勾配に基づく特徴量を用いた音声認識の検討*

◎室井貴司, 滝口哲也, 有木康雄 (神戸大)

1　はじめに

音声認識分野において，特徴量として短時間パワースペクトラム情報に基づく MFCC と，その回帰係数である Δ，ΔΔMFCC を組み合わせたものが広く用いられている。しかし，ΔMFCC は，MFCC の時間変化を近似したものであり，時間特徴の直接的な表現ができていないという問題がある。

本稿では，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対数パワースペクトルの勾配情報に基づく特徴量を用いた音声特徴量抽出手法について検討を行う。勾配情報による特徴量は，画像の分野では SIFT[1] や HOG[2] 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 [3]。また，時間-周波数平面上に表れる特徴を利用した研究 [4] についても様々な報告が成されている。本研究では，SIFT descriptor による局所特徴量を用いて音素認識実験を行い，その有効性を検討する。

2　特徴量の記述

SIFT descriptor による局所特徴量は，時間-周波数平面 $r(t,f)$ 上での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の持つ勾配情報を用い，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 16 方向の勾配ヒストグラムを作成する。

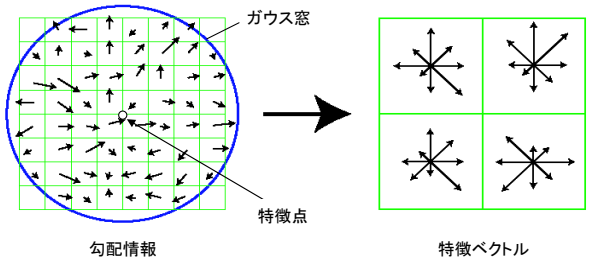

Fig. 1　局所特徴量の記述

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。

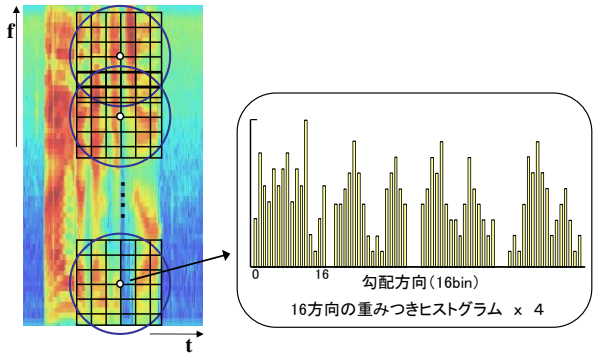

Fig. 2　勾配ヒストグラムの作成

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

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

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

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。

---

*Study on Gradient-Based Feature Extraction Method for Speech Recognition by MUROI, Takashi, TAKIGUCHI, Testuya, ARIKI, Yasuo(Kobe University)

3　音素認識実験

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 6 名，女性話者 4 名の音声を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 5ms であり，時間-周波数平面にはメルフィルタバンク出力 (64 次元) を用いた。特徴点の間隔は時間方向に 2(10ms)，周波数方向に 8 とし，学習・認識には特定話者 HMM を使用した。実験結果は 10 人分の平均値より求めた。

3.2　実験結果

3.2.1　PCA による次元圧縮

音声特徴量 $\mathbf{x}$ の次元数は，周波数方向に特徴点を 8 点取る場合，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。次元数による認識率の変化を図 3 に示す。

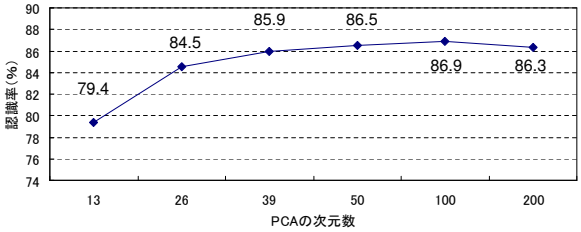

Fig. 3　PCA の次元数による認識率の推移

結果より，39 次元以降は次元数の増加による認識率の改善が緩やかになり，100 次元のときに 86.9% と最も高い認識率が得られた。

3.2.2　単一の特徴量による認識

提案手法，MFCC，ΔMFCC による音素認識実験の結果を表 1 に示す。提案手法の認識率は 100 次元のとき 86.9% であり，MFCC，ΔMFCC に比べて，それぞれ 12.62，11.07*point* の改善が得られた。また，図 3 より，MFCC(ΔMFCC) と同次元 (13 次元) に圧縮した場合でもそれぞれ 5.12，3.57*point* の改善が得られていることがわかる。

Table 1　単一の特徴量による音素認識結果（%）

Proposed	MFCC	ΔMFCC
86.9	74.28	75.83

3.2.3　提案手法と MFCC の組み合わせによる認識

次に，提案手法と MFCC を組合わせたものと MFCC+ΔMFCC の比較実験を行った。この結果を表 2 に示す。ここで，提案手法の次元数が 50 のとき 88.72% となり，MFCC+Δ と比べ，2.05*point* の改善が得られた。また，提案手法 (13 次元) と MFCC を組合わせた特徴量においても 87.26% の認識率が得られ，MFCC+Δ と比較して 0.59*point* の改善が得られている。次元数の増加に伴う提案手法の認識率の増加が，単一で用いた際に比べて小さい理由としては，特徴ベクトルの次元数に大きな差ができ，確率推定の際に MFCC の情報が反映されにくくなることが考えられる。

Table 2　複数特徴量による音素認識結果（%）

MFCC+ΔMFCC	86.67
Proposed(13 次元)+MFCC	87.26
Proposed(26 次元)+MFCC	88.11
Proposed(39 次元)+MFCC	88.50
Proposed(50 次元)+MFCC	88.72

4　おわりに

本稿では，勾配情報に基づく特徴量の音声認識における有効性について報告した。クリーン音声を用いた音素認識実験では，提案手法により，MFCC に比べ高い認識精度が得られた。しかし，提案手法は勾配特徴量であることから，現段階では雑音の影響を受けやすいと考えられる。そのため，今後は雑音に対するロバストな手法へと拡張していく予定である。また，次元圧縮の手法や特徴点の決定方法についても検討していく必要がある。

参考文献

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

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

[3] 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.

[4] Y. Amit, “Robust acoustic object detection,” Journal of the Acoustical Society of America, 118:887-906, 2005.