社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (34)
2006／12／22

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

加藤 俊祐† 滝口 哲也† 有木 康雄†

†神戸大学自然科学研究科 〒657-8501 兵庫県神戸市灘区六甲台町 1-1

E-mail: †kato-shun@me.cs.scitec.kobe-u.ac.jp, †{takigu,ariki}@kobe-u.ac.jp

あらまし 本稿では，高次局所自己相関（HLAC）とフィッシャー重みマップ（FWM）に基づく新しい音声特徴抽出法について提案する．現在，音響・音声認識分野では，MFCCが広く用いられているが，時間特徴が表現できていないという問題がある．この問題を解決するために，時間—周波数平面上の３ｘ３局所領域において，３５種類の局所パターンに対する局所自己相関特徴を計算し，これを局所特徴量とする．ある一定の時間幅を持つ時間—周波数平面(フレーム)において，３５種類の局所パターンごとに，識別効果の高い領域の局所特徴量に重みを付けて加算し，音声特徴ベクトル(３５次元)を形成する．この重みをフィッシャー重みマップと呼ぶ．音素認識において，HLACとFWMの有効性を確認した．

キーワード 局所自己相関特徴，フィッシャー重みマップ，局所特徴量，音素認識

Phoneme Recognition Based on Fisher Weight Map to Local Features

Shunsuke KATO† Tetsuya TAKIGUCHI† and Yasuo ARIKI†

†Graduated School of Science and Technology, Kobe University, Rokkodaicho 1-1, Nada, Kobe, Hyogo 657-8501 Japan

E-mail: †kato-shun@me.cs.scitec.kobe-u.ac.jp, †{takigu,ariki}@kobe-u.ac.jp

Abstract In this paper, we propose a new feature extraction method based on higher-order local auto-correlation (HLAC) and Fisher weight map (FWM). Widely used MFCC features lack temporal dynamics. To solve this problem, 35 types of local auto-correlation features are computed within two-dimensional local regions. These local features are accumulated over more global regions by weighting high scores on the discriminative areas where the typical features among all phonemes are well expressed. This score map is called Fisher weight map. We verified the effectiveness of the HLAC and FWM through total phoneme recognition.

Keyword Local auto-correlation feature, Fisher weight map, Local feature, Phoneme recognition

1. はじめに

音響・音声認識分野では，MFCCが広く用いられているが，時間特徴が表現できていないという問題がある．この問題を解決するために，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[1][2]．しかし，線形回帰係数であるため，フォルマント遷移などを表現するには間接的であり，より直接的な時間変化を表現した特徴が望まれる．

これに対して，時間—周波数平面上の局所的な幾何学構造に基づき，音声の時間変化特徴を表現する方法が提案されている．文献[3][4][5]では，時間—周波数平面上の３ｘ３領域の集合に対して主成分分析を行い，その直交基底により幾何学構造を抽出している．

音声特徴の幾何学構造は，フォルマントやその時間的遷移によく表されていることから，時間—周波数平面上での局所的な高次局所自己相関として，直接的に表現可能である．この点から，本稿では，音声の時間-周波数平面において，高次局所自己相関などの局所特徴に対して，フィッシャー判別基準を用いて音声特徴量を抽出する手法を提案する．

提案手法の基になっている手法は，高次局所自己相関特徴に対してフィッシャー判別基準を用いて特徴抽出を行なうもので，画像処理・認識の分野においては有効性が示されている[6][7]．

本研究は，画像認識で提案されている文献[6][7]の手法を，音声の特徴抽出に適用するものであり，短時間フーリエ変換後の時間-周波数平面上において，局所特徴を求める．さらに，認識のために重要な局所特徴を

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，重みマップを求める．最後に，局所特徴量と重みマップとの積により音声特徴ベクトルを求めるものである．本研究では，さらに，提案手法による音声特徴量とMFCC特徴量，ΔMFCC特徴量を組み合わせた実験も行っている．

2. 局所特徴量

2.1. 局所パターンと局所特徴量

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．例えば，図1にある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となる．この局所特徴量は，フォルマントの遷移など局所的な時間変化特徴を表していると考えられる．

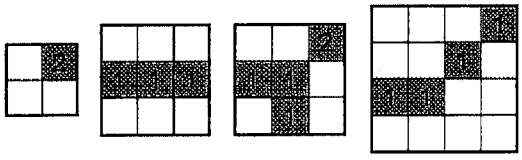

図 1　局所パターンの例

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 $f=1,\cdots,F$）でのパワースペクトルを $I(r)$ とすると，点 r での局所パターン k の局所特徴量 $h_r^{(k)}$ は次式で表される．

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

文献[6]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，和を用いた方が，実験的に精度がよかったため，局所特徴量として式(1)を用いている．局所パターンの変位 $(a_1^{(k)},\cdots,a_N^{(k)})$ を参照点 $r(t,f)$ の近傍3×3の局所領域に限定し，さらに次数 N を高々2までに制限すると，局所パターンの種類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．図2に35種類の局所パターンを示す．

2.2. 局所特徴量行列

図2において，局所パターンの1に対応する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．（但し，図中の2，3は，対応するパワースペクトルの二倍，三倍を意味する．）ここで，ある音素に対する時間-周波数平面上の全ての点 $r(t,f)$ 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（$M=(T-2)\times(F-2)$ 次元）で表記する．

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

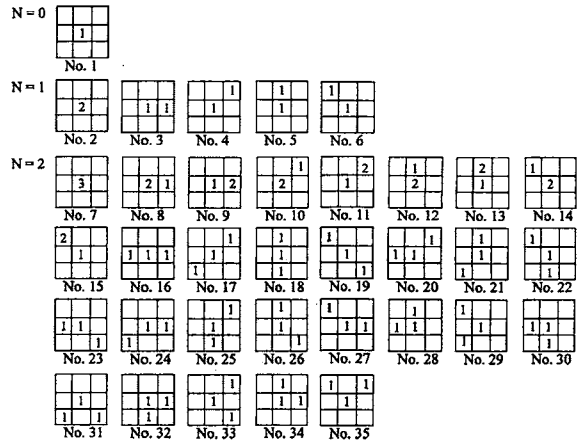

図 2　35種類の局所パターン

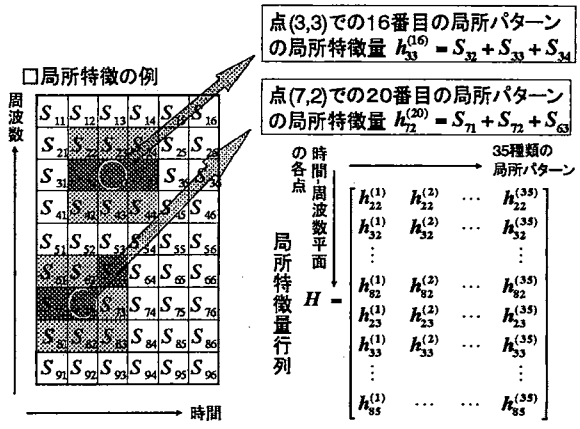

図3　局所特徴量行列

図3には，点(3,3)における16番目の局所パターン局所特徴量 $h_{33}^{(16)}$ と，点(7,2)における20番目の局所パターンの局所特徴量 $h_{72}^{(20)}$ を記載している．

さらに，局所パターンの総数を K 種類（今回は3×3近傍なので $K=35$）として，$h^{(k)}$ を横に並べたものを

$$H = [h^{(1)} \cdots h^{(35)}] \qquad (3)$$

とし，これを局所特量行列 $\mathbf{H}$ とする．

3. フィッシャー重みマップ

3.1. フィッシャーの判別基準

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行い特徴抽出を行う．本稿では，フィッシ

ャーの判別基準を利用し最適な重みマップを決定する．

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特徴量行列を$\{\mathbf{H}_i \in R^{M\times K}\}_{i=1}^{N}$，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$\}_{i=1}^{N}$と書くことにする．特徴ベクトルのクラス内共分散行列$\tilde{\Sigma}_W$とクラス間共分散行列$\tilde{\Sigma}_B$は，

$$\tilde{\Sigma}_W = \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in\omega_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)})(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)})^t \qquad (4)$$

$$\tilde{\Sigma}_B = \frac{1}{N}\sum_{j=1}^{J} N_j(\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}})^t \qquad (5)$$

となる．Jはクラス数（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，

$$J(\mathbf{w}) = \frac{tr\tilde{\Sigma}_B}{tr\tilde{\Sigma}_W} \qquad (6)$$

となる．$tr\tilde{\Sigma}_B$，$tr\tilde{\Sigma}_W$はそれぞれ

$$\begin{aligned} tr\tilde{\Sigma}_W &= \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in\omega_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)})^t(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)}) \\ &= \mathbf{w}^t\left\{\frac{1}{N}\sum_{j=1}^{J}\sum_{i\in\omega_j}(\mathbf{H}_i^{(j)} - \bar{\mathbf{H}}^{(j)})(\mathbf{H}_i^{(j)} - \bar{\mathbf{H}}^{(j)})^t\right\}\mathbf{w} \\ &= \mathbf{w}^t\Sigma_W\mathbf{w} \end{aligned} \qquad (7)$$

$$\begin{aligned} tr\tilde{\Sigma}_B &= \frac{1}{N}\sum_{j=1}^{J} N_j(\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}})^t(\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}}) \\ &= \mathbf{w}^t\left\{\frac{1}{N}\sum_{j=1}^{J} N_j(\bar{\mathbf{H}}^{(j)} - \bar{\mathbf{H}})(\bar{\mathbf{H}}^{(j)} - \bar{\mathbf{H}})^t\right\}\mathbf{w} \\ &= \mathbf{w}^t\Sigma_B\mathbf{w} \end{aligned} \qquad (8)$$

となり，最終的にフィッシャーの判別基準は

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\Sigma_B\mathbf{w}}{\mathbf{w}^t\Sigma_W\mathbf{w}} \qquad (9)$$

となる．このフィッシャーの判別基準$J(\mathbf{w})$を制約条件$\mathbf{w}^t\Sigma_W\mathbf{w} = 1$の下で最大化するので，重み$\mathbf{w}$は固有値問題

$$\Sigma_B\mathbf{w} = \lambda\Sigma_W\mathbf{w} \qquad (10)$$

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をフィッシャー重みマップと呼ぶ．

3.2. 音声特徴量

提案手法の音声特徴量は，式(11)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た$\mathbf{W}$と局所特徴量行列$\mathbf{H}$の転置行列との積として得られる．

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1 \cdots \mathbf{x}_C] \\ &= \mathbf{H}^t[\mathbf{w}_1 \cdots \mathbf{w}_C] \\ &= \mathbf{H}^t\mathbf{W} \end{aligned} \qquad (11)$$

ただし，$\mathbf{X}$は行列なので，最終的には式（12）のように縦一列にならべたベクトル$\mathbf{x}$が最終的な音声特徴量(ベクトル)となる．

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_C \end{bmatrix} \qquad (12)$$

4. 音声認識への適用

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図4に示す．

まず，入力音声をフーリエ変換し，時間-周波数平面に変換する．変換した時間-周波数平面でそのまま局所特徴量を計算すると，時間軸方向の情報がなくなる．このため，時間軸方向に対してある一定幅のフレーム(i)で切り取り，このフレーム内で局所特徴量を計算し，局所特徴量行列$\mathbf{H}_i$をフレームごとに算出する．フレームの切り取り幅は，予備実験により5，シフト幅は1が最適であった．次に，学習用データの局所特徴量$\mathbf{H}_i$ $(i = 1\cdots N : N$は学習データ数$)$から重み$\mathbf{W}$を学習し，学習用，認識用局所特徴量行列$\mathbf{H}_i$に重み$\mathbf{W}$をかけてベクトルに変換し，音声特徴ベクトル$\mathbf{x}_i$を求める．

識別は混合正規分布(GMM, Gaussian Mixture Model)によって識別する．音素ごとに学習用データ$\mathbf{x}_i$ $(i = 1\cdots N : N$は学習データ数$)$を用いて学習を行い，認識用データ$\mathbf{x}$に対して認識を実行する．

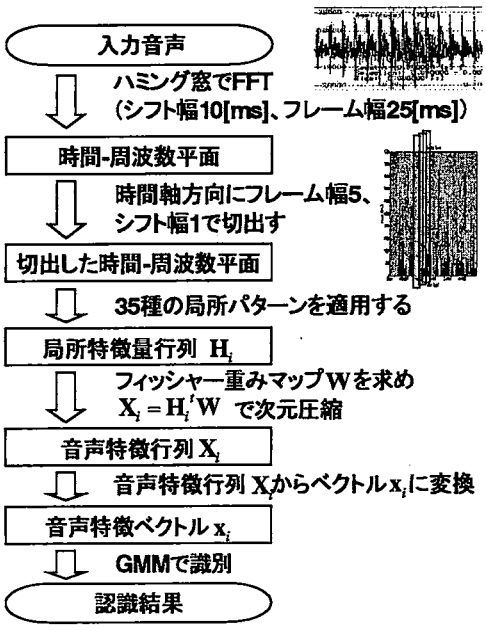

図4　認識の流れ

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（MHT, MTK, MSH, MHO, MMY, MYI），女性話者 4 名（FKN, FTK, FYM, FKS）の音声を使用し，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．GMM の混合分布数は 8 である．音声信号の標本化周波数は 16KHz，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．

5.2. フィッシャー重みマップwの数と音素認識率

予備実験により，時間-周波数平面より時間-メル周波数平面（64 次元）の方が 3%ほど良い識別結果を示した．したがって，以後の実験は全て，時間-メル周波数平面（64 次元）上で行なった．

次に，フィッシャー重みマップ W の数（固有ベクトル）と音素認識率の関係について調べた．結果を図 5 に示す．図より，W の数は 21 以降上昇しないことが分かる．また，累積寄与率は，W の数が 25 前後において 0.99 であった．これより，以後の実験では，W の数を 25 としてフィッシャー重みマップを設計した．

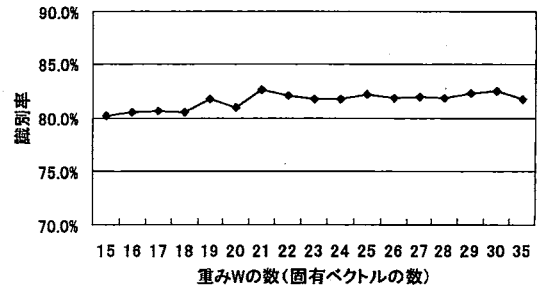

図5　重み W の数と識別率（話者 MHT）

5.3. 音声特徴ベクトルの次元圧縮

音声特徴ベクトル x の次元数は，25（フィッシャー重みマップ W の本数）×35（局所パターンの数）=875 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，音声特徴ベクトル x を主成分分析（PCA）により圧縮し，音素認識実験を行った．結果を表 1 に示す．

表 1　PCA で圧縮して認識した結果

音声特徴	話者10人の平均の識別率
MFCC12次元(対数パワーなし)	71.6%
ΔMFCC12次元(対数パワーなし)	75.8%
PCA圧縮なしフィッシャー(次元875)	74.2%
PCAで圧縮(次元13)	68.7%
PCAで圧縮(次元15)	70.4%
PCAで圧縮(次元20)	71.8%
PCAで圧縮(次元25)	73.3%
PCAで圧縮(次元30)	75.4%
PCAで圧縮(次元35)	76.2%
PCAで圧縮(次元40)	77.0%
PCAで圧縮(次元45)	77.1%
PCAで圧縮(次元50)	77.7%
PCAで圧縮(次元55)	78.1%
PCAで圧縮(次元60)	78.3%
PCAで圧縮(次元65)	78.7%
PCAで圧縮(次元70)	78.7%
PCAで圧縮(次元75)	79.1%
PCAで圧縮(次元80)	79.0%
PCAで圧縮(次元85)	78.9%
PCAで圧縮(次元90)	79.1%
PCAで圧縮(次元95)	79.0%
PCAで圧縮(次元100)	79.2%
PCAで圧縮(次元150)	79.5%
PCAで圧縮(次元200)	78.6%

表より，PCAで圧縮しない場合には，音素認識率は74.2%であったが，PCA で１５０次元に圧縮することにより，79.5%に向上した．

比較のため，MFCC(対数パワー無し)とΔMFCC(対数パワー無し)を用いて認識した結果も，表に併記している．表より，フィッシャー重みマップ＆PCA150 次元が最も高い音素認識率を示した．

5.4. MFCCと組み合わせた認識結果

フィッシャー重みマップ(FWM)＆PCA の特徴量とMFCC を組みあせた実験を行った．実験は，以下の４種類である．

実験 A： FWM＆PCA＋MFCC（対数パワーあり）
（ストリーム重みの比率 FWM&PCA：MFCC=0.6：0.4）

実験 B： FWM＆PCA ＋ MFCC（対数パワーあり）＋ΔMFCC（対数パワーあり）
（ストリーム重みの比率は FWM&PCA：MFCC：ΔMFCC=0.2：0.4：0.4）

実験 C： FWM＆PCA ＋ MFCC（対数パワーあり）＋ΔMFCC（対数パワーあり）
（ストリーム重みの比率は FWM&PCA：MFCC：ΔMFCC=0.0：0.5：0.5）

実験 D: MFCC（対数パワーあり１３次元）＋ΔMFCC（対数パワーあり１３次元）で，２６次元ベクトル

実験結果を表２に示す．実験 A は， FWM&PCA（55次元）と MFCC との組み合わせであり，それぞれの認識結果の確率をストリーム重みとして，0.6：0.4 で重み付けしたものである．この結果，81.5%の認識率が得られた．

実験 B は，FWM&PCA（55 次元）と MFCC，ΔMFCC との組み合わせであり，それぞれの認識結果の確率をストリーム重みとして，0.2：0.4：0.4 で重み付けしたものである．この結果，85.3%の認識率が得られた．

実験 C は，FWM&PCA（55 次元）と MFCC，ΔMFCC との組み合わせであり，それぞれの認識結果の確率をストリーム重みとして，0.0：0.5：0.5 で重み付けしたものである．

実験 D は，MFCC（対数パワーあり１３次元） と ΔMFCC（対数パワーあり１３次元）を結合し，２６次元ベクトルにして認識した場合の結果であり，最も高い 86.7%が得られた．

実験 C を行った目的は，実験 D と比較するためである．実験 C は，MFCC(対数パワーあり１３次元)とΔMFCC(対数パワーあり１３次元)を分けて認識した確率を重み付けしたものであり，実験 D は，これを１つのベクトルにまとめて音素認識したものである．この結果，分離して認識した場合，4.3%の低下が見られる．FWM&PCA と MFCC，ΔMFCC を使ってストリーム重みにより認識した実験 B が実験 D に及ばなかった理由として，３つの特徴量を分離して認識し，ストリーム重み付けを行ったことが原因と推定される．

表 2 ストリーム重みによる認識率

話者	A	B	C	D
MHT	87.9%	91.2%	88.6%	92.2%
MTK	90.2%	91.7%	88.9%	92.7%
MSH	80.9%	84.7%	81.8%	86.0%
MHO	77.1%	80.1%	76.5%	83.0%
MMY	67.8%	76.8%	76.5%	81.3%
MYI	77.8%	82.7%	78.7%	83.4%
FKN	82.6%	86.8%	84.1%	86.6%
FTK	86.2%	88.0%	84.1%	90.1%
FYM	83.2%	85.8%	81.8%	84.7%
FKS	81.2%	85.4%	83.2%	86.7%
平均	81.5%	85.3%	82.4%	86.7%

6. 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，新しい音声特徴量抽出法について提案した．この方法は，時間―周波数平面上の３ｘ３局所領域において，３５種類の局所パターンに対する局所特徴量を計算し，識別効果の高い局所領域の局所特徴量に高い重みを付けて加算し，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は８７５次元と高次元であるため，主成分分析(PCA)により次元を削減することにより，音素認識実験において，MFCC やΔMFCC に比べ，高い認識精度を示すことができた．しかし，FWM&PCA を MFCC やΔMFCC と組み合わせた結果，MFCC とΔMFCC と組み合わせた結果を超えることはできなかった．この理由として，特徴間に相関があるため，独立して認識し，その確率を重み付けして足し合わせることに問題があると考えられる．

今後，３つの特徴をどのように組み合わせるかについて研究を進める予定である．また，提案した音声特徴は，局所特徴の性質から雑音にロバス

トであると考えられるため，雑音下での音声認識に対して有効であるかどうか検証を行なう予定である．また，GMM ではなく，時間情報を扱える HMM に FWM を組み込む予定である．

文　献

[1] S.Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech & Signal Process., vol.ASSP-034, pp.522-529, 1986.

[2] K.Elenius and 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.

[3] T.Nitta, "A novel feature-extraction for speech recognition based on multiple acoustic -feature planes," IEEE proc. ICASSP' 98, pp.29-32, 1998.

[4] T. Nitta,"Feature Extraction for Speech Recognition Based on Orthogonal Acoustic-feature Planes and LDA," Proceedings of IEEE ICASSP'1999, pp.421-424, May 1999.

[5] 新田恒雄，井上雄，正井康之，松浦博，"複合音響特徴平面に基づく音声認識のための局所特徴抽出法，" 電子情報通信学会論文誌，D-II, Vol. J83, No.11, pp. 2341-2349, 2000.

[6] 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.

[7] Yusuke Shinohara, Nobuyuki Otsu, "Facial Expression Recognition Using Fisher Weight Maps," FGR 2004, pp.499-504, 2004.