 箇所 (×50 単語) の推定された音響伝達特性を用いて 0cm 以上，以下のグループ毎に回帰パラメータを学習し，全 13 箇所 (×50 単語) の推定された音響伝達特性の位置を推定する．

回帰分析結果の単語毎の平均及び標準偏差を，Fig. 6，Fig. 7 に示す．図より，全体的に約 15cm の偏差を持っていることが分かる．推定値の平均を表す青線と実測値を表す赤線のずれは，水平方向で平均約 12.3cm，垂直方向で約 3.6cm であった．

4　まとめと考察

本稿では，音源位置毎に異なる音響伝達特性に着目し，音声信号から音響伝達特性をクリーン音声 HMM を用いて推定し，重回帰分析により音響伝達特性から未学習の位置を推定する手法について検討を行った．

インパルス応答そのものを使った予備実験では，実測値に近い推定性能が得られていたのに対し，音声信号から推定した音響伝達特性を使った実験では，特に水平方向での誤差が大きく生じていた．また，予備実験で誤差が生じていた垂直方向の 0cm 以下は，音声信号の実験においては予備実験に比べて高い推定精度が得られていた．

これらのことから，実際のインパルス応答と，それが畳みこまれた音声信号から推定した音響伝達特性との間にはギャップが生じていると考えられる．そのため，このギャップを埋めるために音響伝達特性の正確な推定と，用いる特徴量の検討が必要であると考えられる．

今後は垂直方向・水平方向の両方を動かした場合の評価と，より少ない位置での回帰パラメータの学習，重回帰分析以外の変換方法について検討を行う．

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] B. Raj and M. V. S. Shashanka and P. Smaragdis, “Latent direchlet decomposition for single channel speaker separation,” Proc. ICASSP06, pp. 821-824, 2006.

[3] T. Nakatani and B.-H. Juang, “Speech dereverberation based on probabilistic models of source and room acoustics,” Proc. ICASSP06, pp. I-821-I-824, 2006.

[4] R. Takashima, T. Takiguchi, Y. Ariki, “HMM-based Separation of Acoustic Transfer Function for Single-channel Sound Source Localization,” ICASSP2010, pp. 2830-2833, 2010.

[5] Y. Suzuki, F. Asano, H.-Y. Kim and Toshio Sone, “An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses,” J. Acoust. Soc. Am. Vol. 97(2), pp. 1119-1123, 1995.