音響伝達特性を用いた単一チャネル音源位置推定における特徴量選択の検討*

☆高島遼一，滝口哲也，有木康雄 (神戸大院)

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など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，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 [2]．本稿では，位置毎に推定された音響伝達特性の MFCC (Mel-Frequency Cepstral Coefficient) を SVM (Support Vector Machine) で識別を行うが，その際 MFCC の各次元に異なるカーネル関数を設定して独立にカーネルマトリクスを計算し，MKL (Multiple Kernel Learning) により重み付け統合を行うことで，次元毎に適した識別空間の設計及び特徴次元の自動選択を行う．

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，あらかじめこれを位置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を推定することができる．

本手法は大きく二つのステップに分けられる．まず，ある位置から発話された音声から音響伝達特性を推定する．そして，推定された音響伝達特性を用いて音源位置を学習，識別を行う．提案手法の概要を Fig. 1 に示す．

2.2 音素 HMM による音響伝達特性の推定

第一ステップでは，ある位置で発話された観測信号から音響伝達特性を推定する．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，$O_{cep}(d;n) \approx S_{cep}(d;n) + H_{cep}(d;n)$ と近似される．d はケプストラムの次元を表す．実際の環境では S が未知であり，直接 H を求めることはできないため，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．

本手法におけ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 (Hidden Markov Model) でモデル化しておき，それらを用いて観測信号を音素認識する．そ

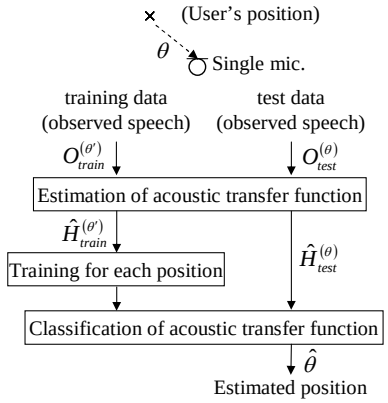

Fig. 1 提案手法の概要

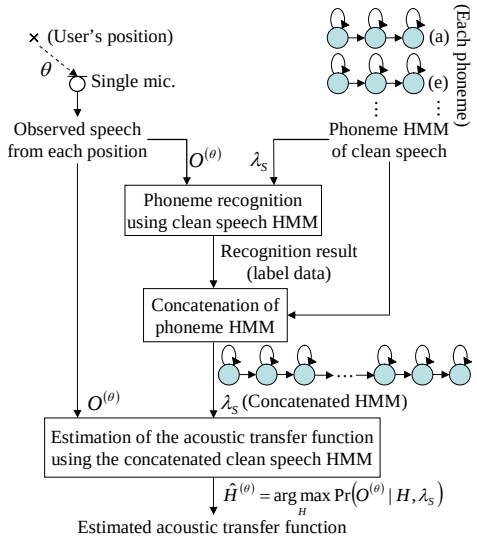

Fig. 2 音素 HMM による音響伝達特性の推定

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の MFCC を推定する．

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|\lambda_S, H) \tag{1}$$

λ_S はクリーン音声のモデルパラメータを表す．ここで，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モデル化する．

$$\mu^{(O)} = \mu^{(S)} + H, \quad \Sigma^{(O)} = \Sigma^{(S)} \tag{2}$$

$\mu^{(O)}$，$\mu^{(S)}$，$\Sigma^{(O)}$，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．このモデル化により，(1) 式の解を EM アルゴリズムを用いて推定する．詳細な更新式とその導出については [2] を参照されたい．

*Feature selection for single-channel sound source localization using the acoustic transfer function

2.3 次元毎に異なる非線形空間を用いた音響伝達特性の識別

第二ステップでは，音源位置 θ 毎に推定された音響伝達特性の MFCC を用いて，SVM で位置の学習を行う．そして，音源位置が不明な評価音声についても，その推定された音響伝達特性の MFCC を識別することで，位置の推定を行う．

通常，SVM は 1 種類のカーネル関数を用いて，特徴量を高次元空間へ射影して識別関数を設計するが，MFCC は次元無相関であるため，必ずしも一つのカーネル関数が全ての次元に対して有利に働くとは限らない．また，MFCC の中には識別に有効な次元と，そうでない次元が含まれていると考えられる．そこで本研究では，MFCC の次元毎に異なるカーネル関数を定義し，MKL で統合を行うことで，次元毎に適した識別空間の設計及び特徴次元の自動選択を行う．

MKL は，複数のサブカーネルを線形結合して新たなカーネルを作成することで，より複雑な非線形空間を作成する手法である．これを用いてサンプル i, j の音響伝達特性 H_i, H_j のカーネル関数を表現すると，

$$k(H_i, H_j) = \sum_l \beta_l k_l(H_i, H_j) \tag{3}$$

となる．β_l は l 番目のサブカーネル k_l の重みである．Varma らは，複数の特徴を用いた画像識別において，サブカーネルを特徴量ごとに定義することで，識別に適した特徴重みを MKL により学習させている [3]．本研究では，MFCC の次元毎に異なるサブカーネルを定義し，MKL により重みを学習させる．

MKL の重みの学習は，SVM の枠組みで解かれるのが一般的である．SVM の枠組みにおける最適化の双対問題を以下に示す．

$$\begin{aligned} \max_{\alpha,\beta} \quad & \textstyle\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_l \beta_l k_l(H_i, H_j) \\ s.t. \quad & \begin{cases} \textstyle\sum_i y_i \alpha_i = 0, & 0 \le \alpha_i \le C \\ \textstyle\sum_l \beta_l = 1, & \beta_l \ge 0 \end{cases} \end{aligned} \tag{4}$$

α_i はラグランジュ係数，y_i はクラスを表す変数 $(-1, 1)$,C は SVM のスラック変数である．

2.4 評価実験

提案手法を評価するために特定話者による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 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 1,000 単語を用いた．なお，クリーン音声の学習，位置の学習，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．識別手法として，16 混合の GMM(Gaussian Mixture Model)，従来の単一カーネル SVM，次元毎に同一のカーネルを定義した MKL-SVM，そして提案手法である，次元毎に異なるカーネルを定義した MKL-SVM を用いて比較を行った．SVM ベースの手法では全てガウシアンカーネルを使用し，提案手法ではガウシアンカーネルの分散値を次元毎に変えることで異なるカーネルを定義している．また各カーネルのパラメータは実験的に定めた．

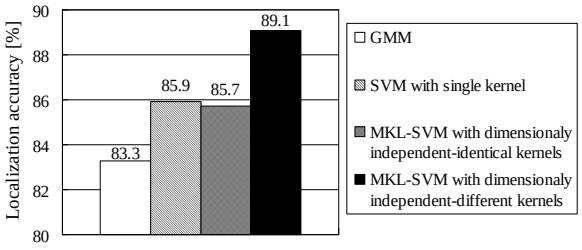

Fig. 3　各手法による音源位置の推定精度

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 種類である．

各識別手法による音源位置推定精度を Fig. 3 に示す．SVM ベースの手法はいずれも GMM による手法よりも高い識別精度を示している．また，次元毎に同一カーネルを定義した場合では，従来の単一カーネル SVM とほとんど精度は変わらなかったが，次元毎にカーネルを変えることで，それらよりも高い精度で識別できていることが分かる．

3　おわりに

本稿では，音響伝達特性の識別によるシングルチャネル音源位置推定の手法において，MFCC の次元毎に適したカーネル関数を定義して MKL で統合することで，識別精度の向上を試みた．実験では，提案手法が従来の SVM よりも高い識別精度を示したが，現状では各カーネルのパラメータを実験的に決めており，最適なパラメータを決定するのに手間がかかるという欠点がある．そのため，カーネルパラメータの自動選択などが今後の課題としてあげられる．また，学習位置から少しずれた位置での発話や，同じ位置でも発話方向が異なる場合についても評価を行っていく．

参考文献

[1] D. Johnson and D. Dudgeon, “Array Signal Processing,” Prentice Hall, 1996.

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

[3] M. Varma, D. Ray, “Learning the discriminative power-invariance trade-off,” ICCV2007, pp. 1150-1157, 2007.