社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

スペクトル平面における勾配ヒストグラムに基づく音声特徴量の検討

室井　貴司†　　滝口　哲也†　　有木　康雄†

† 神戸大学工学研究科　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
E-mail: †muroi@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　本稿では，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．現在，音声特徴量として 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．また，ΔMFCC や ΔΔMFCC は線形回帰係数であるため，時間特徴の直接的な表現でないと言える．これに対し，本研究では，より直接的に時間特徴を表現するため，時間－周波数平面上の局所領域から勾配情報に基づく音声特徴量を抽出する手法を提案する．本稿で提案する手法は，画像認識分野で用いられている SIFT(Scale Invariant Feature Transform) や HOG(Histograms of Oriented Gradients) などの勾配に基づく特徴抽出手法を音声認識に応用したものである．これらは，物体認識や画像識別など様々な画像タスクにおいて効果を挙げている．提案手法に対し，評価実験として音素識別実験を行ったところ，MFCC と比べ，高い識別率が得られた．また，MFCC と組み合わせることにより，さらに識別精度の改善が得られた．
キーワード　勾配ヒストグラム，時間-周波数特徴，音素認識

Study on Spectro-Temporal Features Based on Gradient Histograms

Takashi MUROI†, Tetsuya TAKIGUCHI†, and Yasuo ARIKI†

† Graduate School of Engineering, Kobe University　Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
E-mail: †muroi@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　This paper proposes a novel feature extraction method for speech recognition based on gradient features on 2-D time-frequency matrix. Widely used MFCC features lack temporal dynamics and delta-MFCC is an indirect expression of temporal frequency changes. To extract the temporal dynamics more directly, local gradient features are measured in the region around reference positions. This method was originally proposed as HOG (Histograms of Oriented Gradients) and applied to human body detection in image recognition. In this paper, we develop it into gradient-based acoustic features in speech recognition. The proposed feature was evaluated on a phoneme recognition task and showed the significant improvement for clean speech and even for the noisy speech when combined with MFCC.
Key words　gradient histograms, spectro-temporal features, phoneme recognition

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 や ΔΔMFCC が提案され，音声認識において効果を上げている [1], [2]．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．

これに対し，時間-周波数平面上の幾何的構造に基づいて特徴を抽出することで音声の時間特徴を表現する手法が提案されている [3]．文献 [4] では，様々な形状のフィルタを用いることで幾何的構造を抽出しており，文献 [5] では，スペクトログラムから方向性パターンを抽出することにより単語認識精度を改善させている．我々はこれまでに，時間－周波数平面上の 3 × 3(時間 × 周波数) 領域から幾何学的構造を抽出する手法を提案してきた [6], [7]．このような局所特徴に基づく手法は，画像認識の分野では物体認識や画像識別など様々な用途に用いられている．近年では，SIFT(Scale Invariant Feature Transform) [8] や HOG (Histograms of Oriented Gradients) [9] が提案され，様々なタスクで高い効果を挙げ，注目されている．本研究では，SIFT や HOG のような勾配に基づく特徴抽出手法を音声認識

に応用する手法を提案する．本稿では，評価実験として，25 音素を用いた音素識別実験をクリーン音声，雑音重畳音声に対して行い，提案手法の有効性を示す．

2 章で時間－周波数平面に前処理として行なう，バイラテラルフィルタリング [10] について述べる．3 章では，勾配特徴の抽出手法について説明し，4 章で評価実験により提案手法の有効性を確認する．最後に 5 章でまとめとして今後の課題について述べる．

2. バイラテラルフィルタリング

時間－周波数平面として用いる短時間フーリエ変換後の音声のメルフィルタバンク出力に対し，前処理としてバイラテラルフィルタによる平滑化を行う．バイラテラルフィルタは，注目する点との距離による重み付けだけでなく，注目する点との対数パワースペクトル値の差に応じてガウス関数により重みを付けた平均化を行うフィルタである．バイラテラルフィルタリングを行なうことにより，時間－周波数平面上のフォルマント遷移等の大局的な変化を保存しつつ微細なノイズを除去することができる．

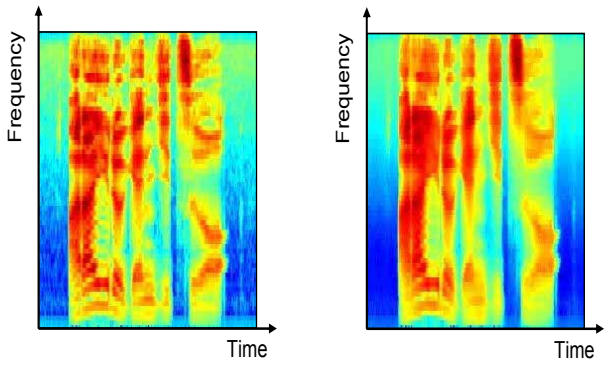
時間－周波数平面上の点 $r_i = (t_i, f_i : i = 1, 2, \ldots, n)$ におけるバイラテラルフィルタの出力 f_i は，

$$f_i = \frac{\sum_{j=1}^{n} w_r(r_i, r_j) w_I(I_i, I_j) I_j}{\sum_{j=1}^{n} w_r(r_i, r_j) w_I(I_i, I_j)} \tag{1}$$

で与えられる．ここで，w_r は空間方向の重み，w_I は対数パワースペクトル値の差に関する重みであり，ガウス関数を用いてそれぞれ次のように表される．

$$\begin{cases} w_r(r_i, r_j) = e^{-\frac{1}{2}||r_i - r_j||^2} \\ w_I(I_i, I_j) = e^{-\frac{1}{2}||I_i - I_j||^2} \end{cases} \tag{2}$$

図 1 に雑音なしの環境における男声発話のメルフィルタバンク出力 (a) と，バイラテラルフィルタリング後の時間－周波数平面 (b) を示す．



(a) (b)

図 1 メルフィルタバンク出力 (a) とバイラテラルフィルタリング後の時間-周波数平面 (b).

Fig. 1 Mel-scale time-frequency matrix (a) and blurred by bilateral filters (b).

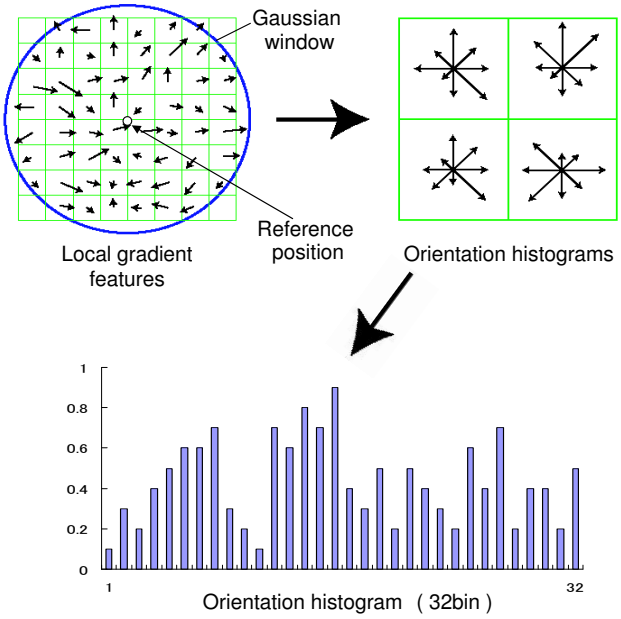


図 2 局所勾配特徴（左図）と重み付き方向ヒストグラム (右図，下図).

Fig. 2 The local gradient features (left) and weighted orientation histograms (right and bottom).

3. 局所特徴量

本研究で提案する局所特徴量は，平滑化後の時間－周波数平面上の参照点における局所勾配特徴の方向ヒストグラムを求めることで得られる．局所勾配特徴と，方向ヒストグラムを図 2 に示す．まず，参照点の周辺領域における局所勾配特徴を計算し（図 2 左），さらに，図 2 右に示すように，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 8 方向の方向ヒストグラムを作成する。得られた方向ヒストグラムを全て繋げることで得られるヒストグラムを局所特徴量とする．

3.1 局所勾配特徴

バイラテラルフィルタリングを行った時間－周波数平面上で，局所勾配特徴を求める．局所勾配特徴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．

時間－周波数平面上の点 (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．

$$m(t, f) = \sqrt{d_t(t, f)^2 + d_f(t, f)^2} \tag{3}$$

$$\theta(t, f) = tan^{-1} \frac{d_f(t, f)}{d_t(t, f)} \tag{4}$$

$$\begin{cases} d_t(t, f) = I(t + 1, f) - I(t - 1, f) \\ d_f(t, f) = I(t, f + 1) - I(t, f - 1) \end{cases} \tag{5}$$

ここで，$I(t, f)$ は平滑化後の時間周波数平面上の点 (t, f) における対数パワースペクトル値を表す．次に得られた勾配強度 $m(t, f)$ に対し，参照点を中心とするガウス窓による重み付け

を行う．勾配強度は図 2 左側上の矢印で示す．これにより，参照点に近い点の特徴がより強く反映され，参照点から離れた位置の情報が低減される．また，勾配方向の微細な変化を取り除く平滑化の効果も得ることができる．

3.2 重み付き方向ヒストグラム

ガウス窓によって重み付けがなされた方向ヒストグラムを図 2 の右側に示す．矢印の向きと長さで，4×4 局所領域における各方向の勾配強度を表している．重み付き方向ヒストグラム h は，局所領域における勾配強度 $m(t,f)$ と勾配方向 $\theta(t,f)$ を用いて以下のように作成する．

$$h_{\theta'} = \sum_x \sum_y G(x,y,\sigma) \cdot m(x,y) \cdot \delta[\theta', \theta(x,y)] \qquad (6)$$

h'_θ は 360 度全方向を 45 度ごとの 8 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，$G(x,y,\sigma)$ は，局所領域と同じ大きさを持つガウス窓であり，これを用いて勾配強度 $m(x,y)$ に対し重み付けを行なう．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 8 方向 $\times 4 = 32$ 次元のベクトルを重み付き方向ヒストグラムとして得る．

3.3 勾配特徴ベクトル

重み付き方向ヒストグラムを算出する参照点を周波数方向に等間隔で配置し，得られた重み付き方向ヒストグラムをフレーム内で縦に並べたベクトル $\mathbf{X}$ を勾配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，勾配特徴ベクトル $\mathbf{X}$ は，(8×4：重み付き方向ヒストグラムの次元数)×(8：特徴点の数) = 256 次元ベクトルとなる．

3.4 特徴抽出の流れ

特徴抽出から音素識別実験までの流れを図 4 に示す．まず，入力音声を短時間フーリエ変換とメルフィルタバンクを用いて時間－周波数平面で表す．時間－周波数平面上の微小な変化を除き，フォルマント遷移等の大きな変化を強調するためにバイラテラルフィルタリングを行う．次に，平滑化された時間－周波数平面上で周波数方向に等間隔で参照点を取り，参照点の周囲の局所領域で勾配特徴を計算する．得られた勾配特徴を量子化し，方向ヒストグラムとして表す．これを同一フレーム内で全て縦に並べることで勾配特徴ベクトル $\mathbf{X}$ が得られる．

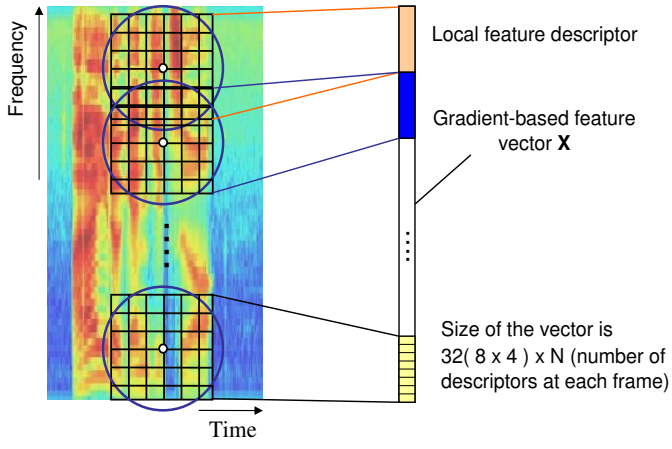


図 3 フレーム内で勾配特徴を並べることにより勾配特徴ベクトルを得る．

Fig. 3 Gradient-based feature vector obtained by packing the local feature descriptors at a frame.

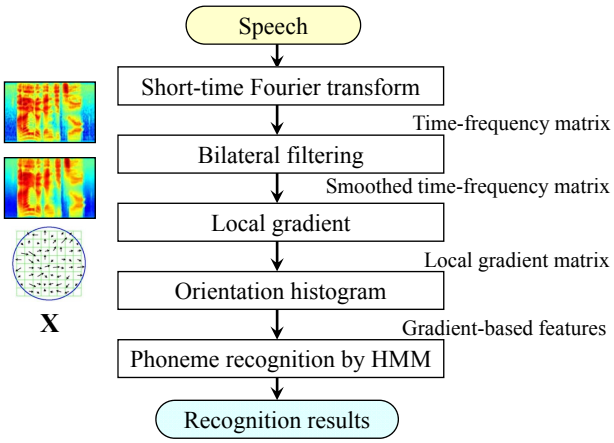


図 4 音声特徴抽出の流れ．

Fig. 4 Flow of the proposed feature extraction.

評価実験では，3 状態 HMM によって音素ごとに学習データを用いて学習を行い，識別用データに対して特定話者音素識別を実行する．

4. 音素識別実験

4.1 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（MHT，MTK，MSH，MHO，MMY，MYI），女性話者 4 名（FKN，FTK，FYM，FKS）の音声データに雑音を重畳したものを使用した．雑音には白色雑音（SNR=∞，10，0dB）と，実環境雑音として，CENSREC-1-C データベース [11] に収録されている高速道路付近（Street）と学生食堂（Restaurant）の 2 環境で SNR（−5 ≦SNR≦ 10dB）で録音されたデータの非音声部分を使用した．各話者のデータは音素ごとに切り出し，音素識別実験を行った．音素は 25 種類，各話者の学習用音声データは 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．音響モデルは HMM を用い，状態数は 3，各状態の混合数は，予備実験より最も高い識別精度を示した 36 とした．HMM の学習には 10 人分の学習用データを全て使用して，特定話者モデルとして学習し， 識別は話者ごとの識別用データを別々に使用して識別実験を行い，話者ごとに識別率を算出した．

音声信号の標本化周波数はクリーン音声および白色雑音重畳音声は 16kHz，実環境雑音重畳音声は 8kHz（CENSREC-1-C に収録されているデータの標本化周波数が 8kHz であるため），フレーム幅，シフト幅はそれぞれ共に 25ms，10ms である．また，予備実験により，フーリエ変換のみの時間-周波数平面（128

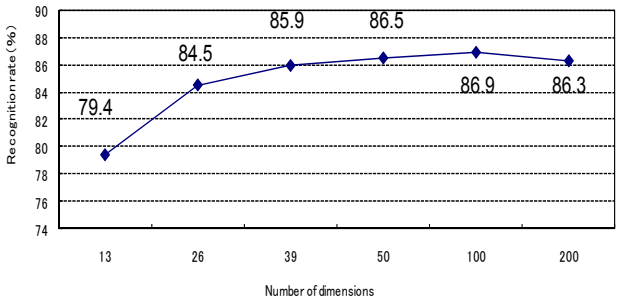


図 5　PCA の次元数による音素識別率の変化.

Fig. 5　Result of phoneme recognition as a function of the number of feature dimensions after PCA.

次元）よりも，64 次元の対数メルフィルタバンク出力を用いた時間-メル周波数平面（64 次元）の方がよい結果を示したため，評価実験は全てメルフィルタバンク出力による時間-周波数平面上で行った．これは，フォルマントなど音声の幾何学的な特徴が低周波数域によく表れるので，低周波数域を強調するメル周波数平面の方が対数パワースペクトルを用いた時間-周波数平面よりも，音素ごとの特徴をより強く示すためと考えられる．

4.2　実験結果

4.2.1　音声特徴ベクトルの次元圧縮

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，256 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（Principal Component Analysis: PCA）により次元圧縮を行う．次元数は，10 名のクリーン音声データを用いた予備実験により決定する．結果は 10 名の識別率の平均値により求める．次元数による識別率の変化を図 5 に示す.

これより，100 次元のときに最も高い識別率が得られたが，高次元であること，50 次元以降の識別率の変化が小さいことから，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．

4.2.2　白色雑音重畳音素識別実験

提案手法を PCA により 50 次元に圧縮した特徴量と，MFCC の特徴量を用いて実験を行なった．ここで，勾配特徴ベクトルは時間周波数平面上の 8×8 近傍の局所領域で作成した重み付き方向ヒストグラムに基づいているため，8 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 8 フレーム分の情報を持つ ΔMFCC を用いた.

音素識別実験の結果を表 1 に示す．表 1 より，SNR によらず提案手法が最も良い識別率を示した．また，図 5 の提案手法の結果と MFCC，ΔMFCC の結果を比較すると，クリーン音声において，MFCC(ΔMFCC) と同次元 (13 次元) に圧縮した場合でもそれぞれ 5.1，3.6point の改善が得られており，MFCC，ΔMFCC と比べ識別精度が高いと言える．

次に，提案手法を 50 次元に圧縮したものに 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．提案手法に MFCC を組み合わせることで，SNR によらず MFCC+ΔMFCC よりも高い識別率が得られた.

また，提案手法の次元数を変化させて MFCC と組み合わせた場合の結果を表 3 に示す．ここで，提案手法の次元数が 50 のとき 88.7% となり，MFCC+ΔMFCC と比べ，2.2point 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 87.3% の識別率が得られ，MFCC+ΔMFCC と比較して 0.7point の改善が得られている．これより，提案手法は ΔMFCC に比べ，識別率の改善に貢献していることがわかる．次元数の増加に伴う提案手法と MFCC を組み合わせた特徴量の識別率の増加が，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.

表 1　単一特徴量での音素識別結果.

Table 1　Results of speaker dependent phoneme recognition by single feature.

(%)

SNR	clean	10dB	0dB
Proposed feature with PCA 50dim	86.5	70.3	48.4
MFCC with Power	74.3	47.1	30.2
ΔMFCC with Power	75.8	57.6	35.4

表 2　特徴量を組み合わせた場合の音素識別結果.

Table 2　Results of speaker dependent phoneme recognition by feature integration.

(%)

SNR	clean	10dB	0dB
Proposed feature (50 dim) + MFCC	88.7	72.3	50.3
MFCC + ΔMFCC with Power	86.6	64.3	43.9

表 3　クリーン音声を用いた音素識別率の提案手法の次元数による変化.

Table 3　Results of speaker dependent phoneme recognition by feature integration for clean speach.

(%)

MFCC+ΔMFCC	86.6
Proposed(13dim)+MFCC	87.3
Proposed(26dim)+MFCC	88.1
Proposed(39dim)+MFCC	88.5
Proposed(50dim)+MFCC	88.7

4.2.3　実環境雑音下での音素識別実験

より実環境に近い条件で評価を行うために，実環境にて録音された雑音をクリーン音声に重畳し，識別実験を行った．雑音は学生食堂，高速道路付近の 2 種類の環境で，SNR（−5 ≦SNR≦ 10dB）で録音されたデータの非音声部分を用いた．雑音データが 8kHz であるため，音声データを 16kHz から 8kHz にダウンサンプリングした後，雑音を重畳した．特徴量を単一で使用した場合と，特徴量を複数組み合わせた場合の実験結果をそれぞれ図 6，図 7 に示す．提案手法を PCA で 50 次元に圧縮した特徴量を単一で用いたとき，学生食堂の環境では 42.1% ととなり，MFCC，ΔMFCC に比べ，それぞれ

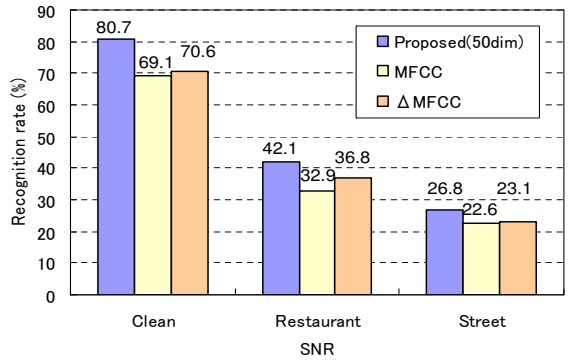


図 6　実環境雑音下における単一特徴量の音素識別結果.

Fig. 6　Results of speaker dependent phoneme recognition by feature integration for noisy speech.

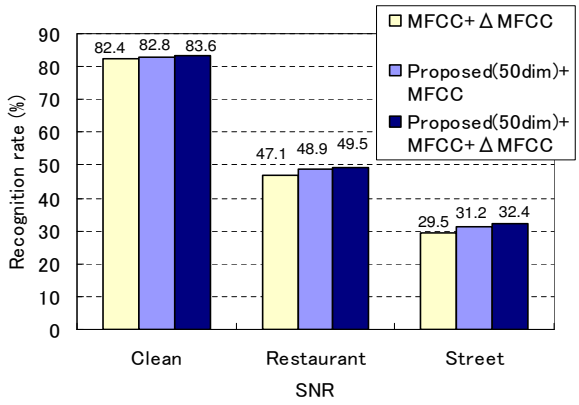


図 7　実環境雑音下において特徴量を組み合わせた場合の音素識別結果.

Fig. 7　Results of speaker dependent phoneme recognition by feature integration for noisy speech.

9.2point，5.3point 識別率が改善した．また，高速道付近の環境では 26.8% となり，それぞれ 4.2point，3.7point の改善が得られた．

提案手法と MFCC を組み合わせた場合は，Proposed+MFCC は MFCC+ΔMFCC に比べ，学生食堂環境で 1.8point，高速道路付近環境では，1.7point の識別率の改善が得られ，提案手法が動的特徴として ΔMFCC よりも識別に貢献していることが分かる．また，MFCC+ΔMFCC とさらに提案手法を加えた特徴量を比較すると，クリーン音声で 1.4point，学生食堂環境で 2.4point，高速道路付近環境では 2.9point の識別率の改善が得られた．これらより，提案手法は MFCC の特徴量が持たない識別に寄与する情報を含んでいると言える．

5. おわりに

本稿では，時間－周波数平面上における対数パワースペクトル値の勾配情報に基づく音声特徴抽出手法を提案し，その有効性について報告した．この手法は，時間周波数平面上の 8 × 8 局所領域において，勾配情報を計算し，ガウス窓によって重み付けを行なった 8 方向の方向ヒストグラムを作成し，音声特徴ベクトルを形成するものである．この音声特徴ベクトルの次元数は 256 次元と高次元であるため，主成分分析 (PCA) により次元削減することで，音素識別実験において MFCC や ΔMFCC に比べ，高い識別精度を示した．また，提案手法と MFCC の特徴量を組み合わせた場合でも，ΔMFCC に比べ，提案手法が識別率の改善に貢献していることがわかった．さらに MFCC+ΔMFCC に提案手法を加えることで，識別率の改善が得られ，提案手法には MFCC の特徴量に含まれていない情報を含むことが示された．しかし，提案手法は PCA による次元削減後でも 50 次元のベクトルであり，MFCC と比べると高次元であることから，勾配特徴ベクトルの持つ情報を損なわずにより低次元へと次元を削減することが求められる．また，より雑音にロバストな性質を持たせるためには，ケプストラム減算法（Cepstrum Mean Subtraction: CMS）のような正規化処理が必要と考えられる [12]．今後は，これらの問題解決に取り組むと同時に，単語認識による評価実験を行なう予定である．

文　献

[1] K. Elinius, M. Blomberg, “Effect of Emphasizing Transitional or Stationary Parts of the Speech Signal in a Discrete Utterance Recognition System,” IEEE Proc. ICASSP ’82, pp.535-538, 1982.

[2] T. Nitta, “A Novel Feature-Extraction for Speech Recognition Based on Multiple Acoustic-Feature Planes,” IEEE Proc. ICASSP ’98, pp.29-32, 1998.

[3] T. Nitta, “Feature Extraction for speech Recognition Based on Orthogonal Acoustic- feature Planes and LDA,” IEEE Proc. ICASSP ’99, pp.421-424, 1999.

[4] Ken Schutte, James Glass, “Speech Recognition with Localized Time-Frequency Pattern Detectors,” Proc. of ASRU 2007, pp.341-344, 2007.

[5] H. Matumura, R. Oka, K. Kogure, Y. Kojima, “Speaker-Independent Spoken Word Recognition by Using the Orientation Patterns Obtained from the Vector Field of Spectrum Pattern,” Transactions of IEICE, Vol.72-D-II, No.4 , pp.487-498, 1989.

[6] Y. Ariki, S. Kato, T. Takiguchi “Phoneme Recognition Based on Fisher Weight Map to Higher-Order Local Auto-Correlation,” Interspeech2006, pp.377-380, 2006.

[7] T. Muroi, T. Takiguchi, Y. Ariki, “Speaker Independent Phoneme Recognition Based on Fisher Weight Map,” The 2nd International Conference on Multimedia and Ubiquitous Engineering (MUE2008), pp.253-257, 2008.

[8] D.Lowe, “Distinctive Image Feature from Scale Invariant Keypoints,” Proc. of International Journal of Conference on ComputerVision (IJCV), 60(2), pp.91-110, 2004.

[9] N. Dalal, B. Triggs, “Histograms of Oriented Gradients for Human Detection,” Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.720-723, 2007.

[10] C. Tomasi, R. Manduchi, “Bilateral Filtering for Gray and Color Images,” Proc. of International Conference on Computer Vision, IEEE (1998), pp. 829-846, 1998.

[11] 北岡教英他, “雑音下音声認識評価ワーキンググループ活動報告：認識に影響する要因の個別評価環境,” 電子情報通信学会技術研究報告, pp. 1-6, 2006.

[12] Cooke, M. P., Green, P. D., Josifovski, L. B., and Vizinho, A., “Robust Automatic Speech Recognition with Missing and Uncertain Acoustic Data,” Speech Communication, 34, pp.267-285, 2001.