対判別フィッシャー重みマップを利用した局所特徴量による音素認識*

◎加藤 俊祐, 滝口 哲也, 有木 康雄 (神戸大・工)

1 はじめに

本稿では、局所特徴を用いたフィッシャー判別基準による音声特徴量抽出手法について検討を行う。これらの手法は、画像の分野では様々な画像に対して有効性が示されてきている [1]。本研究では、短時間フーリエ変換後の時間-周波数平面上において局所特徴を求め、さらに重みマップとの積をとり特徴ベクトルを求めた。重みマップは、認識のために重要な特徴を含んでいる領域に高い重み付けがなされるように、フィッシャーの判別基準を利用して求める。さらに特定のクラスに重みが偏るのを防ぐために、対ごとに重みマップを求め識別を行なう対判別でのフィッシャー判別基準を利用した [2]。

2 局所特徴量とフィッシャー重みマップ

2.1 局所特徴量

時刻 t、周波数 f のパワースペクトルを $I(r)$ とすると、点 r（時間と周波数を表す 2 次元ベクトル）における k 番目の局所特徴量は次式で表される。

$$h_k(r) = I(r)I(r + a_1^{(k)}) \cdots I(r + a_N^{(k)}) \tag{1}$$

変位を参照点 r の近傍 3×3 の局所領域に限定し、さらに次数 N を高々2 までに制限すると、局所パターンの変位（$a_1, \cdots, a_N$）の種類は平行移動により等価なものを除くと全部で 35 種類になる。図 1 に局所パターンの一部を示す。局所パターン中の 1 に対応するパワースペクトル値を積和することにより、各々の局所パターンに対応する局所特徴量が得られる。（但し、図中の 2、3 は、対応するパワースペクトルの二乗、三乗を意味する。）ここで、ある音素に対する時間-周波数平面上の全ての点 r（$M = T$（時間方向の総数）$\times F$（周波数方向の総数））における k 番目の局所パターンを以下のように M 次元ベクトルで表記する。

$$\mathbf{h}_k = [h_k(1,1) \cdots h_k(1,T), \cdots h_k(F,T)]^t \tag{2}$$

さらに、局所パターンの総数を K 種類（今回は $K = 35$）として、$\mathbf{h}_k$ を横に並べたものを

$$\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_K] \tag{3}$$

とし、これを局所特徴量 $\mathbf{H}$ とする。

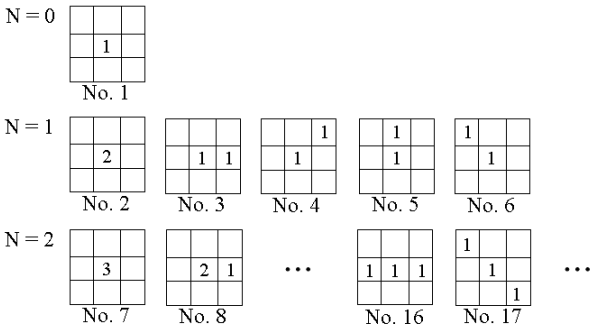

Fig. 1 局所パターンの例

2.2 フィッシャー重みマップ

認識のために重要な特徴を含んでいる領域に高い重み付けをしながら特徴抽出が行われるように、最適な重みマップを決定する。本稿では、フィッシャーの判別基準を利用する [1]。

N 個の学習データがあるとする。各データに対応する局所パターン行列を $\{\mathbf{H}_i \in R^{M \times K}\}_{i=1}^N$、特徴ベクトルを $\{\mathbf{x}_i = \mathbf{H}_i^t \mathbf{w} : \mathbf{w}$ は重みベクトル $\}_{i=1}^N$、クラス内分散行列を $\tilde{\Sigma}_W$、クラス間共分散行列を $\tilde{\Sigma}_B$ で表すと、次式が得られる。

$$\begin{aligned} tr\tilde{\Sigma}_W &= \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in\omega_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)})^t(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)}) \\ &= \mathbf{w}^t\left\{\frac{1}{N}\sum_{j=1}^{J}\sum_{i\in\omega_j}(\mathbf{H}_i^{(j)} - \bar{\mathbf{H}}^{(j)}) \right. \\ &\qquad \left. (\mathbf{H}_i^{(j)} - \bar{\mathbf{H}}^{(j)})^t\right\}\mathbf{w} \\ &= \mathbf{w}^t\Sigma_W\mathbf{w} \end{aligned} \tag{4}$$

$$\begin{aligned} tr\tilde{\Sigma}_B &= \frac{1}{N}\sum_{j=1}^{J}N_j(\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}})^t(\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}}) \\ &= \mathbf{w}^t\left\{\frac{1}{N}\sum_{j=1}^{J}N_j(\bar{\mathbf{H}}^{(j)} - \bar{\mathbf{H}}) \right. \\ &\qquad \left. (\bar{\mathbf{H}}^{(j)} - \bar{\mathbf{H}})^t\right\}\mathbf{w} \\ &= \mathbf{w}^t\Sigma_B\mathbf{w} \end{aligned} \tag{5}$$

ここで、J はクラス数（音素数）、ω_j は j 番目のクラス、N_j はクラス ω_j に属するサンプル数、$\bar{\mathbf{x}}^{(j)}$ はクラス ω_j に属する $\mathbf{x}_i^{(j)}$ の平均、$\bar{\mathbf{x}}$ は $\mathbf{x}_i$ の全平均である。従って、フィッシャーの判別基準は、

$$J(\mathbf{w}) = \frac{tr\tilde{\Sigma}_B}{tr\tilde{\Sigma}_W} = \frac{\mathbf{w}^t\Sigma_B\mathbf{w}}{\mathbf{w}^t\Sigma_W\mathbf{w}} \tag{6}$$

となる。このフィッシャー判別基準を制約条件 $\mathbf{w}^t\Sigma_W\mathbf{w} = 1$ の下で最大化する重み $\mathbf{w}$ は固有値問題

$$\Sigma_B\mathbf{w} = \lambda\Sigma_W\mathbf{w} \tag{7}$$

の固有ベクトルとして求められる。このようにして得られる最適重みベクトルをフィッシャー重みマップと呼ぶ。最終的に、式 (8) のように上位 c 個の固有ベクトルである重みマップを並べ、局所特徴量 $\mathbf{H}$ との積 $\mathbf{X}$ を音声特徴量として識別を行なう。

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1 \cdots \mathbf{x}_c] \\ &= \mathbf{H}^T[\mathbf{w}_1 \cdots \mathbf{w}_c] \\ &= \mathbf{H}^T\mathbf{W} \end{aligned} \tag{8}$$

---

*Phoneme Recognition by Local Features Using Pairwise Discriminant Fisher-Weight-Map. by Shunsuke Kato, Tetsuya Takiguchi and Yasuo Ariki (Kobe University)

3　対判別

2.2 の手法で重み $\mathbf{W}$ を求めると、重みがある特定のクラス（音素）に偏り、ある特定のクラスの識別率が高くなり、別のクラスの識別率は低くなる現象が起こる。本研究ではこれを改善するために、対判別の手法を用いた [2]。

対判別ではまず学習データを用いて、クラスの対ごとに重み $\mathbf{W_{ij}}$ を求め、式 (8) により学習データの音声特徴量 $\mathbf{X_{ij}}$ を求める。この音声特徴量をもとに、クラス i とクラス j の混合分布モデル (GMM) を求めておく。識別では、与えられたパターンベクトル $\mathbf{I}$ に対して式 (8) により音声特徴量 $\mathbf{X_{ij}}$ を求め、クラス i とクラス j における GMM の事後確率の比、$P_{ij}^{(i)}(\mathbf{I})$、$P_{ij}^{(j)}(\mathbf{I})$ を求める。次にクラス i を固定し、クラス j を変えて $P_i(\mathbf{I}) = \min_j P_{ij}^{(i)}(\mathbf{I}) \quad (i \neq j)$ を求め、$\arg\max_i P_i(\mathbf{I})$ を識別されたクラスとする。

図 2 に 4 クラスによる対判別の識別例を示す。クラス A と B の対に対して計算された重み $\mathbf{W_{AB}}$ と局所特徴量の積を $\mathbf{X_{AB}}$ とし、$\mathbf{X_{AB}}$ によるクラス A の尤度とクラス B の尤度の事後確率の比が 0.1 と 0.9 となる。これをそれぞれの対ごとに算出し、各クラスごとに最小値を求める（クラス A の場合は 0.1）。各クラスの最小値の中で最大値を持つクラス（この場合は 0.7 である B）が識別されたクラスとなる。

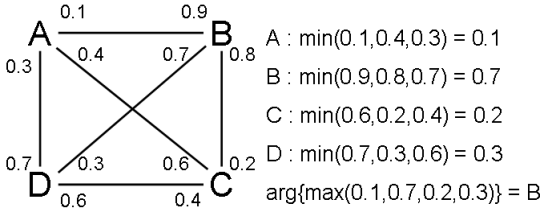

Fig. 2　対判別による 4 クラスの識別例

4　認識実験

4.1　実験条件

評価実験には対判別でない通常の群判別による実験を、5 母音のみと全音素（27 音素）の 2 つのデータセットに対して実験を行った。対判別の実験では、破裂音である p,b,t,d,k,g の 6 音素に対して実験を行った。評価試料には、同一話者が発声したラベル付きの連続音声データセット中から切り出した音素データを使用する。群判別による 5 母音のみの実験、対判別による 6 音素の実験では、各音素学習用に 100 個のデータ、評価用には学習で使用していない 100 個のデータを使用した。群判別による全音素の実験では、学習用に総数 2578 個のデータを使用し、評価用には学習で使用していない 2578 個のデータを使用した。音声データは全て、窓幅 25 ms、シフト幅 10 ms で分析を行い時間-メル周波数領域に変換。局所特徴量 $\mathbf{H}$ は、群判別の実験ではそれぞれ 3,5,7 フレーム（式 (2) において $T = 3, 5, 7$）、対判別では 5 フレームのみで切り出し実験した。識別には全て GMM を使用。

4.2　実験結果

4.2.1　群判別の実験結果

5 母音の認識率は、切り出し幅 5、シフト幅 2、重み $\mathbf{w}$ の軸の本数を 5 としたとき（式 (8) において $c = 5$。35 局所パターン $\times$ 5 = 175 次元）、98.6%と最高になった。MFCC12 次元を用いた認識率は 95.8%となった。従って、MFCC より高い性能が得られていることが分かる。

次に、表 1 に全音素に対する認識結果を示す。全音素の場合は、重み $\mathbf{w}$ の軸の本数を 20、切り出し幅 7、シフト幅 2 とした時に認識率は 81.0%と最高になった。MFCC の認識率は 84.6%であり、比較するとフィッシャー重みマップによる手法の方が低い認識率となっている。この原因の一つとして、フィッシャー重みマップを導出する際、重みがある音素に偏るため認識率が低下することが原因と考えられる。通常、判別分析ではクラス数の増加に伴い識別精度が低下することがある。この問題解決する方法として、対判別等を用いる必要がある。表 1 より切出し幅は 5〜7、シフト幅は切出し幅の半分が妥当なサイズである。

Table 1　全音素の認識率（重み $\mathbf{w}$ の軸の本数 20）（正答率%）

切出し幅	シフト幅						
	1	2	3	4	5	6	7
3	78.5	76.2	73.9				
5	80.7	79.8	79.7	79.0	76.8		
7	78.9	81.0	80.7	80.3	80.4	80.4	75.4

MFCC(12 次元) の認識率：84.6%

4.2.2　対判別の実験結果

p,b,t,d,k,g の 6 音素による対判別による識別率を図 3 に示す。最高で対判別は 92.7%、群判別では 89.9%、12 次元の MFCC では 81.4%となり、対判別は群判別より高い効果が得られることがわかる。一般的に p,b,t,d,k,g の破裂音は特徴が似ており、高い識別率を得ることが難しい。実際に音素別の識別率を見ると、群判別では識別率に差があるが、対判別では差があまりないという結果が得られた。この結果から、今後、全音素に対する対判別により識別率の改善が期待される。

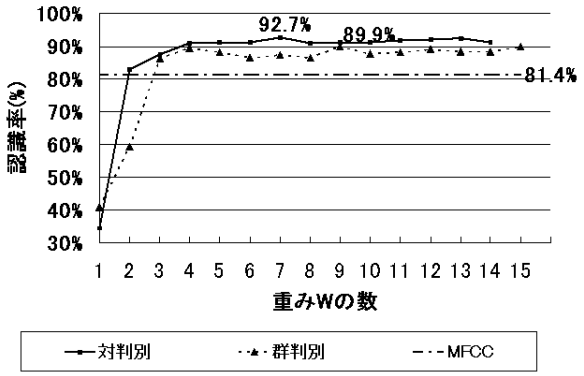

Fig. 3　対判別による p,b,t,d,k,g6 音素の識別率（切出し幅 5、シフト幅 2）

5　まとめ

本稿では、局所特徴を用いたフィッシャー重みマップによる音声特徴量抽出手法による音素認識について報告した。さらに対判別による認識についても報告した。今後は、全音素に対する対判別に拡張していく予定であるが、クラス数が増加すると対の数が膨大になるため、音素をいくつかのグループごとに分け、グループ内で対判別を行なう手法などを検討している。

参考文献

[1] 篠原雄介, 大津展之,“フィッシャー重みマップを用いた顔画像からの表情認識,” 信学技報, PRMU2003-269, Vol.103, No.737, pp.79-84, 2004.

[2] 河原達也, 堂下修司, “対判別に基づく連続型 HMM による認識,” 電子情報通信学会 D-Ⅱ, Vol.J75-D-Ihspace-.1emI, No.10, pp.1641-1648, 1992.

[3] 豊田 崇弘, 長谷川 修, “テクスチャ識別のためのマスクパターンによる特徴抽出法 ,” 電子情報通信学会 PRMU 研究会, PRMU2004-63, pp.77-84, 2004.