音響伝達特性の判別に基づく単一チャネル音源位置推定における MKL-SVM を用いた特徴量重みの自動学習*

☆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，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，車内音声認識などの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].

そこで我々はこれまで，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，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．

提案手法の処理はすべて MFCC を特徴量として行われており，位置毎の音響伝達特性の識別は SVM (Support Vector Machine) で行われる．その際，音響伝達特性のケプストラムの各次元の中には，その位置のインパルス応答の影響を強く受ける次元と，影響を受けにくい次元が存在しており，またこのような次元毎の影響の度合いは，音源の位置によって多少のばらつきがあると考えられる．そのため，その位置の音響伝達特性の特徴をよく表すような次元重みを音源位置毎に学習させれば，位置の識別性能を向上させることができると期待される．

そこで本研究では，MKL (Multiple Kernel Learning) により，音響伝達特性 MFCC の次元重みを位置毎に学習する手法を検討する．この手法では，SVM のカーネル関数を MFCC の次元毎に定義してそれぞれ独立に計算する．そして MKL によって次元毎のカーネルの重みを学習して統合し，SVM で識別を行う．本稿では AdaBoost による特徴次元重み学習との比較を行い，提案手法の有意性を示す.

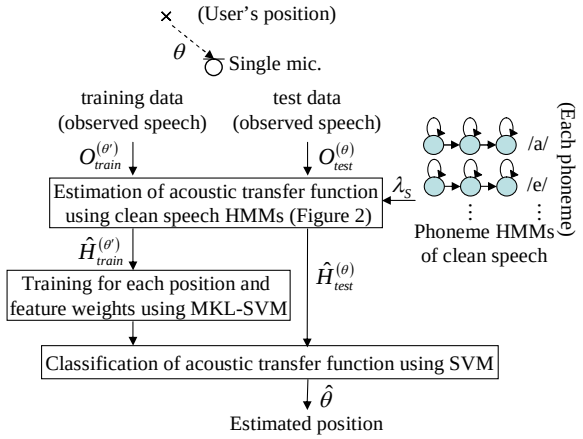

Fig. 1 提案手法の概要

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，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．

提案手法の概要を Fig. 1 に示す．まず，位置毎の音響伝達特性を学習するために，それぞれの位置 θ で発話された音声 $O^{(\theta)}_{train}$ 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，位置毎に推定された音響伝達特性のケプストラム $\hat{H}^{(\theta)}_{train}$ を MKL-SVM により学習する．この際，音響伝達特性ケプストラムの次元重みも同時に学習される．そして評価したい音声 $O^{(\theta)}_{test}$ についても学習データと同様にして音響伝達特性 $\hat{H}^{(\theta)}_{test}$ を推定し，それを SVM で識別することで，音源位置 $\hat{\theta}$ 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，

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

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

*Feature selection based on MKL-SVM for single-channel sound source localization using the acoustic transfer function. by Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

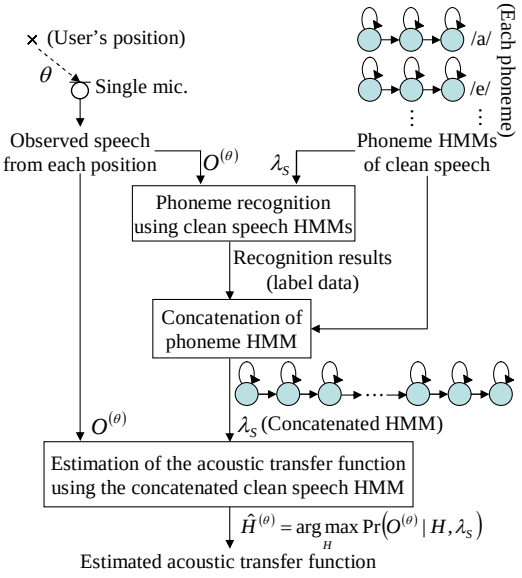

Fig. 2　音素 HMM による音響伝達特性の推定

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．

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

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

D は次元数，$\mu^{(S)}_{p,j,k,d}$，${\sigma^{(S)}_{p,j,k,d}}^2$ 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 H は，$\partial Q(\hat{H}|H)/\partial\hat{H} = 0$ を解くことで求められる．

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

2.3　MKL-SVM による音響伝達特性の次元重み学習及び識別

本節では，MKL-SVM による音響伝達特性の次元重みの学習方法と，識別方法について述べる．本稿における音源位置推定手法では，まず音源位置 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．

その際，音響伝達特性 MFCC の中にはその位置のインパルス応答の影響を強く受ける次元と，影響を受けにくい次元が存在すると考えられる．また影響を受ける次元は，音源の位置によって多少のばらつきがあると考えられる．そこで本研究では，MKL (Multiple Kernel Learning) により，音響伝達特性 MFCC の次元重みを位置毎に学習する手法を提案する．

MKL[5] 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 MFCC $\mathbf{H}_i, \mathbf{H}_j$ より計算されるカーネル関数は，以下のように表現される．

$$k(\mathbf{H}_i,\mathbf{H}_j) = \textstyle\sum_l \beta_l k_l(\mathbf{H}_i,\mathbf{H}_j) \tag{10}$$

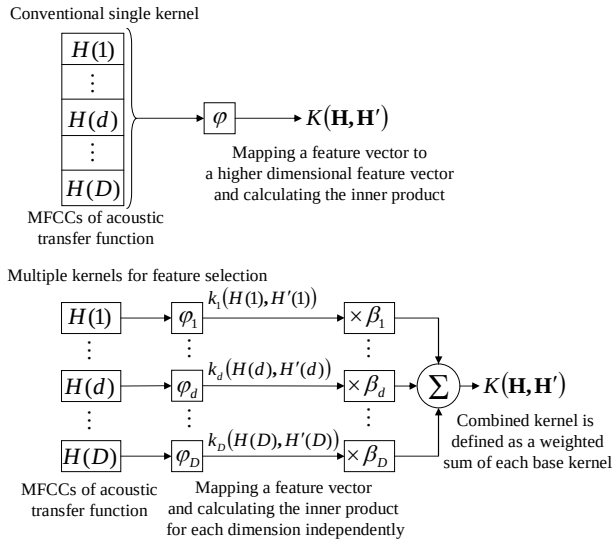

Fig. 3　従来の単一カーネル (上図) とマルチカーネルによる特徴選択 (下図)

β_l は l 番目のサブカーネル k_l の重みである．

MKL-SVM は本来，それぞれのサブカーネルを識別器とみなし，それらを統合することで，通常の SVM の識別能力を向上させることを目的として用いられているが，画像認識の分野などでは，MKL-SVM を用いて特徴選択を行う手法も提案されている [6]．この手法では，複数の特徴量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．本研究では，MFCC の次元毎にサブカーネルを定義し，MKL により重みを学習させる．

$$k(\mathbf{H}_i, \mathbf{H}_j) = \textstyle\sum_d \beta_d k_d(H_i(d), H_j(d)) \qquad (11)$$

従来の単一カーネルと，特徴選択のためのマルチカーネルの図を Fig. 3 に示す．特徴ベクトルの次元毎に独立してサブカーネルを計算させた場合，次元間の相関関係を表す情報は失われてしまう．しかし MFCC は次元の相関性が弱いため，次元毎にサブカーネルを定義しても識別能力に大きく影響はしないと考えられる．

MKL の重み β_d の学習は，SVM の枠組み，すなわちマージン最大化の枠組みで解かれるのが一般的である [5]．SVM の枠組みにおける最適化の双対問題を以下に示す．

$$\max_{\alpha,\beta} \quad \textstyle\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \sum_d \beta_d k_d(H_i(d), H_j(d))$$
$$s.t. \quad \begin{cases} \sum_i y_i\alpha_i = 0, & 0 \le \alpha_i \le C \\ \sum_d \beta_d = 1, & \beta_d \ge 0 \end{cases} \qquad (12)$$

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 はマージンと学習データの誤り率とのトレードオフを決定する変数である．(12) 式を満たす α_i，β_d は 2 ステップの反復による解法を用いて求められる．まず第一ステップでは β_d を固定して α_i を通常の SVM の解法により更新する．そして第 2 ステップでは α_i を固定して β_d を更新する．本手法では，α_i の更新には SVM^{light} を用い，β_d の更新には projected-gradient を用いた．これらのステップを繰り返すことにより，特徴次元の重みと，識別境界が同時に学習される．

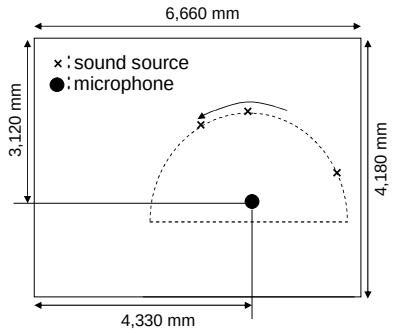

Fig. 4　実験環境

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．

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 [7] 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．Fig. 4 にインパルス応答の収録環境を示す．

3.2　実験結果

音響伝達特性の識別手法として，従来の単一カーネル SVM と提案手法である MKL-SVM を用いて比較を行い，また次元重みの学習法として，AdaBoost による特徴選択手法 [8] との比較も行った．この手法は，特徴量毎に学習させた Decision Stump を弱識別器とし，AdaBoost によって学習される各弱識別器の貢献度を用いて特徴選択を行う手法である．文献 [8] では弱識別器の貢献度が低い特徴量を取り除くことで特徴選択を行っているが，本稿では MFCC の次元毎に得られる貢献度を次元重みとして用い，SVM により識別を行っている．各手法におけるカーネル関数は Gaussian kernel を用い，(12) 式における C の値は 1 とした．

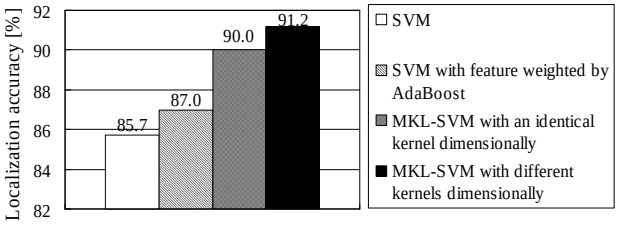

Fig. 5　識別手法毎の音源位置推定精度

MKL-SVM を用いる提案手法では，次元毎に定義する Gaussian kernel のパラメータを同一のものとした場合と，次元毎に異なるパラメータを用いた場合の 2 種類で実験を行った．これは，特徴ベクトルの MFCC が次元無相関であるため，識別に最適なカーネルのパラメータは次元によって異なるかもしれないと考えたためである．これら 2 種類の提案手法，及び従来の SVM のカーネルのパラメータは実験的に定めた．

各識別手法による音源位置推定精度を Fig. 5 に示す．次元重みを学習させることで，従来の単一カーネル SVM の識別精度を向上させることができているが，MKL-SVM を用いる提案手法では AdaBoost による次元重み学習法よりも高い識別精度が得られている．さらに次元毎に異なるカーネルのパラメータを設定することで，次元毎に同一のカーネルを定義した場合に対して若干の識別性能の向上が見られた．

提案手法では，SVM を用いてマルチクラスの識別を行うために，one-vs-rest 法を用いてクラス (位置) 毎に識別境界を学習している．一方，識別境界を学習する度に特徴次元の重みも学習されるため，結果として識別境界と同じ数の種類だけ特徴次元の重みが得られる．これは，音源位置毎に最適な次元重みを学習していることを意味している．

Tbl. 1 は次元毎に同一のカーネルパラメータを設定した提案手法を用いて得られた，音源位置毎の次元重みを表している．また Fig. 6 は，(2) 式へ実際のクリーン音声 $S_{cep}(d;n)$ を代入して得られた $H_{cep}(d;n)$ の単語毎のフレーム平均値を，位置毎にプロットした図である．これらの表と図を見ると，ある位置において高い重みが得られている次元では，音響伝達特性がその位置を識別しやすい分布をしていることが分かる．例えば Tbl. 1 では，90° において 7 次元目が最も高い重みを得ており，一方 Fig. 6 では，7 次元目が 90° の音響伝達特性を判別されやすい分布になっている．同様に 30° では 10 次元目が，130° では 8 次元目が高い重みを得ており，それぞれの位置を判別しやすい分布になっている．一方，1 次元目はいずれの位置においても重みがほぼ 0 となっており，Fig. 6 を見ると 1 次元目の値はほとんど位置の違いの影響が現れていないことが分かる．これらのことから，それぞれの位置において，その位置の音響伝達特性を判別しやすい次元に対する重みが，MKL によって学習できていることが分かる．

Table 1　MKL によって学習された位置毎の次元重みの例．**太字**はその位置において最も高い重みを表す．

	1^{st}	4^{th}	7^{th}	8^{th}	10^{th}
30 degrees	0.00	0.07	0.07	0.07	**0.08**
90 degrees	0.00	0.06	**0.10**	0.07	0.07
130 degrees	0.01	0.07	0.06	**0.11**	0.07

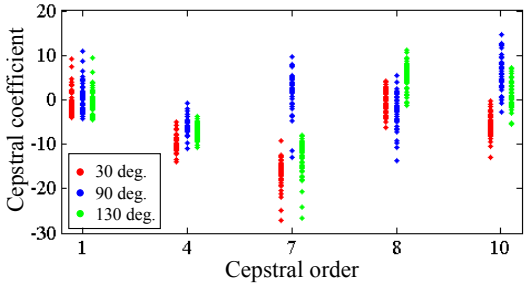

Fig. 6　特定の次元における音響伝達特性の分布

4　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 の次元毎にカーネル関数を定義することにより，特徴次元の重みを MKL により学習させた．提案手法では音源位置毎に，異なる次元重みのセットを学習することができ，従来の単一カーネル SVM よりも高い識別精度を得ることができた．今後は雑音環境下や，実環境下での識別性能の評価や，音響伝達特性の推定方法についても検討していく．

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] A. Rakotomamonjy, F. Bach, S. Canu and Y. Grandvalet, “More Efficiency in Multiple Kernel Learning,” Proc. ICMKL, pp. 775-782, 2007.
[6] M. Varma, D. Ray, “Learning the discriminative power-invariance trade-off,” Proc. ICCV2007, pp. 1150-1157, 2007.
[7] S. Nakamura, “Acoustic sound database collected for hands-free speech recognition and sound scene understanding,” International Workshop on Hands-Free Speech Communication, pp. 43-46, 2001.
[8] 土屋成光, 藤吉弘亘, “Boost 学習に基づく特徴量の貢献度を用いた特徴選択手法,” 画像の認識・理解シンポジウム, MIRU2008, IS2-1, pp. 599-604, 2008.