残響適応パラメータを用いた単一チャネル音源位置推定の検討*

☆高島遼一，滝口哲也，有木康雄 (神戸大)

1 はじめに

これまでに提案されてきた音源方向や位置の推定方法は，マイクロホンアレーにおける各観測信号の位相差を用いた手法が多く，複数のマイクロホンが必要であった [1]．単一マイクロホンで音源位置を推定することができれば，コスト削減やシステムの縮小化など様々な利点が期待できる．

我々はこれまでに位置毎に発話された残響音声から音響伝達特性を推定し、それらを判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．以前の研究ではクリーン音声を GMM (Gaussian Mixture Model) でモデル化し，それを用いて音響伝達特性パラメータを推定していたが [2]，本研究ではクリーン音声の HMM (Hidden Markov Model) により推定された音響伝達特性を用いることで音源位置推定の精度向上を図る．

2 音源位置の推定

2.1 HMM による音響伝達特性の推定

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，あらかじめこれを位置毎に学習しておけば，テストデータに対してもその音響伝達特性を判別することで音源位置を推定することができる．そのために，まず観測された信号から音響伝達特性を推定する必要がある．ある場所で発声されたクリーン音声 s は，音響伝達特性 h の影響を受ける．このとき，観測信号 o はフレーム n 毎に短時間フーリエ変換を適用することで $O(\omega;n) \approx H(\omega;n) \cdot S(\omega;n)$ と近似することができる．さらに対数をとり，逆フーリエ変換を適用することによりケプストラムが得られる．

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

ここで，d はケプストラムの次元を表す．ケプストラムは音声認識の分野で広く用いられていることから，音響伝達特性の特徴量として使用する．実際の環境では S が未知であるため，(1) 式から直接 H を求めることはできないが，S を HMM で学習しておき，事前知識として用いることで最尤推定法により O から H を推定することができる．

本手法における位置毎の音響伝達特性の推定と学習の流れを Fig. 1 に示す．あらかじめ特定話者のクリーン音声を音素 HMM でモデル化しておき，それを用いて観測信号を音素認識する．そして音素認識の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を推定する．

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|\lambda_S, H) \tag{2}$$

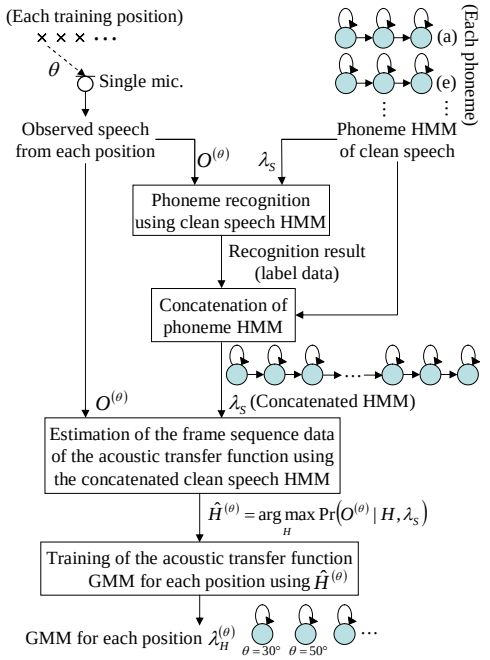

Fig. 1 音素 HMM による位置毎の音響伝達特性の推定と学習の流れ

ここで，λ_S はクリーン音声のモデルパラメータを表す．(1) 式より，O は S と H の加算とみなされるため，O の事後確率はクリーン音声 HMM を以下の式により残響適応したものとしてモデル化することができる．

$$\mu^{(O)} = \mu^{(S)} + H, \quad \Sigma^{(O)} = \Sigma^{(S)} \tag{3}$$

$\mu^{(O)}$，$\mu^{(S)}$，$\Sigma^{(O)}$，および $\Sigma^{(S)}$ はそれぞれ O と S の平均ベクトルと共分散行列 (対角行列) を表す．(2) 式の解は EM アルゴリズムにより推定される．そのとき，Q 関数は次式のように導出される [3]．

$$\begin{aligned} Q(\hat{H}|H) \;=\; & -\textstyle\sum_p \sum_j \sum_k \sum_n \gamma_{p,j,k}(n) \\ & -\textstyle\sum_{d=1}^{D} \left\{ \frac{1}{2}\log(2\pi)^D \sigma^{(S)^2}_{p,j,k,d} \right. \\ & \left. + \frac{(O(d;n)-\mu^{(S)}_{p,j,k,d}-\hat{H}(d;n))^2}{2\sigma^{(S)^2}_{p,j,k,d}} \right\} \end{aligned} \tag{4}$$

$$\gamma_{p,j,k}(n) = \Pr(O(n), p, j, k|\lambda_S) \tag{5}$$

ここで，$\mu^{(S)}_{p,j,k,d}$ と $\sigma^{(S)^2}_{p,j,k,d}$ はそれぞれ音素 p，状態 j，混合要素 k における平均ベクトルと共分散行列の対角成分の d 次元目の値を表す．この Q 関数を最大にする $\hat{H}$ は，$\hat{H}$ について偏微分して解くことにより求めることができる．

*Single-channel sound source localization using adaptation parameter for reverberant speech

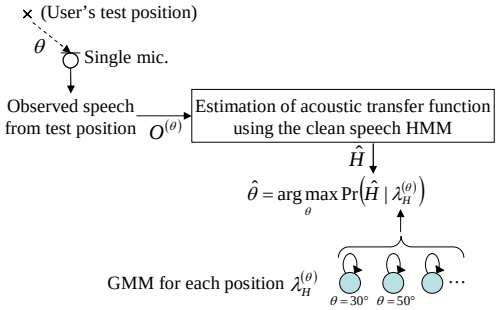

Fig. 2　音響伝達特性パラメータの判別による音源位置の推定

$$\hat{H}(d;n) = \frac{\sum_p \sum_j \sum_k \gamma_{p,j,k}(n) \frac{O(d;n) - \mu^{(S)}_{p,j,k,d}}{\sigma^{(S)2}_{p,j,k,d}}}{\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k}(n)}{\sigma^{(S)2}_{p,j,k,d}}}. \quad (6)$$

2.2　GMM による音源位置の判別

音源位置 θ 毎に観測された学習用の発話データを用いて音響伝達特性を (6) により推定し，それらを GMM でモデル化しておく．そして未知の位置で発話されたテストデータに対しても同様に音響伝達特性を推定し，学習した GMM との尤度を比較することで位置の推定を行う．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \Pr(\hat{H}|\lambda^{(\theta)}_H) \quad (7)$$

ここで，$\lambda^{(\theta)}_H$ は位置 θ に対応する音響伝達特性 GMM を表す．Fig. 2 に音響伝達特性の判別による音源位置推定の概要を示す．

2.3　評価実験

提案手法を評価するためにシミュレーション実験を行った．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 5 名の単語音声を用いてそれぞれ特定話者実験を行っている．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．クリーン音声のモデルは 2,620 単語を用いて，54 種類の音素 HMM を学習しており，各音素 HMM の状態数は 3，混合数は 32 として実験を行っている．推定された音響伝達特性の学習には 50 単語を用いて，混合数が 16 の GMM でモデル化し，1,000 単語を用いて評価を行った．なお，クリーン音声の学習データ，音響伝達特性の学習データ，評価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．音響伝達特性の学習データと評価データは，RWCP 実環境音声・音響データベースより音源とマイクロホンの距離が 2 m，残響時間が 300 msec のインパルス応答をクリーン音声に畳み込むことで作成した．音源位置は 30°，90°，130° の 3 種類の場合と，10°，50°，90°，130°，170° の 5 種類の場合で実験を行った．比較実験として，以前に提案していた 64 混合のクリーン音声 GMM を用いて音響伝達特性を推定する手法と比較し，今回提案する手法ではクリーン音声音素

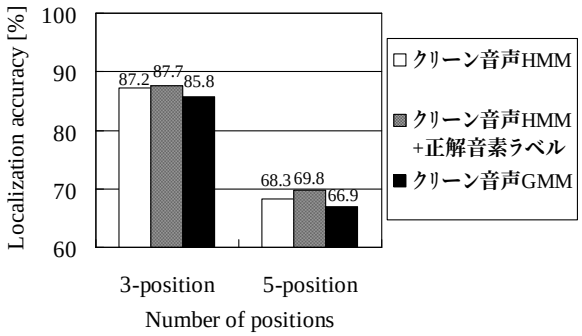

Fig. 3　各手法による音源位置の推定精度

Table 1　各手法により推定された音響伝達特性の二乗誤差

	HMM	HMM+正解音素ラベル	GMM
MSE	2096.14	1968.36	2264.33

HMM による認識結果をラベルとして用いているのに対して，正解の音素ラベルを与えた場合とも比較を行った．

各手法の音源位置の推定精度を Fig. 3 に，また各手法により推定された音響伝達特性の二乗誤差を Table. 1 に示す．二乗誤差は (1) 式に正解の S を与えて得られた H を正解の音響伝達特性として計算している．クリーン音声 GMM を用いた場合と比べてクリーン音声 HMM を用いた手法の方が音響伝達特性が精度よく推定できており，位置の推定精度も 1.4 % ほど向上している．また，正解の音素ラベルを与えることによってさらに音響伝達特性の推定精度及び音源位置の推定精度が向上している．このことから，音素認識の精度を上げることでさらに音源位置の推定精度を上げることができることが分かる．

3　おわりに

本稿では，単一マイクロホンのみによる音源位置推定の方法として，HMM で分離された音響伝達特性を用いた尤度比較による推定方法を提案した．実験により，以前に提案していたクリーン音声 GMM を用いる手法と比べて音源位置の推定精度の向上が得られた．しかし位置の数が増えると以前の手法と同様に大幅に精度が下がるため，今後は別の残響信号のモデル化や特徴量について検討する．

参考文献

[1] D. Johnson and D. Dudgeon, "Array Signal Processing," Prentice Hall, 1996.
[2] T. Takiguchi, Y. Sumida, R. Takashima, Y. Ariki, "Single-Channel Talker Localization Based on Discrimination of Acoustic Transfer Functions," EURASIP Journal on Advances in Signal Processing Vol. 2009, 9 pages, 2009.
[3] T. Takiguchi, M. Nishimura, "Improved HMM Separation for Distant-talking Speech Recognition," IEICE TRANS. INF. & SYST., VOL.E87-D, NO.5 MAY 2004.