音響伝達特性モデルを用いたシングルチャネル音源位置推定の検討*

◎高島遼一, 住田 雄司, 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位置の推定方法は，マイクロフォンアレーにおける各観測信号の時間差を用いた手法が多く，複数のマイクロフォンが必要であった．単一マイクロフォンによる音源位置推定を可能にするため，住田らはこれまでに位置毎の音響伝達特性を判別することにより音源位置を推定する方法を提案し，音源が一つの場合においてその優位性を示してきた [1].

本稿では位置毎の観測信号の音響伝達特性をあらかじめ推定，学習しておき，クリーン音声モデルとの合成により複数音源における観測信号モデルを作成し，合成モデルを用いた尤度判定により音源位置を推定する手法を提案する．

2 観測信号のモデル化

音源 i $(i = 1 \cdots M)$ が位置 θ_i から発したクリーン音声 s_i は位置毎に異なる音響伝達特性 $h_i^{\theta_i}$ の影響を受け，さらに他音源からの信号 $x_j^{\theta_j} = s_j * h_j^{\theta_j}$ $(i \neq j)$ と足しあわされて観測される．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行うことにより次のように表される．

$$O(\omega; n) = \sum_{i=1}^{M} X_i^{\theta_i}(\omega; n) \tag{1}$$

$$X_i^{\theta_i}(\omega; n) = S_i(\omega; n) \cdot H_i^{\theta_i}(\omega; n) \tag{2}$$

本手法では全ての i，θ_i についてクリーン音声と音響伝達特性をそれぞれ GMM(Gaussian mixture model) と single Gaussian model で学習しておき，モデル合成により観測信号のモデルを作成する．

2.1 音響伝達特性の推定

音源，位置毎の音響伝達特性モデルを作成するため，それぞれの音源から個別に音声を発声させ，その音声の音響伝達特性を推定し，それを用いて音響伝達特性を single Gaussian model で学習する．(2) 式はケプストラム領域では次のように表される．

$$X_{i,\mathrm{cep}}^{\theta_i}(d; n) \approx S_{i,\mathrm{cep}}(d; n) + H_{i,\mathrm{cep}}^{\theta_i}(d; n) \tag{3}$$

d はケプストラムの次元を表す．実際の環境では S_i は未知であるので (3) 式から直接 $H_i^{\theta_i}$ を求めることはできない．そこで，あらかじめ S_i の GMM を作成しておき，$X_i^{\theta_i}$ に対してその GMM の尤度が最大となるように $H_i^{\theta_i}$ を推定する．なお，簡単のため i，θ_i，cep は省略して表記する．

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(X|\lambda_S, H) \tag{4}$$

λ_S はクリーン音声のモデルパラメータであり，混合要素 k 毎に重み w_k，平均 $\mu_k^{(S)}$，共分散行列 $\Sigma_k^{(S)}$(対角成分を $\sigma_{k,d}^{(S)^2}$ とする対角行列) によって表される．EM アルゴリズムを用いて，$\hat{H}$ の更新式は次式のように導出される [1].

$$\hat{H}(d; n) = \frac{\sum_k \gamma_k(n) \frac{X(d;n) - \mu_{k,d}^{(S)}}{\sigma_{k,d}^{(S)^2}}}{\sum_k \frac{\gamma_k(n)}{\sigma_{k,d}^{(S)^2}}} \tag{5}$$

$$\gamma_k(n) = \Pr(X(n), k|\lambda_S) \tag{6}$$

2.2 モデル合成による観測信号のモデル化

次に 2.1 節で求めた $\hat{H}_i^{\theta_i}$ を用いて音源，位置毎に音響伝達特性を single Gaussian model でモデル化する．これにより求まったモデルパラメータ $\lambda_{H_i}^{\theta_i} = \left\{\mu^{(H_i^{\theta_i})}, \Sigma^{(H_i^{\theta_i})}\right\}$ とクリーン音声のモデルパラメータ λ_{S_i} を用いて観測信号モデル λ_O^{Θ} $(\Theta = \{\theta_1, \ldots, \theta_M\})$ を作成する [2]．音源数が 2 のときの観測信号のモデル化の流れを Fig. 1 に示す．

まずケプストラム領域で学習した S_i と $H_i^{\theta_i}$ のモデルパラメータを用いて残響音声モデル $\lambda_{X_i}^{\theta_i}$ を作成する．(3) 式より，残響音声モデルの平均，共分散行列はそれぞれ

$$\mu_{\mathrm{cep}}^{(X_i^{\theta_i})} = \mu_{\mathrm{cep}}^{(S_i)} + \mu_{\mathrm{cep}}^{(H_i^{\theta_i})},\ \Sigma_{\mathrm{cep}}^{(X_i^{\theta_i})} = \Sigma_{\mathrm{cep}}^{(S_i)} + \Sigma_{\mathrm{cep}}^{(H_i^{\theta_i})} \tag{7}$$

と表せる．次に各音源の残響音声モデル $\lambda_{X_i}^{\theta_i}$ を用いて観測信号モデル λ_O^{Θ} を作成する．観測信号は (1) 式よりスペクトルの加算によって表されるため，モデルパラメータをケプストラム領域からスペクトル領域に変換する必要がある．ケプストラムは逆離散コサイン変換を行うことで対数スペクトルに変換できる．

$$\mu_{\log}^{(X_i^{\theta_i})} = \Gamma^{-1} \mu_{\mathrm{cep}}^{(X_i^{\theta_i})},\ \Sigma_{\log}^{(X_i^{\theta_i})} = \Gamma^{-1} \Sigma_{\mathrm{cep}}^{(X_i^{\theta_i})} (\Gamma^{-1})^T \tag{8}$$

*Monaural sound-source-direction estimation using model of acoustic transfer function, by Ryoichi Takashima, Yuji Sumida, Tetsuya Takiguchi and Yasuo Ariki (Kobe Univ.)

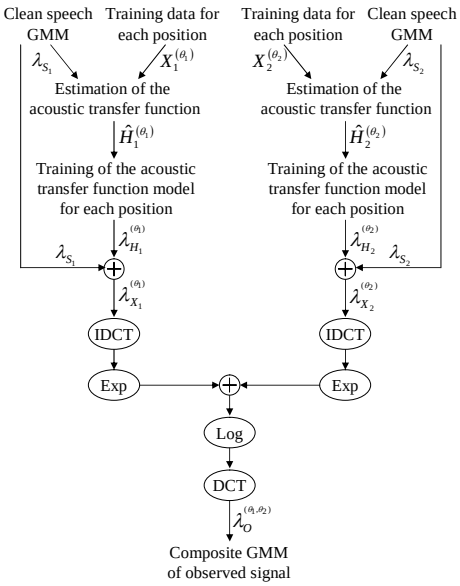

Fig. 1　観測信号のモデル化の流れ

Γ はコサイン変換の変換行列を表す．その後指数変換を行いスペクトルに変換し，加算することで観測信号モデルパラメータをスペクトル領域で求めることができる．

$$\mu_{\mathrm{lin},p}^{(X_i^{\theta_i})} = \exp\left\{\mu_{\mathrm{log},p}^{(X_i^{\theta_i})} + \sigma_{\mathrm{log},pp}^{(X_i^{\theta_i})^2}/2\right\}$$

$$\sigma_{\mathrm{lin},pq}^{(X_i^{\theta_i})^2} = \mu_{\mathrm{lin},p}^{(X_i^{\theta_i})} \cdot \mu_{\mathrm{lin},q}^{(X_i^{\theta_i})} \cdot \exp\left\{\sigma_{\mathrm{log},pq}^{(X_i^{\theta_i})^2} - 1\right\} \quad (9)$$

$$\mu_{\mathrm{lin}}^{(O^{\Theta})} = \sum_{i=1}^{M} \mu_{\mathrm{lin}}^{(X_i^{\theta_i})}, \quad \Sigma_{\mathrm{lin}}^{(O^{\Theta})} = \sum_{i=1}^{M} \Sigma_{\mathrm{lin}}^{(X_i^{\theta_i})} \quad (10)$$

ここで p, q は行列の要素インデックスを表す．そしてこれらに対数変換，離散コサイン変換を行うことでケプストラム領域に戻す．

$$\sigma_{\mathrm{log},pq}^{(O^{\Theta})^2} = \log\left\{\frac{\sigma_{\mathrm{lin},pq}^{(O^{\Theta})^2}}{\mu_{\mathrm{lin},p}^{(O^{\Theta})} \cdot \mu_{\mathrm{lin},q}^{(O^{\Theta})}} + 1\right\}$$

$$\mu_{\mathrm{log},p}^{(O^{\Theta})} = \log \mu_{\mathrm{lin},p}^{(O^{\Theta})} - \sigma_{\mathrm{log},pp}^{(O^{\Theta})^2}/2 \quad (11)$$

$$\mu_{\mathrm{cep}}^{(O^{\Theta})} = \Gamma \mu_{\mathrm{log}}^{(O^{\Theta})}, \ \Sigma_{\mathrm{cep}}^{(O^{\Theta})} = \Gamma \Sigma_{\mathrm{log}}^{(O^{\Theta})} \Gamma^T \quad (12)$$

2.3　観測信号モデルによる音源位置の判別

全ての音源と位置の組み合わせについて (12) 式により観測信号モデルを作成しておき，未知の位置から発声された音声の観測信号について，観測信号との尤度が最も高い位置の組み合わせを出力する．

$$\hat{\Theta} = \underset{\Theta}{\mathrm{argmax}} \Pr(O|\lambda_{O^{\Theta}}) \quad (13)$$

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．音源は男声，女声を各 1 音源，音源位置は音源距離約 2 m，角度 30°，90°，130° の 3 種類とした．

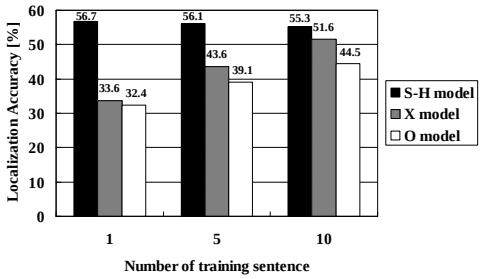

Fig. 2　両方の位置が正解した割合

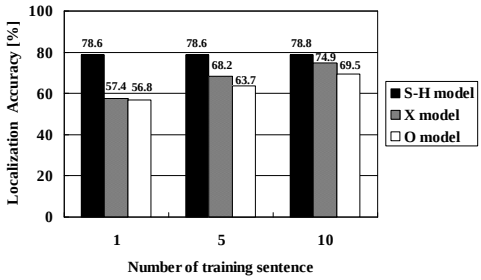

Fig. 3　少なくとも片側の位置が正解した割合

また，サンプリング周波数 12 kHz，特徴量に MFCC 16 次元を用いている．

比較実験として 2 種類の手法と比較を行った．一つは各音源の残響音声から音響伝達特性を分離せずに直接残響音声モデル $\lambda_{X_i}^{\theta_i}$ を学習により求めた手法 (X Model) で，もう一つは全ての組み合わせの混合音声から λ_O^{Θ} を直接学習した手法 (O Model) である．音源位置の学習には 10 文，5 文，1 文を用い，提案手法 (S-H Model) におけるクリーン音声モデルは 40 文を 64 混合で学習した．両方の音源位置が正解した場合と少なくとも片方の音源位置が正解した場合の結果を Fig. 2，Fig. 3 に示す．提案手法以外の方法では位置の学習に用いる文章数が少なくなるにつれて正解率が低下したが，提案手法では文章数による正解率の差はほとんど見られなかった．ゆえに，あらかじめクリーン音声モデルを作成しておけば位置の学習には少量のデータで十分であると考えられる．

4　おわりに

本稿では音源毎のクリーン音声モデルと位置毎の音響伝達特性モデルとのモデル合成により観測信号モデルを作成し，それとの尤度比較により複数音源における音源位置推定を行う手法を提案した．今後はさらに音源や位置の数を増やし実験を行う予定である．

参考文献

[1] 住田他，音講論（春），771-772，2008.
[2] Tetsuya Takiguchi, et al., IEICE Trans. INF. and SYST., vol. E89-D, 908-914, 2006.