フィッシャー重みマップに基づく音声特徴量のロバストネスに関する考察*

◎室井貴司, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

本稿では，時間-周波数平面上において，高次局所自己相関特徴に対しフィッシャー判別手法を用いる音声特徴量抽出手法 [1] について検討を行う。これらの手法は画像処理・認識の分野においては有効性が示されている [2]。本研究では，短時間フーリエ変換後の時間-周波数平面上において局所特徴量を求め，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，認識においてクラスの分離が良い特徴空間を構築するため，フィッシャーの判別基準を利用している。また，提案手法の雑音に対する頑健性を検証するため，白色雑音を重畳した音声に対して認識実験を行った。

2 局所特徴量

2.1 局所パターンと局所特徴行列

時間-周波数平面の時刻 t，周波数 f の点 $r(t,f)$ のパワースペクトルを $I(r)$ とすると，点 r における k 番目の局所特徴量 $h_r^{(k)}$ は次式で表される。

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

文献 [2] では，高次局所自己相関として積を用いているが，本研究では，実験的に精度が良かったため和を用いている。次数 N を高々2 までとし，変位を参照点 r の 3×3 の局所領域に限定すると，局所パターンの変位（a_1，…，a_N）は平行移動により等価なものを除くと 35 種類得られる。局所パターンの対応する点のパワースペクトル値を加算することにより，各々の局所パターンに対応する局所特徴量が得られる。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。

$$h^{(k)} = \left[h_{2,\,2}^{(k)} \ldots\ldots\ldots h_{F-1,\,T-1}^{(k)}\right] \quad (2)$$

さらに，局所パターンの総数を K 種類（今回は $K = 35$）として，h_k を横に並べたものを

$$H = [h_1 \ldots\ldots h_K] \quad (3)$$

とし，これを局所特徴行列 H とする。

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，フィッシャーの判別基準を利用して最適な重みマップを決定する。学習データに対して，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，フィッシャーの判別基準は，

$$J(\mathbf{w}) = \frac{tr\tilde{\Sigma}_B}{tr\tilde{\Sigma}_W} \quad (4)$$

と表せる。制約条件 $\mathrm{w}\tilde{\Sigma}_W\mathrm{w} = 1$ の下で $J(\mathbf{w})$ を最大化する重み w は固有値問題

$$\tilde{\Sigma_B}\mathbf{w} = \lambda\tilde{\Sigma}_W\mathbf{w} \quad (5)$$

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 W をフィッシャー重みマップと呼ぶ。この W と局所特徴行列 H との積 X，

$$X = H^t W \quad (6)$$

を音声特徴量として識別を行う。

3 認識実験

3.1 実験条件

評価実験データは ATR の音素バランス文 B セットの男性話者 5 名，女性話者 4 名の音声に SNR=∞, 20，10，5，0dB の白色雑音を重畳させ，各話者のデータを音素ごとに切出し，音素認識の実験を行なった。音素は全部で 25 音素，各話者の学習用音声データは全音素合わせて 2578 個，評価用音声データは，学習で使用していない 2578 個のデータを使用した。音声信号の標本化周波数は 20KHz，フレーム幅は 25ms，シフト幅は 10ms であり，時間-周波数平面上でのフレーム幅は 5 フレーム，シフト幅は 1 フレームである。認識には GMM を使用した。フィッシャー重みマップ W の本数は 35 本とした。

3.1.1 音声特徴ベクトルの次元圧縮

音声特徴量 X の次元数は，35（フィッシャー重みマップ W の本数）× 35（局所パターンの数）=1225 次元であり，高次元であることから，GMM の確率推定に問題が生じる可能性がある。そこで，音声特徴量 X を主成分分析（PCA）により圧縮し，音素認識実験を行った。

*Study on Phoneme Recognition of Noisy Speech Based on Fisher Weight Map by MUROI, Takashi, TAKIGUCHI, Testuya, ARIKI, Yasuo(Kobe University)

Table 1　単一の特徴量による認識率 (%)

SNR	FWM100 次元	MFCC	ΔMFCC	ΔΔMFCC
∞	81.1	71.8	76.0	76.9
20dB	74.6	67.2	69.6	66.3
10dB	62.6	55.1	57.9	52.3
5dB	53.3	45.3	47.5	43.2
0dB	41.4	35.1	35.4	32.6

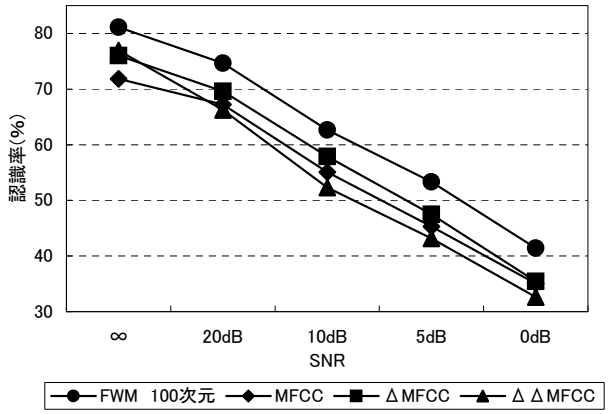

Fig. 1　単一特徴量の認識率の推移

3.2　実験結果

3.2.1　単一の特徴量による認識結果

フィッシャー重みマップ（FWM）を PCA により 100 次元に圧縮した特徴量と MFCC の特徴量を用いて実験を行った。ここで，FWM は時間-周波数平面上で 3 × 3 近傍の局所パターンを 1 フレームずつずらしているため，5 フレーム分の情報を持っていることになる。そのため，単一フレームの情報のみを持つ MFCC は比較対象として適切でないと考えられる。そこで比較対象として，MFCC の他に複数フレームの情報を持つ ΔMFCC，ΔΔMFCC を用いた。

実験結果を表 1，図 1 に示す。これらより，SNR によらず FWM で最も良い認識率が得られた。

3.2.2　MFCC と組み合わせた認識結果

FWM を 50 次元に PCA により次元削減した特徴量と MFCC，ΔMFCC，ΔΔMFCC の 4 つの特徴量を組み合わせて音素認識実験を行った。

実験結果を表 2，図 2 に示す。FWM に MFCC と ΔMFCC を組み合わせることで，MFCC と ΔMFCC を組み合わせた特徴量よりも 1 %程度高い認識率が得られた。しかし，MFCC と ΔMFCC，ΔΔMFCC を組み合わせた特徴量の結果を超えることは出来なかった。この理由として，FWM を PCA で次元削減することで，特徴量が特定の音素の特徴に偏ってしまっていることなどが考えられる。

Table 2　組み合わせた特徴量による認識率（%）

SNR	FWM+MFCC +ΔMFCC	MFCC +ΔMFCC	MFCC+ΔMFCC +ΔΔMFCC
∞	87.3	87.2	90.1
20dB	80.9	79.2	83.0
10dB	69.8	68.5	72.0
5dB	60.5	58.8	62.1
0dB	47.4	47.3	49.7

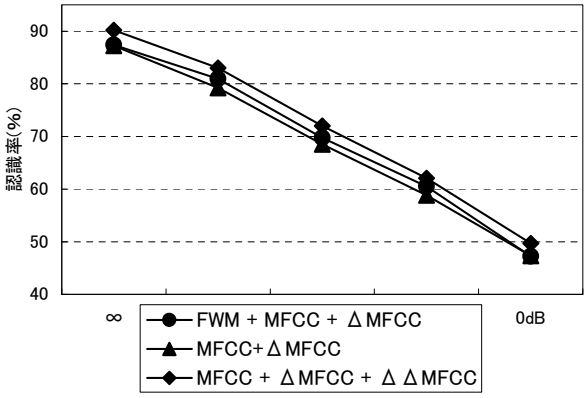

Fig. 2　組み合わせた特徴量の認識率の推移

4　まとめ

本稿では，局所特徴量とフィッシャー重みマップ（FWM）に基づく音声特徴量の音素認識について報告した。局所特徴より得られる音声特徴量を PCA により次元を削減することで，音素認識実験において MFCC の特徴量と比べ，高い認識精度を示すことが出来た。しかし，複数の特徴量を組み合わせた実験では，MFCC と ΔMFCC，ΔΔMFCC を組み合わせた特徴量の方が良い結果となった。また，音声と白色雑音の SN 比に関わらず FWM を用いた特徴量の認識率と MFCC の特徴量の認識率の優位さは一定の傾向を示した。今後は，FWM が音素の特徴をより良く表すような次元削減方法を取り入れていく予定である。

参考文献

[1] 加藤俊祐，滝口哲也，有木康雄，“局所特徴量によるフィッシャー重みマップに基づく音素認識,” 情報処理学会研究報告.SLP，音声言語情報処理 Vol.2006，No.136(20061221)，pp.197-202，2006.

[2] 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，PRMU 2003-269，vol.103，No.737，pp.79-84，2004.