勾配ヒストグラムに基づく時間-周波数特徴を用いた単語認識*

◎室井貴司, 滝口哲也, 有木康雄 (神戸大)

1　はじめに

音響，音声認識分野では，特徴量として MFCC が広く用いられているが，時間特徴が表現されていないという問題がある。この問題に対し，特徴量の線形回帰係数である ΔMFCC が提案され，音声認識において効果を上げている。しかし，これらは線形回帰係数であるため，フォルマント遷移などの音声の時間変化を表現するには間接的であり，より直接的に時間特徴を表現する特徴量が望まれる。

これまで，我々は，より直接的な時間特徴の表現として，時間-周波数平面上における勾配情報に基づく音声特徴量抽出手法を提案してきた [1]。勾配情報による特徴量は，画像の分野では SIFT(Scale Invariant Feature Transform)[2] や HOG (Histograms of Oriented Gradients)[3] に用いられ，様々な画像の認識に対して有効性が示されており，音声特徴量抽出においても，その有効性が報告されている [4]。本稿では，HOG を音声認識に応用した特徴抽出手法を提案し，単語認識実験を行い，その有効性を示す。

2　特徴量の記述

本稿で提案する HOG による局所特徴量は，時間-周波数平面 $r(t,f)$ 上で勾配ヒストグラムを作成することで得られる。局所特徴量の記述には参照点の周辺領域における勾配情報を用る。使用する勾配情報は参照点を中心とする一定の半径を持つ円領域内から求める。さらに，図 1 に示すように，周辺領域を 4 点 ×4 点の領域を持つ 4 つのブロックに分割し，各ブロックごとに勾配ヒストグラムを作成する。

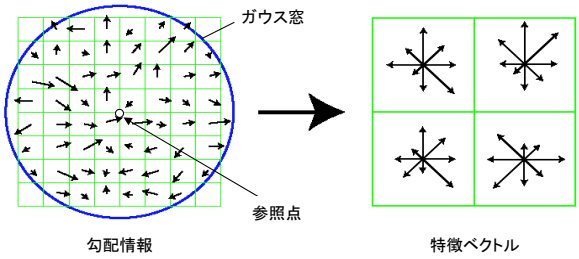


Fig. 1　局所特徴量の記述

2.1　局所特徴量

勾配ヒストグラムを求めるために，まず時間-周波数平面 $r(t,f)$ における勾配強度 $m(t,f)$ と勾配方向 $\theta(t,f)$ を次のように求める。

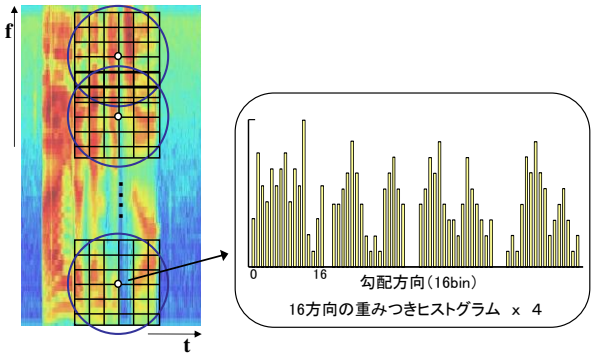


Fig. 2　勾配ヒストグラムの作成

$$m(t,f) = \sqrt{d_t(t,f)^2 + d_f(t,f)^2} \quad (1)$$

$$\theta(t,f) = tan^{-1}\frac{d_f(t,f)}{d_t(t,f)} \quad (2)$$

$$\begin{cases} d_t(t,f) = r(t+1,f) - r(t-1,f) \\ d_f(t,f) = r(t,f+1) - r(t,f-1) \end{cases} \quad (3)$$

次に，局所領域における勾配強度 $m(t,f)$ と勾配方向 $\theta(t,f)$ から重み付き方向ヒストグラム h を以下のように作成する。

$$h_{\theta'} = \sum_x \sum_y G(x,y,\sigma) \cdot m(x,y) \cdot \delta[\theta', \theta(x,y)] \quad (4)$$

h'_θ は全方向を 16 方向に量子化したヒストグラムであり，点 (x,y) は局所領域内の各ブロックに含まれる点を表す。ここで，勾配強度 $m(x,y)$ に対し，局所領域と同じ大きさを持つガウス窓 $G(x,y,\sigma)$ による重み付けを行うことにより，参照点に近い点の特徴がより強く反映される。δ は Kronecker のデルタ関数で，勾配方向 $\theta(x,y)$ が量子化後の θ' に含まれるとき 1 を返す。得られた 4 つのヒストグラムの値を並べた 16 方向 ×4=64 次元のベクトルを局所特徴量とする。これを図 2 に示す。

2.2　音声特徴量ベクトル

局所特徴量を時間，周波数方向に等間隔で算出し，得られた局所特徴量をフレーム内で縦につなげたベクトル $\mathbf{x}$ を音声特徴ベクトルとする。時間-周波数平面において周波数軸上で 8 点の特徴点をとるとき，音声特徴ベクトル $\mathbf{x}$ は，(16×4：ヒストグラムの次元数)×(8：特徴点の数）=512 次元となる。

---

*Gradient-Based Spectro-Temporal Features for Word Recognition, by MUROI, Takashi, TAKIGUCHI, Testuya, ARIKI, Yasuo (Kobe University)

3　評価実験

3.1　実験条件

評価実験として孤立単語音声認識を行った．評価データは ATR の Speech Database SET-A の男性話者 5 名，女性話者 5 名のデータを用い，これに CENSREC-1-C に収録されている食堂と道路の雑音を重畳したものを使用した。学習には各話者 2,620 単語，テストには 1,000 単語を用いた。音声信号の標本化周波数は 8KHz，フレーム幅は 25ms，シフト幅は 10ms であり，時間-周波数平面としてメルフィルタバンク出力 (64 次元) を用いた。また，特徴点の周波数方向の間隔は 8 であり，学習・認識には特定話者 HMM を使用し，実験結果は 10 名の平均値より求めた。

3.2　PCA による次元圧縮

提案手法 (HOG) の音声特徴量 $\mathbf{x}$ の次元数は，512 次元と高次元であることから HMM の確率推定に問題が生じる可能性があるため，主成分分析（PCA）により次元圧縮を行う。1 話者 (MAU) のデータを用いた場合の次元数による認識率の変化を図 3 に示す。

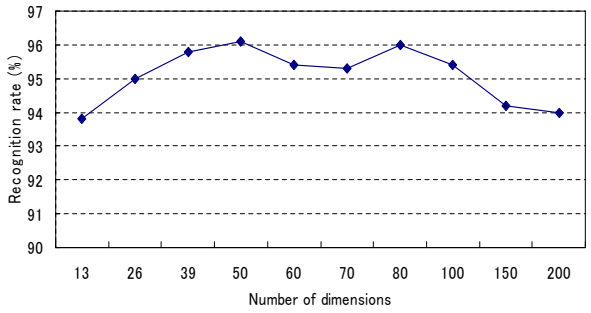


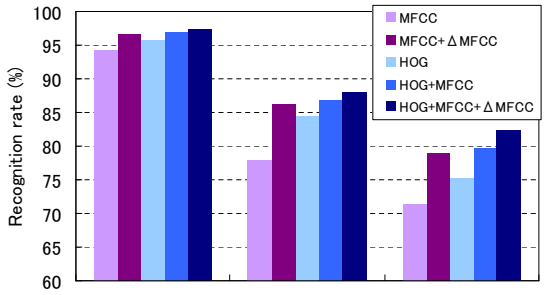
Fig. 3　次元数による認識率の推移（Clean 音声）

結果より，50 次元以降は次元数の増加による認識率の改善が見られず，50 次元のときに 96.1% と最も高い認識率が得られた。これより，以後の実験では 50 次元に圧縮したものを用いることとする。

3.3　単語認識実験結果

提案手法 (HOG)，MFCC，ΔMFCC を用いた認識実験の結果を図 4 に示す。HOG の認識率は，単一で使用した場合，MFCC よりも高い認識率が得られたが，MFCC+ΔMFCC がさらに良い結果を示した。しかし，HOG に MFCC を加えることで MFCC+ΔMFCC よりも高い認識率が得られ，HOG に MFCC+ΔMFCC を加えることにより，さらに認識率の改善が得られた。

勾配に基づく特徴抽出手法は雑音の影響を受けやすいと考えられるが，実験結果は食堂，道路のどちらの環境においてもクリーン音声と同様の傾向を示した。また，今回は HOG と MFCC を組み合わせる場合においても学習・認識はシングルストリームで行ったが，マルチストリームで学習・認識を行うことで，認識率の改善が期待できる。



	Clean	Restaurant	Street
MFCC	94.2	77.9	71.4
MFCC+ΔMFCC	96.6	86.2	78.9
HOG	95.7	84.5	75.1
HOG+MFCC	96.9	86.8	79.6
HOG+MFCC+ΔMFCC	97.3	88.1	82.3

Fig. 4　孤立単語認識実験結果

4　おわりに

本稿では，音声認識において勾配情報に基づく特徴量がどの程度有効であるか実験を行った。この手法は，時間－周波数平面上のパワースペクトル値の勾配方向を 16 方向に量子化し，ヒストグラムとして表現することにより特徴抽出を行うものである。特定話者モデルを用いた孤立単語認識実験では，提案手法により，MFCC に比べ高い認識精度が得られた。また，提案手法と MFCC を組み合わせることにより，さらに認識精度の改善が得られ，提案手法が有効であることが示された。今後は，不特定話者モデルを用いた実験を行っていく予定である。

参考文献

[1] 室井 他, 音講論（秋）, pp.139-140, 2008.

[2] D.Lowe, “Distinctive image feature from scale invariant keypoints,” International Journal of Conference on ComputerVision (IJCV), 2004.

[3] N. Dalal, “Histgrams of oriented gradients for human detection,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[4] 岡隆一, 電子情報通信学会技術報告書. SP, 音声, vol.99, No.577(20000121) pp.13-20, SP99-139, 2000.