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.

AAMを用いた唇領域特徴による音声発話認識

駒井　祐人†　　宮本　千琴††　　滝口　哲也†††　　有木　康雄†††

† 神戸大学工学部情報知能工学科　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
†† 神戸大学大学院工学研究科　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
††† 神戸大学自然科学系先端融合研究環　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
E-mail: †{komai,miyamoto}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　雑音環境下で頑健に音声認識を行う手法の一つとして，音声情報に唇動画像情報を併用して認識を行うマルチモーダル音声認識が注目され，近年研究が進められている．マルチモーダル音声認識では音声情報のみでなく画像情報も大きな役割を果たすため，画像に対してどのような特徴量を用いるかが重要な論点となる．従来から音声特徴量は MFCC などある程度定まった特徴量が用いられているのに対し，画像特徴量はその抽出法の違いから，さまざまな特徴量が提案されている．本研究では Active Appearance Model を用いることで唇領域を自動抽出し，座標値と輝度値の情報を含んだ Active Appearance Model の combined パラメータを用いて発話認識することにより，特徴量としての有効性を確認する．

キーワード　唇領域，Active Appearance Model，combined パラメータ，音声と画像の統合

Speech Recognition Based on Lip Area Feature Captured by AAM

Yuto KOMAI†, Chikoto MIYAMOTO††, Tetsuya TAKIGUCHI†††, and Yasuo TARIKI†††

† Department of Computer and System Engineering, Faculty of Engineering, Kobe University　Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
†† Graduate School of Engineering, Kobe University　Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
††† Organization of Advanced Science and Technology, Kobe University　Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
E-mail: †{komai,miyamoto}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　As one of the techniques for robust speech recognition under the noise environment, multimodal speech recognition using lip dynamic scene information together with audio information is attracting attention and the research is advanced in recent years. Since audio information together with visual information plays a great role in multimodal speech recognition, image features you use becomes a significant point. As for the visual features, various features have been proposed because of the difference of the extraction methods while the feature such as MFCC is used to a certain degree for audio features so far. This paper proposes, for spoken word recognition, to utilize c combined parameter extracted by Active Appearance Model applied to a face image including the lip area. Active Appearance Model contains information of the coordinate value and the brightness value as the image feature.

Key words　Lip area, Active Appearance Model, combined parameter, integration of audio and visual

1. ま え が き

現状の音声認識システムは，雑音の少ない環境，もしくは口元にマイクロフォンを設置するような比較的クリーンな環境で使用されており，雑音の多い環境下では，高精度な認識は難しい．一方，音声の発話時に生じる唇の動きは雑音に影響を受けないため，唇の動きから発話内容を認識する読唇は雑音環境下での認識が可能とされている．そこで，雑音環境下で頑健に音声認識を行う手法の一つとして，音声と唇動画像を用いたマルチモーダル音声認識が注目され，近年研究が進められている．マルチモーダル音声認識では，音声と画像の特徴ベクトルを連結する初期統合 [1] [2] や，音声と画像を別々の過程で処理し，

その結果の尤度に重み付けを行う結果統合 [3] [4]，各状態での出力確率の積を求める合成統合 [5] などがある．これらの処理では音声特徴量はもちろん画像特徴量も認識率に大きく影響するため，画像特徴量のみで読唇を行う研究も盛んに行われている．画像特徴量のみで認識する読唇技術に関しては，唇領域を抽出するにあたって，RGB 値分布 [6]，エッジ抽出 [7]，口腔部分の暗色領域利用 [8]，テンプレートマッチングによる抽出 [9]，SNAKE [10]，Active Shape Model [11]，Active Appearance Model [12] [13] [14] [15] など，さまざまな手法が提案されており，特徴量に関しても，主成分スコア [2] [3]，唇の幅や高さ，歯の画素数 [15]，オプティカルフロー [16]，DCT [14] [17] など多くの手法が用いられている．

本研究では，Active Appearance Models (以下 AAM）を用いることで，顔画像から唇領域を自動的に抽出し，唇領域の座標値と輝度値を含んだ特徴量として，AAM の combined パラメータを抽出する．このパラメータに含まれている shape 情報が唇領域の輪郭の動きを，texture 情報が唇領域内の歯など輝度値が大きく変化する部分を表現できると考え，この combined パラメータを用いて HMM を作成し，音声特徴量と統合する方法を提案する．顔領域抽出法としては Haar-like 特徴を用いた AdaBoost 法 [18] を利用し，音声と画像の統合法として，今回は音声と画像のフレームレートの問題 [19] を考慮する必要のない結果統合を用いた．

本論文は次のように構成されている．まず **2.** で 手法の流れについて述べ，**3.** で AAM を用いた特徴量抽出について述べる．**4.** で 認識手法について述べ，**5.** で 216 単語と 100 単語に対する認識結果を示す．最後に **6.** で本論文をまとめる．

2. 処理の流れ

図 1 に全体の簡単な流れを示す．まず，入力動画に対して Haar-like 特徴を用いた AdaBoost 法による顔領域検出を行う．これは AAM による特徴点探索では，特徴座標点の抽出精度が AAM の初期探索点に大きく依存するため，AdaBoost 法で検出した顔領域を AAM の初期探索点として与えることで，特徴点の正確な抽出が行えるためである．次に検出した顔領域に対して AAM を適用し，入力画像と最も類似する画像を生成する AAM のパラメータを決定し，このパラメータを画像特徴量として抽出する．学習では，この特徴量と音声から抽出した特徴量を用いて，HMM を画像と音声で個別に作成する．認識では，画像用の HMM から出力された尤度と音声用の HMM から出力された尤度を統合することで，最終的な認識結果を出力する．

3. 特徴量抽出

3.1 顔領域検出

本研究では，顔領域検出において一般的に用いられている手法であり，安定して高速に顔領域の検出が可能である Haar-like 特徴を用いた AdaBoost 法を利用することで，発話対象者の顔領域の位置を自動的に検出する．この手法では，図 2 に示す矩形領域を図 3 のように当てはめ，白色領域と黒色領域の平均輝度値の差を Haar-like 特徴として抽出し，顔判別に有効な矩形

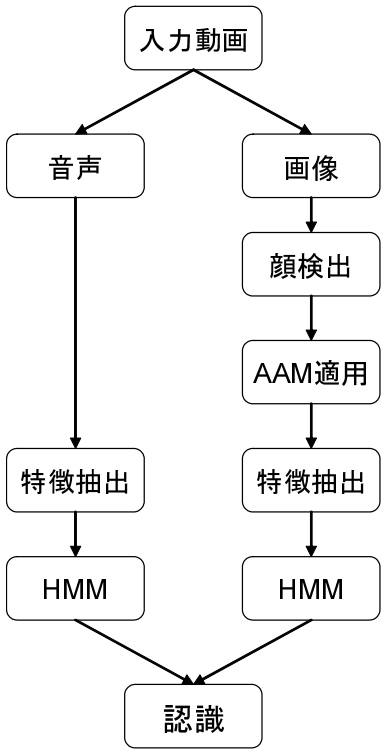

図 1　処理の流れ

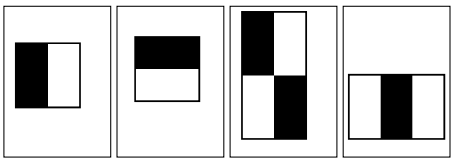
図 2　Haar-like 特徴量

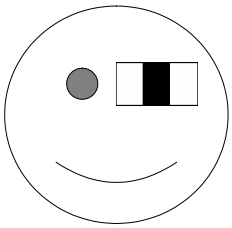
図 3　Haar-like 特徴量の矩形の当てはめ

の位置，種類，縦横比，スケールを弱識別器として AbdaBoost によって学習させる．作成された弱識別器から，顔領域の識別に有効なものを選出し，線形結合することで強識別器を構成し，顔領域を検出する．

3.2 Active Appearance Models

AAM は，Cootes らによって提案された手法であり，特徴点の形状である shape と特徴点の輝度値である texture を主成分分析して部分空間を構成し，比較的低次元なパラメータにより顔モデルを表現する手法である．

顔画像の各点の特徴点座標を並べた shape ベクトルを $\mathbf{s}$ と置き，学習画像に与えられたベクトル $\mathbf{s}$ を正規化することで，学習画像集合から平均形状 $\bar{\mathbf{s}}$ を求める．また，$\mathbf{s}$ の内部の texture を平均形状に正規化し，その輝度値を並べた texture ベクトルを $\mathbf{g}$ とすると，$\mathbf{s}$，$\mathbf{g}$ は，式 (1)，(2) のように与えられる

$$\mathbf{s} = (x_1, y_1, ..., x_n, y_n)^T \quad (1)$$

$$\mathbf{g} = (g_1, ..., g_m)^T \quad (2)$$

ここで，x_i, y_i $(i \leqq n)$ は各特徴点の座標を表している．g_j $(j \leqq m)$ は，平均形状 $\bar{\mathbf{s}}$ に画像を正規化したときの $\bar{\mathbf{s}}$ 内部での各画素の輝度値であり，学習画像集合から平均輝度値 $\bar{\mathbf{g}}$ を求めることができる．$\mathbf{s}$，$\mathbf{g}$ は，$\bar{\mathbf{s}}$，$\bar{\mathbf{g}}$ からの偏差を主成分分析して得られる固有ベクトル $\mathbf{P_s}$，$\mathbf{P_g}$ を用いて，式 (3)，(4) のように表すことができる．

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P_s b_s} \quad (3)$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g b_g} \quad (4)$$

$\mathbf{b_s}$，$\mathbf{b_g}$ はそれぞれ shape パラメータ，texture パラメータと呼ばれ，平均からの変化を表すパラメータであり，これらを変化させることで shape と texture を変化させることができる．また，shape と texture に相関があることから，$\mathbf{b_s}$ と $\mathbf{b_g}$ をさらに主成分分析することで，式 (5)，(6) のように表現できる．

$$\mathbf{b} = \begin{pmatrix} \mathbf{W_s b_s} \\ \mathbf{b_g} \end{pmatrix} = \begin{pmatrix} \mathbf{W_s P_s}^T(\mathbf{s} - \bar{\mathbf{s}}) \\ \mathbf{P_g}^T(\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix} = \mathbf{Qc} \quad (5)$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q_s} \\ \mathbf{Q_g} \end{pmatrix} \quad (6)$$

ここで，$\mathbf{W_s}$ は shape ベクトルと texture ベクトルの単位の違いを正規化する行列，$\mathbf{Q}$ は固有ベクトル，$\mathbf{c}$ は shape と texture の両方を制御するパラメータで combined パラメータと呼ばれる．$\mathbf{c}$ を用いて $\mathbf{s}$，$\mathbf{g}$ を表現すると式 (7)，(8) のようになる．

$$\mathbf{s(c)} = \bar{\mathbf{s}} + \mathbf{P_s W_s}^{-1} \mathbf{Q_s c} \quad (7)$$

$$\mathbf{g(c)} = \bar{\mathbf{g}} + \mathbf{P_g Q_g c} \quad (8)$$

このようにして，パラメータベクトル $\mathbf{c}$ を制御することによって，shape と texture を同時に扱い，顔の変化を表現することが可能となる．

3.3 combined パラメータ

AAM の学習モデルに口の開閉が含まれている画像を用いた場合，図 4 に示すように $\mathbf{c}$ を変化させる事により，多様な唇の動きが表現できることが分かる．$\mathbf{c}$ には唇の詳細な形状と輝度値に関する情報が含まれているため，本研究ではパラメータベクトル $\mathbf{c}$ を画像特徴量として用いることを提案する．$\mathbf{c}$ の抽出法としては，入力画像 $\mathbf{I_i}$ をアフィン変換させて得られる画像を $\mathbf{I_i(W(p))}$ とすると，モデル画像 $\mathbf{g(c)}$ との誤差 $\mathbf{e}$ は式 (9) のようになり，$\mathbf{e}$ が最小となるように $\mathbf{c}$ と $\mathbf{p}$ を最急降下法によって求める．

$$\mathbf{e(c,\ p)} = \|\mathbf{g(c)} - \mathbf{I_i(W(p))}\| \quad (9)$$

ただし，$\mathbf{p}$ はアフィン変換するための拡大縮小，回転，平行移動に関するパラメータであり，$\mathbf{W}$ はアフィン変換を実行する

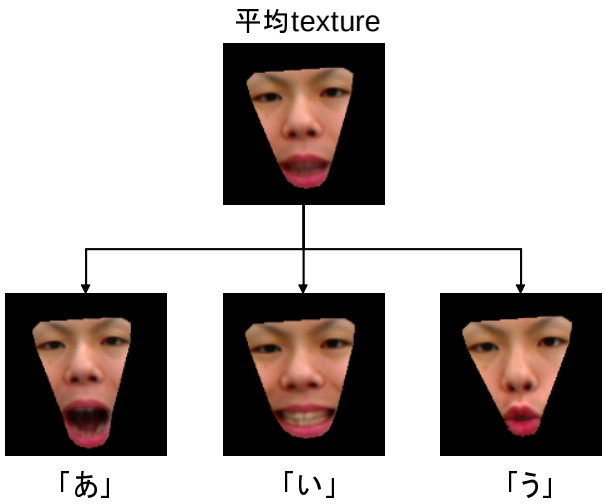

図 4　c パラメータを変化させたときのモデル画像の例（左から順に発話内容「あ」「い」「う」を生成したモデル）

図 5　AAM のモデル構築に用いた 63 点の特徴点

関数である．$\mathbf{c}$ の次元数は shape と texture の主成分分析の累積寄与率が 95% となるように計算しているため，特徴点の個数と学習画像の枚数によって可変である．本研究ではモデルに与えた特徴点は図 5 に示すように，両目，両眉にそれぞれ 8 点，鼻に 11 点，外側の唇輪郭点に対して 12 点，内側の輪郭点に対して 8 点の合計 63 点を与えている．唇以外に特徴点を与えるのは，発話時の口の急激な変化にカメラのフレームレートが追いつかず，画像がぼやけてしまい，AAM による唇領域の抽出精度が劣化するためである．モデルの学習画像を 78 枚用意した結果，$\mathbf{c}$ の次元数は 12 次元となった．さらに，画像のフレームレートは音声の約 3 分の 1 であり，このフレームレートで特徴量抽出を行うと認識率の低下を招く恐れがあるため，フレーム間を 3 次スプライン関数で補間して内挿した．こうして得られた $\mathbf{c}$ と $\mathbf{c}$ の $\Delta, \Delta\Delta$ 係数，計 36 次元を最終的に画像特徴量として使用した．

3.4 追加特徴量

画像特徴量として，c パラメータとの比較を行うため，唇領域に対して 2 次元 DCT を適用する．処理としては，AAM によっ

て唇領域を特定し，その領域において最大幅をもつ上下左右端を求め，唇領域を抽出し，32×32 の正方領域に正規化を行う．その領域に対して $N \times N$ のブロック分割を行い，各ブロックで平均輝度を求め，N^2 個のデータ列 $x(i,\ j)(i,\ j = 0, ..., N-1)$ を求める．これに対して，式 (10) のように 2 次元 DCT を適用する [17] [20]．今回は N の値は 16 とした．

$$X(k_1, k_2) = \frac{4C(k_1)C(k_2)}{N_2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \mathbf{x}(i,\ j) \times \cos\frac{(2i+1)k_1\pi}{2N} \cos\frac{(2j+1)k_2\pi}{2N}$$

$$\begin{cases} C(k) = \frac{1}{\sqrt{2}},\ k = 0 \\ C(k) = 1, \quad k \neq 0 \end{cases} \quad (k_1, k_2 = 0, ..., N-1) \qquad (10)$$

ここで，ブロック分割を行った際の周波数成分の次元は 16×16 の 256 次元となり次元が大きいこと，また，DCT は変換後，低周波成分に情報が集中するという性質から，4×4 の低周波成分 16 次元を切り出し，その Δ と ΔΔ の計 48 次元を特徴量とした．

4. 認 識 手 法

マルチモーダル音声認識ではサブワード型 HMM がよく用いられているが，画像のみでの読唇の研究ではそのほとんどでワード型 HMM が用いられているため，本研究ではワード型 HMM とサブワード型 HMM の両方を比較する．音声特徴量としては MFCC12 次元と対数パワー，及びこれらの $\Delta, \Delta\Delta$ 成分，計 39 次元を用いた．音声と画像の統合法は **1.** で示したように結果統合を行い，最終的な尤度の計算は式 (11) のように行った [4].

$$L_{A+V} = \alpha L_A + (1-\alpha) L_V\ , \quad 0 \leqq \alpha \leqq 1 \qquad (11)$$

ここで L_{A+V} は統合後の尤度，L_A，L_V，は音声と画像それぞれの尤度，α は重みである．

5. 実　　験

5.1 実 験 条 件

本研究では，発話単語として ATR 音素バランス単語 216 語 ×10 セットと，ATR 音素バランス文よりランダムに選出した 100 単語 ×1 セットを用いた．撮影機器は Logicool Qcam Orbit MP で，解像度は 960×720，フレームレートは 30fps，マイクは SONY ECM-PC50 を使用した．

撮影条件として，不特定/特定話者，時期差，顔方位，ぞんざい/ていねいな口調，カメラとの距離，雑音の強さなどがあるが，今回はカメラから約 40cm の距離で固定し，特定話者 1 名に正面顔ではっきりとした口調で発話させた．時期差は考慮しないため全て同一時期に撮影し，雑音は音声抽出の後に，SN 比が 5dB，0dB，-5dB となるよう雑音を加えた．

実験は，216 単語 ×10 セットに対して leave-one-out 法を適用し，9 セットで学習，1 セットを認識して，10 セットの平均を認識率とした（以下 closed 条件）．また，216 単語 ×10 セットで学習した，未知データ 100 単語 ×1 セットを認識する方法

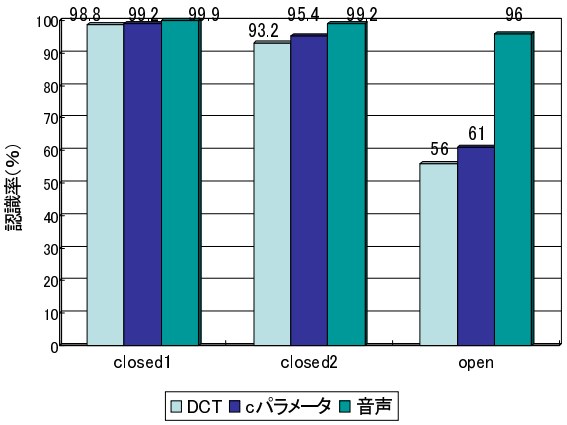

図 6　画像と音声それぞれの特徴量で条件ごとに認識した結果

（以下 open 条件）も行った．ワード型 HMM は closed 条件で状態数 5，混合数 2，サブワード型 HMM は monophone 型で closed 条件と open 条件の両方の実験を行った．closed 条件では状態数 5，混合数 16，open 条件では状態数 5，混合数 2 とした．混合数は実験的に最も良いものを選んだ．

5.2 実 験 結 果

図 6 に画像特徴量と音声特徴量を別々に用いて発話認識した結果を示す．closed1 はワード型 HMM での認識率，closed2 は closed 条件でのサブワード型 HMM の認識率，open は open 条件でのサブワード型 HMM の認識率を表す．closed1 では c パラメータが DCT より 0.4% ，closed2 では 2.2% ，open では 5% 高い認識率を得ている．特に closed1 ではワード型 HMM を用いた読唇の従来研究 [6] [11] [15] [16] よりも高い認識率を得ており，c パラメータが効果的な特徴量であることが確認できる．ただし，open 条件では音声では 96% 認識できているのに対し，c パラメータでは認識率が 61% と closed 条件に比べて低下している．closed1，closed2 の実験が音声，画像ともに認識率が高いことから，音素を学習する際に，子音の学習がうまく行われていない可能性があると考えられる．

音素の特徴は，周囲の音素の影響を受けて大きく変化することが知られており，closed 条件ではテストデータの単語は学習データに含まれている単語のため，closed 条件に適合した音素が学習されているが，open 条件ではテストデータが学習されていない未知データのため，同じ音素でも，周囲の音素の影響で別の音素と認識されている可能性がある．音声が open 条件で認識率が高いのは，音声のフレームレートは画像の約 3 倍であり，音声の方が画像よりも音素の前後の情報をより反映しているためで，画像の 3 次スプライン関数によるフレームレート間の補間では，連続する 2 フレーム間の急激な変化を充分に補間できていないからだと考えられる．

次に，雑音状況下での音声との統合を図るため，SN 比が 5dB，0dB，-5dB となるよう音声に雑音を加え，c パラメータによる HMM の出力尤度と音声による HMM の出力尤度を式 (11) で計算し，音声と画像の重みを 0.1 単位で変化させたときの認識結果を図 7，8，9 に示す．図 7，8，9 はそれぞれ，ワード型 HMM（closed1），closed 条件でのサブワード型 HMM

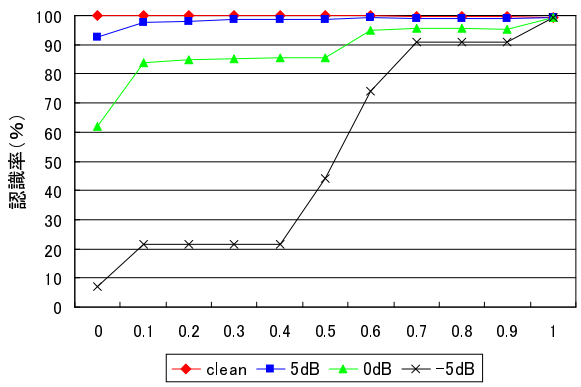

図 7　音声と画像の統合結果 (closed1)

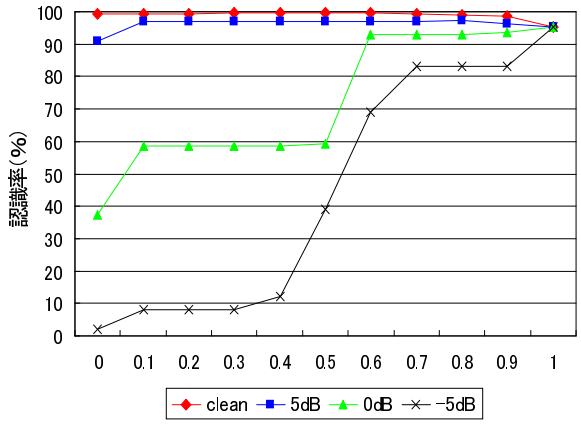

図 8　音声と画像の統合 (closed2)

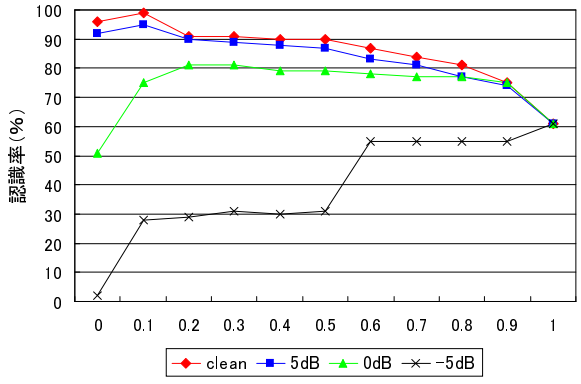

図 9　音声と画像の統合 (open)

(closed2)，open 条件でのサブワード型 HMM（open）の結果であり，横軸は画像の重みを表わしている．重みが 0 のときは音声のみでの認識率，1 のときは画像のみでの認識率である．実験結果を見ると，どの条件でも雑音を加えない clean な環境や SN 比が 5dB の環境では音声も画像も比較的認識率がよいため，重みがどの値でも高い認識を示しており，さらに音声のみの認識率よりもわずかではあるが認識率が改善されている．SN 比が 0dB，-5dB など雑音が大きい環境では音声での認識率が大幅に下がり，特に SN 比が-5dB の状況下では，音声のみではほとんど認識せず，画像の重みを大きくしていくことで認識率

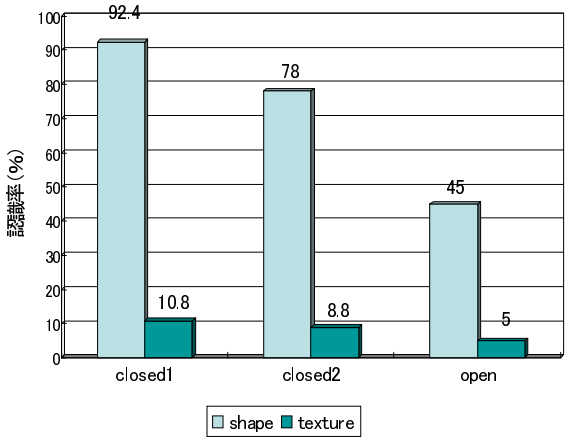

図 10　shape と texture での実験

を改善できていることが分かる．上記の結果より，雑音状況下で，画像情報を取り入れることにより音声情報のみに比べ認識率が改善され，画像情報が有効に働いていることが確認できる．

5.3　combined パラメータの解析

本研究では，c パラメータに含まれている shape 情報が唇領域の輪郭の動きを，texture 情報が唇領域内の歯など輝度値が大きく変化する部分を表現していると考え c パラメータを用いた．c パラメータの shape と texture のそれぞれがどの程度貢献しているのか調べるため，shape と texture の各々のパラメータで実験を行った．HMM の状態数，混合数をはじめ，その他の条件は **5.1** で述べたものと同じであり，shape と texture は式 (7)，(8) で表現したものである．実験結果を図 10 に示す．図 10 を見ると，shape ではどの条件でも，図 6 に示す c パラメータによる認識率の 7 割から 9 割程度の認識率が得られており，shape が唇領域の動きを表現できていると言える．しかし，texture のみでは認識率が非常に低く，shape が c パラメータの中でかなりの重要性を占めていることが分かる．これは，抽出した c パラメータをそのまま使っているため，唇領域以外の texture の部分が認識率の低下を招いている可能性があり，今後，唇領域のみでの c パラメータの抽出法を検討する必要がある．また，図 6 と図 10 を比較すると，texture のみによる認識率は低いが，shape と組み合わせた c パラメータによる認識率は，shape のみの認識率よりもかなり高い．したがって，texture が反映されていないというわけではなく，座標情報と輝度情報を共に用いることで相乗効果が働いている可能性がある．今後，shape と，画像の別の特徴量とを組み合わせる実験が必要であると考えられる．

6.　ま　と　め

本研究では AAM により唇領域を自動抽出し，その際得られた combined パラメータを特徴量として，音声と統合することでその有効性を確認した．また，c パラメータの shape と texture のみで実験を行ったところ，shape は認識率に大きく影響していたが，texture はあまり認識に影響していないことが判明し，唇領域のみでの c パラメータの抽出法や，別の特徴

量との組み合わせなどを考える必要があることが分かった．

本研究では，はっきりとした口調の発話を対象とし，特定話者 1 名での実験であった．今後の課題としては，複数名での認識，c パラメータの改善，音声と画像の新たな統合法，重み最適化手法の検討，自然な口調に対する認識，顔方位のある画像に対する AAM の適用，連続音声認識への展開，などが挙げられる．また今回の実験はデータ数の数から，monophone 型 HMM を選択したが，データ数を増やし triphone 型 HMM を用いることで，さらなる認識率の改善が期待できる．

文　献

[1] Potamianos, G. Graf, H.P. AT&T Labs., Florham Park, NJ, "Discriminative Training Of HMM Stream Exponents For Audio-Visual Speech Recognition", Proc. ICASSP98, Seattle, U.S.A., vol.6, pp.3733-3736, 1998.

[2] 石川剛，澤田裕子，全炳河，南角吉彦，宮島千代美，徳田恵一，北村正，"初期統合によるバイモーダル大語彙連続音声認識"，情報科学技術フォーラム全国大会 pp.203-204, Sep, 2002.

[3] 石川剛, 全柄河, 南角吉彦, 宮島千代美, 徳田 恵一, 北村 正，"音響尤度のリスコアリングによる結果統合を用いたバイモーダル連続音声認識"，音響学会講演集, pp.193-194, Apr. 2003.

[4] 松政宏典，滝口哲也，有木康雄，李義昭，中林稔堯"メタモデルと音響モデルの統合による構音障害者の音声認識"，電子情報通信学会技術研究報告，WIT2008-7，pp.37-42, 2008.

[5] 熊谷建一，中村哲，猿渡洋，鹿野清宏，"HMM 合成を用いたバイモーダル音声認識"，2000 年秋季音講論，pp.111-112, 2000.

[6] 中田康之，安藤護俊，"色抽出法と固有空間法を用いた読唇処理"，電子情報通信学会，Vol.J85-D- ，No12, pp.1813-1822, 2002.

[7] 若杉智和，西浦正英，山口修，福井和広，"色分布間の分離度を用いた唇輪郭抽出"，電子情報通信学会，Vol.J89-D, No9, pp.2025-2032, 2006.

[8] Rainer Stiefelhagen, Uwe Meiger, Jie Yang, "Real-Time Lip-Tracking For LipReading", Eurospeech 97, (1997).

[9] 関岡哲也，横川勇仁，船曳信生，東野輝夫，山田朋弘，森悦秀，"関数合成による唇輪郭抽出法の提案"，電子情報通信学会，Vol.J84-D- ，No.3, pp.459-470, 2001.

[10] 若杉智和，西浦正英，福井和弘，"多次元分布間の分離度を用いたロバストな唇輪郭抽出"，電子情報通信学会技術研究報告，PRMU2003-276, pp.121-126, (2004-3)

[11] Juergen Luettin, Neil A. Thacker, Steve W.Beet, "Visual Speech Recognition using Active Shape Models And Hidden Markov Models"，ICASSP-96, Vol.2, pp.817-820, 1996.

[12] COOTES, T.F., "Active Appearance Model", Proc. European Conference on Computer Vision, Vol2, pp.484-498, 1998.

[13] COOTES, T.F., K Walker, C.J.Taylor, "View-based Active Appearance Models"，Image and Vision Computing 20, pp.657-664, 2002.

[14] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, Final Workshop 2000 Report, Oct. 2000.

[15] 齋藤剛史，久木貢，森下和敏，小西亮介，"複数の口唇領域を用いた単語認識"，画像認識・理解シンポジウム (MIRU2008), IS-17, pp.434-439, 2008.

[16] 大槻恭士，大友照彦，"オプティカルフローと HMM を用いた駅名発話画像認識の試み"，電子情報通信学会技術研究報告，Vol.PRMU2002-124, pp.25-30, 2002.

[17] 山口健，山本俊一，駒谷和範，緒方哲也，奥乃博，"多方向の唇画像を利用した音声認識"，人工知能学全国大会 (JSAI2004), 1E2-02, pp.1-4, 2004.

[18] p.Viola, M. Jones, "Rapid Object Detection Using Boosted Cascade of Simple Features"，In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.1-9, 2001.

[19] 田村哲嗣，石川雅人，速水悟，"マルチモーダル音声認識における音声と画像の同期に関する調査"，電子情報通信学会技術研究報告，SP2008-70, pp.1-6, 2008.

[20] 稲田佳子，肖業貴，尾田政臣，"空間周波数を用いたベクトルマッチングによる顔画像の表情認識"，電子情報通信学会技術研究報告，Vol.101, No.385, pp.25-32, 2001.