唇領域の AAM を用いた発話認識における画像特徴量の音素解析

駒井　祐人†　　宮本　千琴†　　滝口　哲也††　　有木　康雄††

† 神戸大学大学院システム情報学研究科　〒 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　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おり，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．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，わずかではあるが，従来使われてきた DCT や主成分スコアといった特徴量よりも，認識率を改善することができた．また，音声は音素，画像は口形素を用いて統合することで，より精度の高い統合を行うことができた．

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，口形素

Phoneme Analysis of Image Feature in Utterance Recognition Using AAM in Lip Area

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

† Graduate School of System Informatics, 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　Lip-reading, that recognizes the content of the utterance from the movement of the lip enables the recognition under a noise environment. Multimodal speech recognition using lip dynamic scene information together with acoustic information is attracting attention and the research is advanced in recent years. Since visual information plays a great role in multimodal speech recognition, what to select as the image feature becomes a significant point. This paper proposes, for spoken word recognition, to utilize c combined parameter as the image feature extracted by Active Appearance Model applied to a face image including the lip area. Combined parameter contains information of the coordinate value and the intensity value as the image feature. The recognition rate was improved by the proposed method compared to the conventional method such as DCT and the principal component score. Moreover, we integrated with high accuracy the phoneme score from acoustic information and the viseme score from visual information.

Key words　Lip area, Active Appearance Model, Combined parameter, Integration of audio and visual, Viseme

1. はじめに

近年，音声認識技術の発達により，携帯電話での音声による検索機能，音声認識に対応したカーナビゲーションシステムなど，さまざまな音声認識技術が実用化されている．しかし，現在の音声認識技術には，実環境など雑音の大きい状況下では認識性能が著しく低下してしまうという問題点があり，音声認識の実用化に向けて，大きな課題となっている．

一方，人間は発話内容を理解する際，種々の情報を統合的に利用している．音声が聞き取り難い場合，発話者の顔，特に唇の動きに注目して発話内容を理解しようとし，逆に，唇の動きと音声が不一致の場合，唇の動きに影響されて発話内容を誤って理解してしまう．

このように，人間による発話内容の理解には，唇の画像と音声の情報の統合的利用が極めて重要であり，また，

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，従来では頭部固定といった制約があった中，動画像内の発話者の位置に関わらず, 唇領域を自動的に抽出する．また，唇領域の特徴点抽出を行った際に得られる，AAM の combined パラメータ (以下 **c** パラメータ) を特徴量として抽出する．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 **c** パラメータを用いて HMM を作成し，音声特徴量と統合する手法を提案する．顔領域抽出法としては Haar-like 特徴を用いた AdaBoost 法 [18] [19] 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 [20] 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．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．6. で，認識単語の音素正解率と音素正解精度を求めた結果を示す．最後に 7. で本論文をまとめる．

2. 提案手法の流れ

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，特徴座標点の抽出精度が向上するためである．

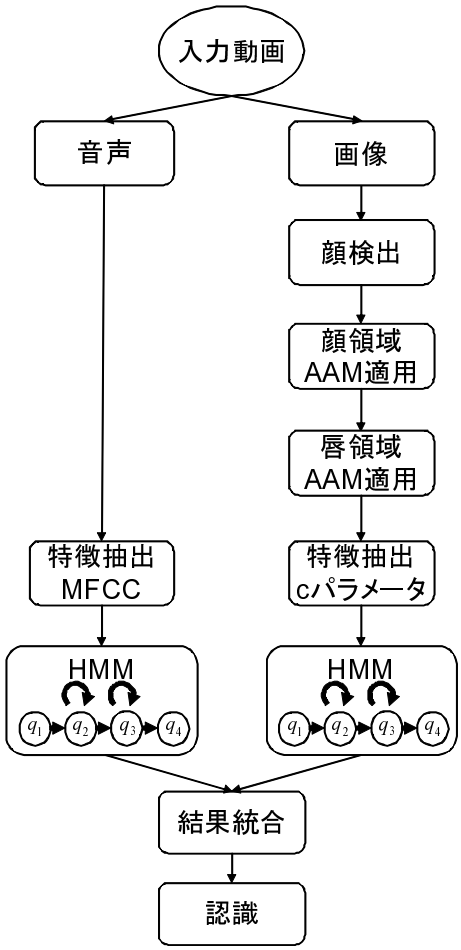

図 1　処理の流れ

次に検出した顔領域に対して AAM を適用する．この AAM は，あらかじめ顔全体に特徴点を与えた学習画像集合から構築した顔全体の AAM と，その唇領域の特徴点から構築した唇領域のみの AAM の 2 種類を用いる．顔全体の AAM を用いる理由として，顔領域を検出しただけでは，唇領域の正確な位置を特定することができず，顔全体の AAM を適用することで，唇領域を正確に抽出することができるためである．一方で，顔全体の AAM から抽出した **c** パラメータを認識パラメータとして用いると，唇領域以外の texture の部分により，認識率の低下を招く恐れがある．そのため，特徴量の不要な次元を取り除く手法も提案されているが [21]，本研究では，唇領域のより正確なパラメータ抽出を行うため，顔全体の AAM と唇領域の AAM の 2 種類を用いた．唇領域の AAM は，顔全体の AAM を適用後，その唇領域を初期探索点として与えることで，正確に特徴点探索を行うことができる．

唇領域の AAM を入力画像に適用した際，入力画像と最も類似する唇領域の画像を生成する **c** 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，同じ動画の音声情報から抽出した音声特徴量を用いて，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 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

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

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

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．

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P_s b_s} \tag{3}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g b_g} \tag{4}$$

$\mathbf{b_s}$，$\mathbf{b_g}$ 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．

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

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q_s} \\ \mathbf{Q_g} \end{pmatrix} \tag{6}$$

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．

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

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

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．

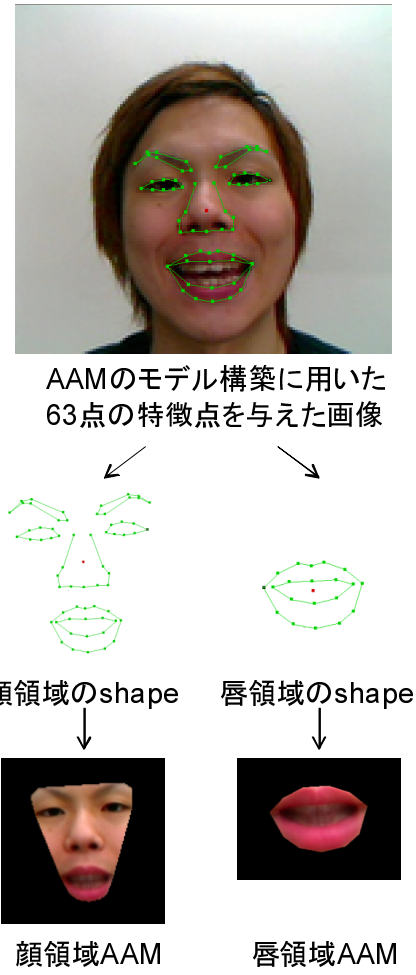

図 2　2 種類の AAM の構築

3.2 モデル構築

本研究では，2. で述べたように，2 つの AAM を用いる．顔全体の AAM は，顔全体に特徴点座標を与えた学習画像集合から，shape 情報とその内側の texture 情報を読み取り構築する．特徴点は図 2 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．唇領域の AAM は，その特徴点座標の唇領域の部分だけを抽出し，唇領域の shape 情報とその内側の texture 情報から構築する．

3.3 Combined パラメータ

AAM の学習データとして，口の開閉が含まれている画像を用いた場合，図 3 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，AAM から生成した画像（これをモデル画像と呼ぶ）$\mathbf{g(c)}$ 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．

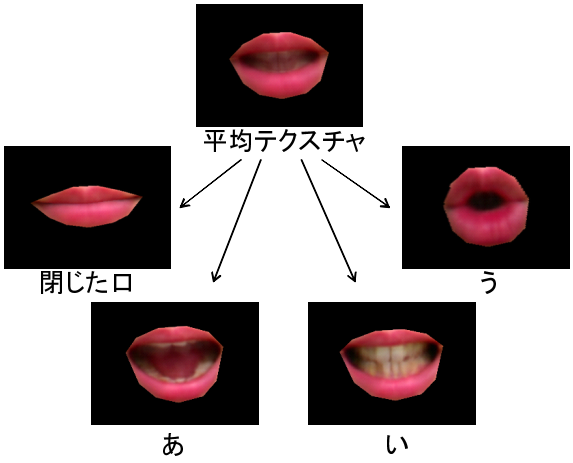

図 3　c パラメータを変化させたときの生成されるモデル画像の例（上から反時計回りに平均テクスチャ，口を閉じたときのモデル，発話内容「あ」,「い」,「う」のモデルによって生成された画像）

$$\mathbf{e(c,\ p)} \ = \ \|\mathbf{g(c)} \ - \ \mathbf{I_i(W(p))}\|^2 \qquad (9)$$

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．モデルの学習画像を 78 枚用意した結果，$\mathbf{c}$ の次元数は 10 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 30 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 [22] と画素値を用いた．処理としては，AAM によって唇領域を特定し，その領域を画面内のサイズの変動の影響を受けないように，縦横の比率を一定にしたまま正方領域に正規化を行う．また，カラー画像のままでは次元数が大きいため，グレイスケール化し，この領域に対して特徴抽出を行った．正方領域の大きさは，画素値は 32×32 とし，特徴抽出後，次元削減のため，主成分分析を適用した．2 次元 DCT は，正規化のサイズが 32×32 の場合と 16×16 の場合で認識率に大きな差がなかったため，計算時間を考慮した 16×16 を選択した．次元数は，画素値は主成分分析適用後，累積寄与率 90% で 10 次元となり，その Δ と $\Delta\Delta$ の計 30 次元を特徴量とした．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 $\Delta\Delta$ 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 の両方を用いる．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 2. 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 (10) のように行った [4].

$$L_{A+V} \ = \ \alpha L_A \ + \ (1-\alpha) L_V \ , \quad 0 \leqq \alpha \leqq 1 \qquad (10)$$

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 画素，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．顔の動きは固定することなく，自然な状態にしてもらった．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（以下 open 条件）も行った．HMM の状態数は 3，混合数 4 とし，ワード型 HMM は closed 条件で，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．混合数は実験的に open 条件の最も良いものを選んだ．

5.2　特徴量ごとでの認識結果

図 4 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．図 4 の 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 の認識率を表している．また DCT は 2 次元 DCT，PCA は画素値を主成分分析したもの，C parameter(face) は顔領域の AAM から抽出した $\mathbf{c}$ パラメータ，C parameter(lip) は唇領域の AAM から抽出した $\mathbf{c}$ パラメータを表して

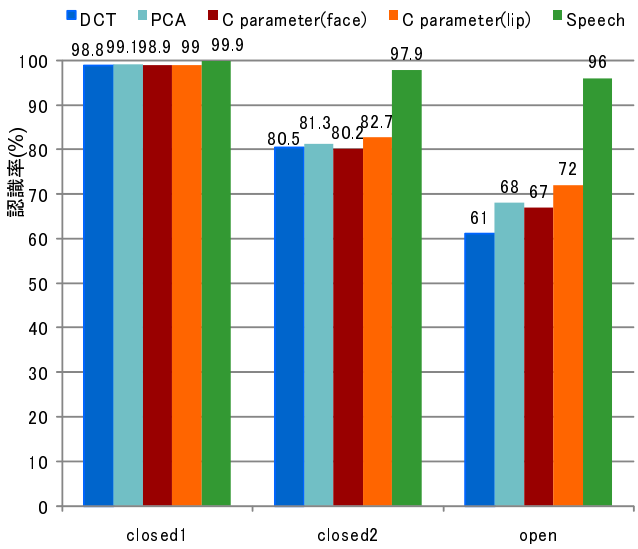

図 4　画像と音声それぞれの特徴量で条件ごとに認識した結果

いる．

特徴量ごとに比較してみると，closed1 では，従来使われてきた特徴量でも，**c** パラメータでも，高い認識率を得られることが分かる．closed2 では，唇領域の **c** パラメータが DCT より 2.2 ポイント，PCA より 1.4 ポイント，顔領域の **c** パラメータより 2.5 ポイント高い認識率を得ている．また，open では DCT より 11 ポイント，PCA より 4 ポイント，顔領域の **c** パラメータより 5 ポイント高い認識率を得ており，closed2，open では，従来使われてきた特徴量よりも，**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．

条件ごとに比較してみると，音声ではどの条件でも高い認識率を得ているのに対し，画像特徴量では，closed2 と open では，closed1 に比べて低下している．

closed1 と closed2 の条件の違いは，ワード型 HMM とサブワード型 HMM の違いであり，サブワード型 HMM では，音素の連結学習が必要なため，ワード型 HMM よりも認識率が低くなっていると考えられる．

open 条件では，closed2 よりもさらに認識率が低くなっている．図 5，図 6 に，混合数を変化させたときの，画像特徴量による HMM の単語認識率の変化を示す．図 5，図 6 を見ると，closed2 では，混合数を増やしていくと認識率が上がっていくことが分かる．混合数を増やすということは，より複雑なモデルになっていくということであり，closed2 では，学習データの単語とテストデータの単語が同じ単語のため，混合数を増やすと，closed2 で用いた単語集合に対して特化し，過学習が起こったモデルになっていると考えられる．しかし，open では混合数を上げると認識率が下がっていく傾向にある．このことから，closed2 では，closed2 の音素環境での音素が学習されており，認識率が open よりも高いと考えられる．この差を埋めるには，あらゆる形の音韻の並びが含まれ

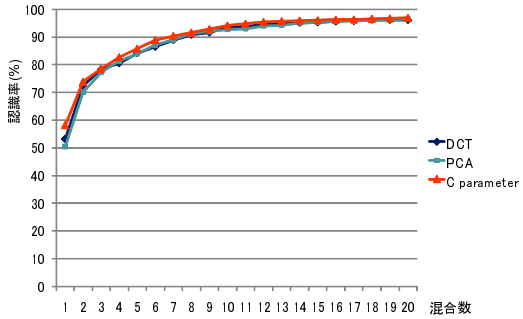

図 5　混合数を変化させたときの認識率の変化 (closed2)

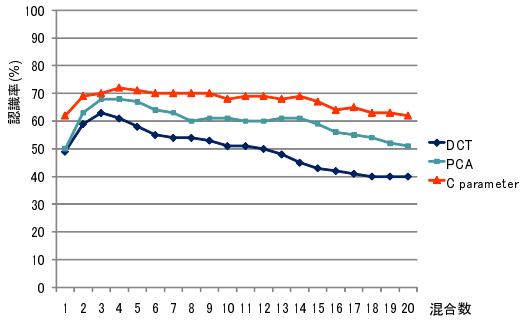

図 6　混合数を変化させたときの認識率の変化 (open)

ているデータベースが必要である．

5.3　音声と画像の統合結果

次に，雑音状況下において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 (10) 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（closed2），open 条件でのサブワード型 HMM（open）の統合結果であり，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

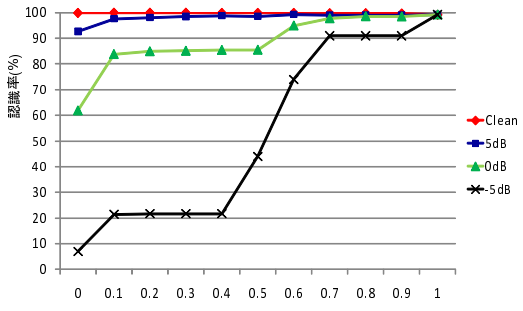

図 7　音声と画像の統合結果 (closed1)

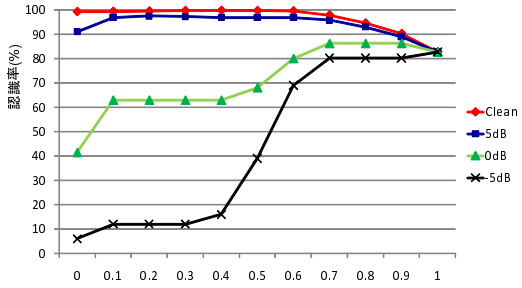

図 8　音声と画像の統合結果 (closed2)

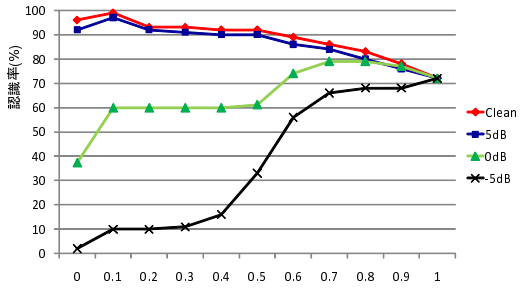

図 9　音声と画像の統合結果 (open)

しており，重みの最適値を取ることで，音声のみの認識率よりも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が改善できている．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．

6.　画像特徴量の音素解析

6.1　連続音素認識

音声と画像とで，音素ごとの認識精度を調べるため，単語を 1 つの文，音素を 1 つの単語とみなした，辞書なし連続音素認識を行った．言語モデルは，子音の次には母音が，母音の次には子音が等確率で出現する，といった日本語の制約に基づく，バイグラムモデルを用いた．音響モデルは，216 単語× 10 セットを HMM で学習したモデル，認識単語は，open 条件の 100 単語を用い，認識は大語彙連続音声認識エンジン Julius を用いた．画像の特徴量は **c** パラメータを用いた．結果を図 10，11 に示す．図 10 は，音声の open での音素の認識率及び誤り傾向を表す confusion matrix，図 11 は，**c** パラメータの open での音素の認識率及び誤り傾向を表す confusion matrix を表しており，図の「挿入」は，挿入誤り数,「欠落」は，削除誤り数を表している．また，音素認識精度を評価するために，音声認識精度に用いられる単語正解率（Correct）と単語正解精度（Accuracy）を用いて，音素を単語とみなしたときの，全音素での音素正解率と音

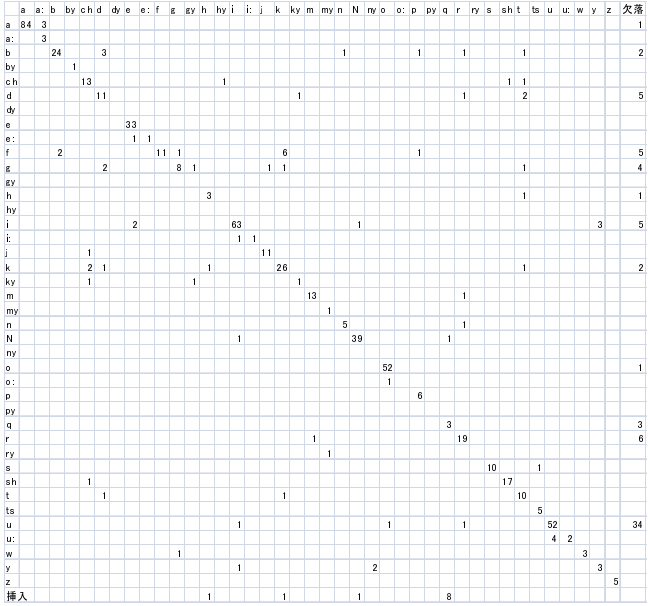

図 10　音声の音素 confusion matrix(open)

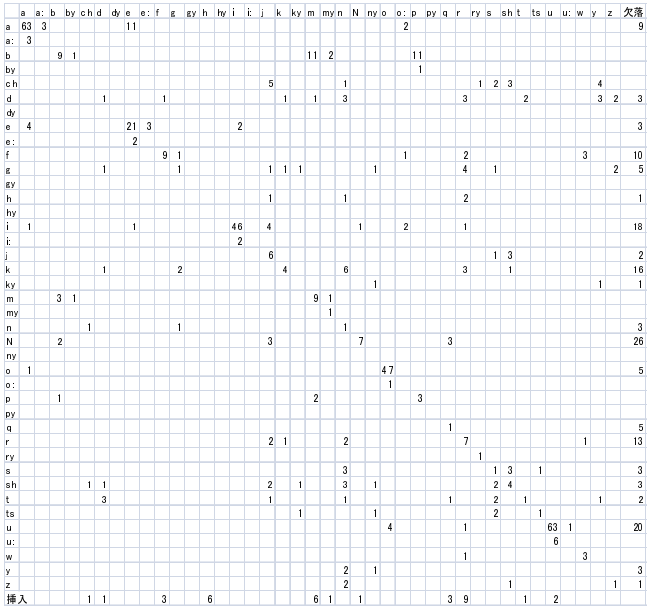

図 11　*c* パラメータの音素 confusion matrix(open)

表 1　音素正解率と音素正解精度（%）

	音声		画像			
	open		open		closed2	
	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct
母音	82.91	82.91	67.81	68.38	65.46	66.21
子音	72.4	75.38	11.85	21.58	37.46	45.87
全体	77.65	79.26	40.74	45.74	52.86	57.05

素正解精度，母音，子音ごとの音素正解率と音素正解精度を求めた．その結果を表 1 に示す．

表 1 を見ると，音声では，母音，子音ともに 8 割前後正解していることが分かる．しかし，画像の open では，母音は 7 割弱正解しているが，子音が 1〜2 割程度しか正解していないため，音素全体で 4 割程度しか正解しないことが分かる．このことから，画像特徴量では子音が

認識しにくいといえる．

6.2 音素の誤認識解析

図 10，11 を見ると，音声では，母音，子音ともに，認識精度が高いことが分かる．**c** パラメータでは，母音は，削除誤りが若干あるものの，ある程度認識できていることが分かる．「i」や「u」に削除誤りが多いが，これは，「ii」や「uu」のように，同じ母音が連続して出現するような箇所で「ii」が「i」，「uu」が「u」と誤認識するような間違いが多かった．これは，「ii」や「uu」と発話してる間，口の形が全く変わらないためであると考えられる．子音を見ると，音声に比べ置換誤りが多く，挿入誤り，削除誤りもかなり多いことが分かる．

挿入誤りはさまざまな音素で起こっているが，その中で多いのは「r」である．「r」は舌の動きだけで表す子音のため，「a」と「ra」では口の形が同じになり，区別できていないと考えられる．

削除誤りの中で多いのは「N」である．発話の前後は口を閉じるようにしているため，単語の最後に「N」が出現すると，口は閉じた形になるため，そこで無音区間と誤認識され，削除誤りになっているといった間違いがいくつかあった．また，「N」が単語内に出現すると，前の母音の口の形を保ったまま「N」を発話するため，「N」にはさまざまな口の形が存在する．そのため，分散が大きい，スパースな特徴であるため，削除誤りが多くなっていると考えられる．

置換誤りもさまざまな音素で起こっている．例えば「k」は，「g」，「n」，「r」といった子音と多く間違っているが，これらは，子音に口の動きがないため，置換誤りを起こしていると考えられる．また，「b」は「m」，「p」といった子音と多く間違っているが，これらは，子音に口の動きはあるが，口の形が同じため，置換誤りを起こしていると考えられる．

6.3 口形素での実験

6.2 で述べたような誤認識が起こる理由は，音素が，音を分ける最小単位であり，これを画像特徴量に当てはめたとしても，「ka」と「ga」のように，口の形が同じであるような音素の区別ができないからであると考えられる．そのため，口の形の最小単位である口形素 (viseme) で closed2 の認識を行い，その口形素認識精度が音声の音素認識精度と近くなれば，それが，音素を用いた時の画像特徴量のほぼ上限値であると考えられる．そのため，本研究では，山口ら [17] を参考に，口形素を定義し，5. で行ったように HMM で認識を行い，音声と統合した．混合数は 12 とし，口形素を用いたときの最もよいものを選んだ．口形素を用いると，「eikyou」「eigyou」のように，音素では区別できても，口形素では 2 つとも「eisyou」となってしまい，区別できないような単語は，統合の際，音声で出力された 2 つの単語に対して同じ尤

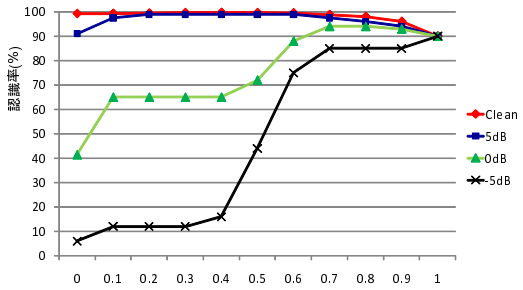

図 12　画像は口形素，音声は音素を用いた時の統合結果 (closed2)

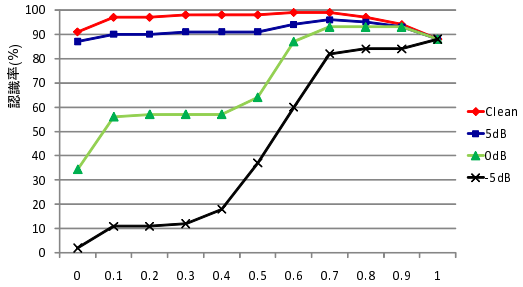

図 13　画像は口形素，音声は音素を用いた時の統合結果 (open)

度で統合した．図 12 に closed2 での統合結果を，図 13 に open での統合結果を示す．

図を見ると，closed2，open ともに，重みが 1 のとき，すなわち画像のみでの認識率が，図 8，9 の，音素で認識したときよりも高いため，全体的に図 8，9 よりも認識率が高くなっていることが確認できる．このことから，音声と画像を結果統合する際，音声は音素で，画像は口形素で認識を行うことで，より精度の高い統合が行えると考えられる．

また，連続音素認識実験と同様に，連続口形素認識実験も行った．図 14 にその confusion matrix を，表 2 に口形素の正解率と正解精度を示す．

	a	i	u	e	o	p	r	sy	w	t	s	y	vf	N	欠落
a	73			10											8
i		57													21
u			68		6										27
e	3	3		26											4
o					50										5
p						54									1
r						1	11	2					3		9
sy		1						44		2		2			2
w			1						26						3
t							5	2	2	15	1	2	5		7
s								5		7	8				1
y							1	2				2			
vf							2	8		7	3		18		17
N						2								10	29
挿入						10	18	1		3		1			

図 14　*c* パラメータの口形素 confusion matrix(open)

表 2 と表 1 を比べると，母音，子音ともに口形素は，音素に比べ認識精度を大幅に改善していることが分か

表 2　口形素正解率と口形素正解精度（%）

	open		closed2	
	Accuracy	Correct	Accuracy	Correct
母音	75.9	75.9	78.21	78.58
子音	47.69	57.85	63.28	68.44
全体	62.54	67.35	71.59	74.08

る．しかし，まだ音声に比べ，closed2 では 10 ポイント程度低い結果となっている．図 14 を見ると，音素と同様,「N」に削除誤りが多い．これは 6.1 で述べたような間違いがほとんどであった．また,「t」はさまざまな口形素と置換誤りを起こしている．口形素の「t」は音素でいう「t」「d」「n」で定義しており，これらはその他の子音と比べて，発話の際に舌の動きを用いるため，これらを識別するためには，舌の動きをとらえることが重要であると考えられる．さらに,「t」を識別できれば,「t」に置換誤りをしている「vf」などの精度も上がると考えられる．

すなわち，画像特徴量ではまだ特徴量に改善の余地があると考えられ，今後，上記で述べたような舌の動きをとらえられるような特徴量，また,「N」をはっきりと認識できる特徴量などを検討する必要があると考えられる．

7. 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，音素の confusion matrix を計算することで，音素を用いたときの，音声情報による音素認識精度と画像情報による音素認識精度の差を明らかにした．さらに，音声は音素，画像は口形素を用いて統合することで，より精度の高い統合ができることを明らかにした．

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 名による実験であった．今後の課題としては，複数名での認識，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，時期差の検討，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] G Potamianos, H.P. Graf, “ 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- , No.12, 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, No.9, 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.

[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 ”，画像の認識・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] 三田雄志，金子敏充，堀修, “ 顔検出に適した Joint Haar-like 特徴の提案 ”，画像の認識・理解シンポジウム (MIRU2005)，pp.104-111，2005.

[20] 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.

[21] 高谷学，滝口哲也，有木康雄, “ 過学習を考慮した AAM パラメータの選択と回帰分析による顔・視線同時推定 ”，画像の認識・理解シンポジウム (MIRU2009)，pp.769-776，2009.

[22] 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.