フィッシャー重みマップに基づく不特定話者音素認識の検討*

◎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-メル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。本稿では、不特定話者で作成した音素モデルに対して、提案手法の有効性を示す。

2 局所特徴量

局所特徴量とは，ある点周辺でのあるパターンの値の強さを表した特徴量のことである。Fig1 の局所パターンを、時間-周波数平面に適用したものが局所特徴量の一例である。

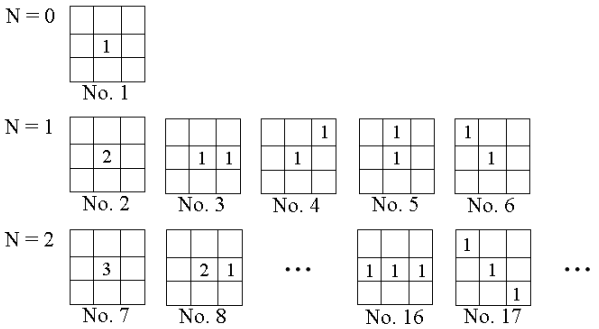

Fig. 1 局所パターンの例

本稿では局所パターン k の変位 $(a_1^{(k)}, \cdots, a_N^{(k)})$ を、点 r の周囲 3×3 の範囲に限定し、次数 N を 2 以下にすると、局所パターンの数 K 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（Fig1）。局所パターンの 1 に対応するパワスペクトルの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。ただし、2 は 2 倍、3 は 3 倍を意味する。

具体的には、時間-周波数平面において点 $r(t, f)$（時刻 $t = 1, \cdots, T$、周波数 $f = 1, \cdots, F$）でのパワスペクトルを $I(r)$ とすると、点 r での局所パターン k の局所特徴量 $h^{(k)}(r)$ は式 (1) で表される。

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

また、k 番目の局所パターンでの時間-周波数平面における全ての点での局所特徴量を、以下のように M 次元ベクトル（$M = (T-2) \times (F-2)$）で表すと

$$\mathbf{h^{(k)}} = [h^{(k)}(2,2) \cdots h^{(k)}(F-1, T-1)]^t \quad (2)$$

となり、$\mathbf{h^{(k)}}$ を横一列に並べたものを

$$\mathbf{H} = [\mathbf{h}^{(1)} \cdots \mathbf{h}^{(K)}] \quad (3)$$

とし、これを局所特徴量行列 $\mathbf{H}$ とする。

3 フィッシャー重みマップ

次に、局所特徴量行列 $\mathbf{H}$ に対して認識に重要な特徴を含んでいる領域に重み付けをし、新たな特徴量 $\mathbf{X}$ を抽出する。本稿では、フィッシャーの判別基準 [1] を利用して重み付けを実行する。

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 $\{\mathbf{H}_i \in R^{M \times K}\}_{i=1}^N$、重み付けをした新たな特徴ベクトルを $\{\mathbf{x}_i = \mathbf{H}_i^t \mathbf{w} : \mathbf{w}$ は重みベクトル $\}_{i=1}^N$、クラス内共分散行列を $\tilde{\Sigma}_W$、クラス間共分散行列を $\tilde{\Sigma}_B$ で表すと、次式が得られる。

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

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

ここで、J はクラス数（ここでは音素数）、ω_j は j 番目のクラス、N_j はクラス ω_j に属する総サンプル数、$\bar{\mathbf{x}}^{(j)}$ 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、

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

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

$$\Sigma_B\mathbf{w} = \lambda\Sigma_W\mathbf{w} \quad (7)$$

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 $\mathbf{w}$ 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。

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

ただし、$\mathbf{X}$ は行列なので、最終的には式（9）のように、縦一列に並べたベクトル $\acute{\mathbf{x}}$ が最終的な音声特徴量となる。

---

*Study on Speaker Independent Phoneme Recognition Using Fisher-Weight-Map. by Shunsuke Kato, Tetsuya Takiguchi and Yasuo Ariki (Kobe University)

$$\dot{\mathbf{x}} = \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_c} \end{bmatrix} \quad (9)$$

4 音声認識への適応

局所特徴量とフィッシャー重みマップを用いた音声認識の流れを Fig2 に示す。まず、入力音声をフーリエ変換し、時間-メル周波数平面に変換する。64 次元のメル周波数なのは事前実験により時間-周波数平面より良い結果が出たためである。次に、事前実験 [2] の結果より、時間軸方向に対してフレーム幅 5、シフト幅 1 に切り出し、フレームごとに局所特徴量 $\mathbf{H}$ に変換する。学習データ分の局所特徴量行列から重み $\mathbf{W}$ を学習し、行列の特徴量行列 $\mathbf{X}$ を求めベクトルの特徴量 $\dot{\mathbf{x}}$ を求める。さらに、学習データ分の特徴量 $\dot{\mathbf{x}}$ から GMM の分布を学習し、それによって識別する。

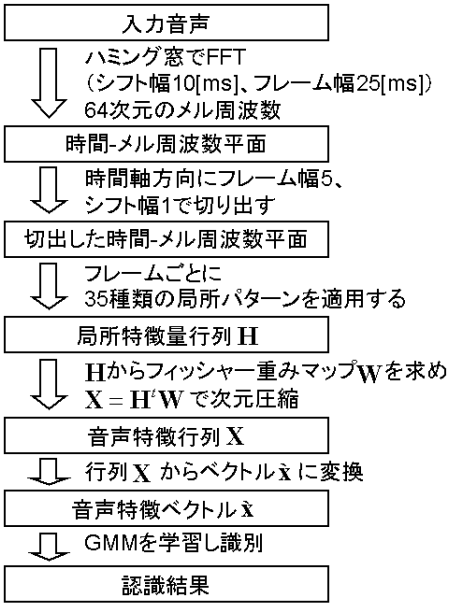

Fig. 2 認識の流れ

5 認識実験

5.1 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 01〜10 の男性話者 6 名、女性話者 4 名の音声を使用した。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。

音声特徴量として、提案手法、提案手法の特徴量にさらに PCA 圧縮をかけたもの、MFCC（12 次元＋対数パワ）、MFCC+Δ MFCC（24 次元＋対数パワ＋Δ対数パワ）、提案手法（PCA 圧縮あり）と MFCC（12 次元＋対数パワ）を組み合わせた特徴量、提案手法（PCA 圧縮あり）と MFCC+Δ MFCC（24 次元＋対数パワ＋Δ対数パワ）を組み合わせた特徴量の６つを利用した。二つの特徴量を組み合わせた特徴量は、識別時にストリーム別に分けて重みをかけ認識をした。

5.2 特定話者モデルでの認識実験

フィッシャー重み $\mathbf{W}$、GMM の分布、PCA 圧縮時の重みは各話者ごとに作成して実験した。結果は Table1 に示す。事前実験 [3] により、提案手法の重み $\mathbf{W}$ の本数 c は 25 本（25×35=875 次元）、PCA 圧縮の時の次元数は 150 次元とした。また、組み合わせた特徴量のストリーム重みは、提案手法:MFCC=0.6:0.4、提案手法:MFCC+Δ MFCC=0.3:0.7 で実験をした。

Table 1 特定話者モデルでの認識実験

特徴量	識別率 (%)
提案手法 (PCA なし)	74.2
提案手法 (PCA あり)	79.5
MFCC	74.5
MFCC+Δ MFCC	86.7
提案手法 (PCA あり)+MFCC	88.4
提案手法 (PCA あり) +MFCC+Δ MFCC	88.3

Table 2 不特定話者モデルでの認識実験

特徴量	識別率 (%)
提案手法 (PCA なし)	78.8
提案手法 (PCA あり)	82.8
MFCC(13 次元、パワあり)	72.1
MFCC+Δ MFCC	86.0
提案手法 (PCA あり)+MFCC	84.3
提案手法 (PCA あり) +MFCC+Δ MFCC	88.2

提案手法 (PCA あり) は MFCC より良い識別率が得られた。また、提案手法 (PCA あり)+MFCC は、MFCC+Δ MFCC より 1.7 ポイント良い結果となった。

5.3 不特定話者モデルでの認識実験

次に、フィッシャー重み W、GMM の分布、PCA 圧縮時の重みは全ての話者で作成して実験した。結果は Table2 に示す。事前実験により、提案手法の重み $\mathbf{W}$ の本数 c は 35 本（35×35=1225 次元）、PCA 圧縮の時の次元数は 50 次元とした。また、組み合わせた特徴量のストリーム重みは、提案手法:MFCC=0.6:0.4、提案手法:MFCC+Δ MFCC=0.3:0.7 で実験をした。

MFCC は特定話者モデルのときより 2.4 ポイント識別率が低下し、MFCC+Δ MFCC は 0.7 ポイント低下したが、提案手法（PCA あり）では特定話者モデルのときより 4.2 ポイント上昇した。これは学習データが増えることによって上手く重み $\mathbf{W}$ を推定できた為であると推測できる。また、提案手法+MFCC+Δ MFCC は不特定話者モデル時でも一番識別率が高く 88.2%となり、特定話者モデルの時とほぼ同じ識別率となった。

6 まとめ

本稿では、局所特徴量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ついて報告し、MFCC、Δ MFCC と提案手法を組み合わせた特徴量での有効性を示した。また、不特定話者モデルによるによる認識の有効性も示した。今後は、連続音声認識での認識へ拡張していきたい。

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] 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,” SIG-SLP64 SP2006-118, pp.19-24, 2006-12

[3] 加藤俊祐, 滝口哲也, 有木康雄, “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,” 音響学会，1-P-10，pp.163-164，2006-03