フィッシャー重みマップを利用した高次局所自己相関特徴による音素認識*

◎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、高次局所自己相関特徴を用いた音声特徴量抽出手法について検討を行う。これまで画像の分野では、高次局所自己相関特徴は様々な画像の認識に対して有効性が示されてきている [1]。また音声特徴量抽出においても局所特徴に注目した論文が報告されている [2]。本研究では、短時間フーリエ変換後の時間-周波数平面上にて、局所特徴を重みマップを用いて積分し、特徴ベクトル（高次局所自己相関特徴）を求める。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た。

2 局所特徴量と高次局所自己相関特徴

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。

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

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。）

2.2 高次局所自己相関特徴

高次局所自己相関特徴は局所特徴量を時間-周波数平面上にわたって積分したものである。従って、高次局所自己相関特徴 x_k は次式で表される。

$$\begin{aligned} x_k &= \int h_k(r)dr \\ &= \int I(r)I(r + a_1^{(k)}) \cdots I(r + a_N^{(k)})dr \end{aligned} \qquad (2)$$

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。

$$\mathbf{h}_k = [h_k(1,1) \cdots h_k(1,T), \cdots h_k(F,T)]^t \qquad (3)$$

すると、k 番目の局所パターンの高次局所自己相関特徴 x_k は、

$$x_k = \mathbf{h}_k^t \mathbf{1} \qquad (4)$$

となる。さらに、局所パターンの総数を K 種類として、$\mathbf{h}_k$ を横に並べたものを

$$\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_K] \qquad (5)$$

とし、高次局所自己相関特徴 x_k の K 個の局所パターンを縦に並べたベクトルを

$$\mathbf{x} = [x_1 \cdots x_K]^t \qquad (6)$$

とすると、$\mathbf{H}$ と $\mathbf{x}$ の間に以下のような関係式が成立する。

$$\mathbf{x} = \mathbf{H}^t \mathbf{1} \qquad (7)$$

$M \times K$ の行列 $\mathbf{H}$ は、非常に高次元なため、M 次元ベクトル $\mathbf{w}$ により次元削減を行う。

$$\mathbf{x} = \mathbf{H}^t \mathbf{w} \qquad (8)$$

$\mathbf{w}$ は各点毎の重み付けを行う事になるので、この $\mathbf{w}$ を重みマップと呼ぶ。

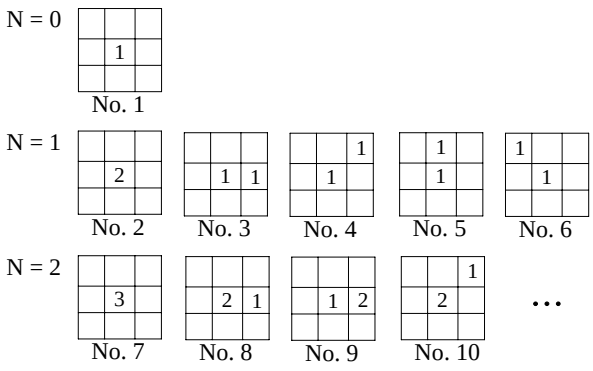

Fig. 1 局所パターン例

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

3　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 $\{\mathbf{x}_i\}_{i=1}^{N}$、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。

$$\begin{aligned} tr\tilde{\Sigma}_W &= \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in\omega_j}(\mathbf{x}_i^{(j)}-\bar{\mathbf{x}}^{(j)})^t(\mathbf{x}_i^{(j)}-\bar{\mathbf{x}}^{(j)}) \\ &= \mathbf{w}^t\left\{\frac{1}{N}\sum_{j=1}^{J}\sum_{i\in\omega_j}(\mathbf{H}_i^{(j)}-\bar{\mathbf{H}}^{(j)})\right. \\ &\qquad \left.(\mathbf{H}_i^{(j)}-\bar{\mathbf{H}}^{(j)})^t\right\}\mathbf{w} \\ &= \mathbf{w}^t\Sigma_W\mathbf{w} \end{aligned} \tag{9}$$

$$\begin{aligned} tr\tilde{\Sigma}_B &= \frac{1}{N}\sum_{j=1}^{J}N_j(\bar{\mathbf{x}}^{(j)}-\bar{\mathbf{x}})^t(\bar{\mathbf{x}}^{(j)}-\bar{\mathbf{x}}) \\ &= \mathbf{w}^t\left\{\frac{1}{N}\sum_{j=1}^{J}N_j(\bar{\mathbf{H}}^{(j)}-\bar{\mathbf{H}})\right. \\ &\qquad \left.(\bar{\mathbf{H}}^{(j)}-\bar{\mathbf{H}})^t\right\}\mathbf{w} \\ &= \mathbf{w}^t\Sigma_B\mathbf{w} \end{aligned} \tag{10}$$

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、

$$J(\mathbf{w}) = \frac{tr\tilde{\Sigma}_B}{tr\tilde{\Sigma}_W} = \frac{\mathbf{w}^t\Sigma_B\mathbf{w}}{\mathbf{w}^t\Sigma_W\mathbf{w}} \tag{11}$$

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

$$\Sigma_B\mathbf{w} = \lambda\Sigma_W\mathbf{w} \tag{12}$$

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マップをフィッシャー重みマップと呼ぶ。

4　認識実験

4.1　実験条件

評価実験には、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。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。5 母音のみのセットでは、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。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、窓幅 25 ms、シフト幅 10 ms で分析を行い時間-周波数領域に変換される。局所特徴量 H は、本実験では 5 フレームで切り出しを行い計算した（式 (3) において $T=5$）。識別には GMM を使用した。

Table 1　音素認識率

	母音	全音素
提案手法	98.6%	79.9%
MFCC	95.8%	84.6%

4.2　実験結果

表 1 に認識結果を示す。5 母音の認識率は、重み $\mathbf{w}$ の軸の本数を 3 とした時（35 局所パターン $\times$ 3 = 105 次元）98.6%となった。MFCC12 次元と比較すると同等の性能が得られているのが分かる。一方、全音素の場合は、重み $\mathbf{w}$ の軸の本数を 20 とした時に認識率は 79.9%となり、MFCC と比較して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だと考えられる。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事があり、対判別等を用いる必要がある [3]。

5　おわりに

本稿では、最適化されたフィッシャー重みマップを用いた高次局所自己相関特徴による音素認識について報告した。クラス数の増加に伴い判別分析では分離性能の劣化が大きくなるため、今後は、本手法を対判別へと拡張していく予定である。

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] 新田恒雄, 井上雄, 正井康之, 松浦博, “複合音響特徴平面に基づく音声認識のための局所特徴抽出法,” 電子情報通信学会論文誌, D-II, Vol. J83, No. 11, pp. 2341-2349, 2000。

[3] 河原達也, 堂下修司, “対判別に基づく連続型 HMM による音声認識,” 電子情報通信学会論文誌, D-II, Vol. J75, No. 10, pp. 1641-1648, 1992。