AAM を用いた顔方位に依存しない発話認識

駒井祐人† 楊楠† 有木康雄†† 滝口哲也††

† 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,yang.nan}@me.cs.scitec.kobe-u.ac.jp, ††{ariki,takigu}@kobe-u.ac.jp

あらまし 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発話認識は，雑音環境下での認識が可能である．しかし，唇情報は，顔が横を向いてしまうと，認識精度が大きく劣化してしまうため，従来のリップリーディングでは正面顔での発話に限定されていることが多かった．本研究では Active Appearance Model を用いて，様々な角度の顔方位を正面に変換してリップリーディングを行う手法を提案する．提案手法では，顔方位に関する回帰モデル式を選択的に扱うことで，正面顔と横顔との変動のミスマッチを抑えつつ，任意の角度で横顔を正面顔に変換することができる．実験では，正面方向の発話のみを学習し，正面，横 15 度，横 30 度の 3 方向の角度において認識を行った結果，3 方向全てにおいて，従来手法と比べ認識精度を改善することができた．

キーワード リップリーディング，Active Appearance Models，Combined パラメータ, 顔方位推定，顔方位変換

1. はじめに

近年，音声認識技術の発達により，スマートフォンを用いた音声による文書作成，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，コンピュータへの新しいインターフェースとして実用化されている．音声入力による文書作成は，高齢者にとって複雑なキーボード操作を伴わないため，情報バリアフリーの点からも意義深い．しかし，現在の音声認識技術には，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ともある．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る．

唇の動きから発話内容を読み取る技術はリップリーディング（読唇）と呼ばれ，聴覚障害者のコミュニケーションの手段の一つとして期待されている．音声認識は雑音に弱く，雑音環境下では認識率が著しく低下するが，リップリーディングは雑音に影響されることがない．また，リップリーディングは，監視カメラなどに収録された会話映像のように，音声が聞き取りにくい場合でも，発話内容の分析を行って，犯罪防止に役立てることなども期待できる．そのため，雑音環境下で頑健に発話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，研究が進められている．

しかし，リップリーディングは，顔が横を向いてしまうと，認識精度が大きく劣化してしまう問題点や，照明変動に弱いという問題点がある．特に顔方位の問題が解決されると，例えば，カーナビゲーションシステムでは，運転者は運転方向を見ながらシステムを扱うことができるし，ロボットとコミュニケーションを行う際に，わざわざロボットに向かって話す必要がなくなる．さらに，監視カメラでは，発話者の顔が正面であこと自体稀である．

そのため，発話者の顔方向が変化しても，発話内容を認識できることが重要であると考えられる．そこで，本研究では，顔方向に依存しないリップリーディング，及びマルチモーダル音声認識の実現を目的としている．

2. 関連研究

顔方位を考慮した研究に関しては，方位毎に発話内容を学習し，方位毎に認識する手法 [1] や，横顔から正面顔への変換行列を学習し，横顔を正面顔に変換して認識する手法 [2] [3]，顔方位が変化しても影響を受けにくい特徴量を提案している手法 [4] [5] などがある．しかし, [1] の手法では，正面だけでなく，横方向に関しても発話内容を学習しなければならないため，膨大な学習データが必要となる. [2] [3] の手法では，変換行列を学習した特定の方位でしか認識することができない. [4] [5] の手法では，方位不変特徴量を提案しているが，認識内容は連続数字の発話に限定されている．

そのため，顔方位に頑健な発話認識を実現するためには，

- 横方向の発話を学習する必要がない
- 特定の方位に限定されない
- 単語もしくは文章単位で認識可能

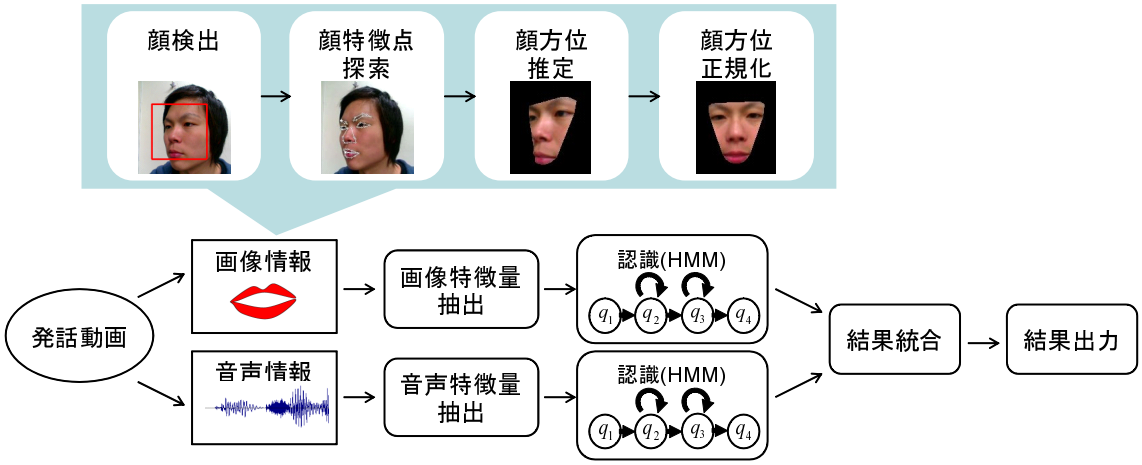

図 1　提案手法の流れ

といったことが必要である．

3. 提 案 手 法

本研究では，Active Appearance Models（以下 AAM）を用いることで，顔が横を向いてしまっていても，発話者の顔方位を推定し，推定した角度の量だけ顔方位を正面に変換することで，任意の顔方位に対して，リップリーディングを行うことができる手法を提案する．その後，音声情報と統合することで，雑音環境下で，顔方位が変動しても認識可能なマルチモーダル音声認識システムを構築する．

提案手法では，AAM により発話者の位置に関係なく，唇領域を自動で抽出することができる．また，唇が見える程度の角度であれば，任意の顔方位に対して発話認識を行うことができる．さらに，顔方向を正面に変換するため，横方向の発話内容を学習する必要がなく，正面での発話を学習するだけで横方向の発話を認識することができる．

以上のことから，本研究の提案内容は，前節で述べた，顔方位を含む発話を認識する際の問題点に対応しているシステムであると言える．

3.1 提案手法の流れ

図 1 に提案手法の流れを示す．発話動画が入力されると，画像情報に対しては，まず，Haar-like 特徴を用いた AdaBoost 法 [6] により，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である．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を向上させることができる．次に検出した顔領域に対して AAM を適用する．AAM によって顔特徴点の探索を行い，入力画像に最も近いモデルのパラメータを生成する．生成したパラメータから，3.6 で述べる手法を用いて，発話者の顔方位を推定する．方位推定後，3.7，3.8 で述べる手法により，推定した角度の量だけ顔方位を正面に正規化することで，横顔を正面顔に変換する．その後，画像特徴量を抽出し，HMM で発話認識を行う．また，発話動画の音声情報に対しては，音声特徴量を抽出し，HMM で発話認識を行う．最後に，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結果統合 [7] することで，最終的な認識結果を出力する．

画像特徴量は，唇の輝度値と特徴点座標の両方を含む，Combined パラメータ（c パラメータ）を用いる．本研究では，顔方位を正確に推定するため顔領域の AAM を用いているが，これにより，c パラメータには目や鼻といった唇以外の顔情報も含まれてしまう．そのため，方位正規化後，c パラメータの中で唇の情報が集約されている次元のみを抽出する．抽出した特徴量の時系列に対して，3 次スプライン補間を適用し，特徴ベクトルのサンプリング数を増やす．さらに，補間後の特徴ベクトルの時系列に対して回帰係数（動的特徴）を求め，これを元の特徴量に加える．音声特徴量は，音声認識において最も広く用いられている MFCC を用いる．

3.2 AAM の構築

AAM は，特徴点の形状である形状と特徴点の輝度値であるテクスチャ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．

学習用の各顔画像に対して特徴点座標を手動で付与し，これを並べた形状ベクトルを $\mathbf{s} = (x_1, y_1, \cdots, x_n, y_n)^T$ と置く．学習画像に与えられたベクトル $\mathbf{s}$ を正規化し，平均形状 $\bar{\mathbf{s}}$ を求める．また，$\mathbf{s}$ の内部のテクスチャを平均形状に正規化し，その輝度値を並べたテクスチャベクトルを $\mathbf{g} = (g_1, \cdots, g_m)^T$ とする．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 $\bar{\mathbf{s}}$ に顔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 (1)，(2) のよ

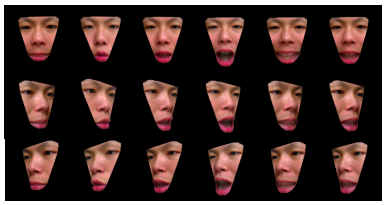

図 2　生成されるモデル画像の例

うに表すことができる．

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P_s b_s} \tag{1}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g b_g} \tag{2}$$

$\mathbf{b_s}$，$\mathbf{b_g}$ はそれぞれ形状パラメータ，テクスチャ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形状とテクスチャを変化させることができる．また，形状とテクスチャ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 (3)，(4) のように表現できる．

$$\mathbf{b} = \begin{pmatrix} \mathbf{W_s b_s} \\ \mathbf{b_g} \end{pmatrix} = \begin{pmatrix} \mathbf{W_s P_s}^T(\mathbf{s} - \bar{\mathbf{s}}) \\ \mathbf{P_g}^T(\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix} = \mathbf{Qc} \tag{3}$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q_s} \\ \mathbf{Q_g} \end{pmatrix} \tag{4}$$

ここで，$\mathbf{W_s}$ は形状ベクトルとテクスチャベクトルの単位の違いを正規化する行列，$\mathbf{Q}$ は固有ベクトル行列，$\mathbf{c}$ は形状とテクスチャ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 (5)，(6) のようになる．

$$\mathbf{s(c)} = \bar{\mathbf{s}} + \mathbf{P_s W_s}^{-1}\mathbf{Q_s c} \tag{5}$$

$$\mathbf{g(c)} = \bar{\mathbf{g}} + \mathbf{P_g Q_g c} \tag{6}$$

このように，パラメータベクトル c を制御することによって，形状とテクスチャを同時に扱い，顔の変化を表現することが可能となる．学習モデルに顔の方位や口の開閉が含まれている画像を用いた場合，図 2 に示すように，$\mathbf{c}$ を変化させる事により，多様な顔の方位や唇の動きが表現できる．

3.3　AAM の探索

入力顔画像 I が与えられた時，構築した AAM の中で，入力顔画像に最も似ている画像を表現する combined パラメータを決定することで，特徴座標点を抽出することができる．

最適な combined パラメータ $\mathbf{c}$ を決定するために，入力顔画像 I をアフィン変換して得られる画像 $I(\mathbf{W(p)})$ と，式 (6) のテクスチャベクトル $\mathbf{g(c)}$ との誤差 $\mathbf{e}$ を式 (7) のように表し，$\mathbf{e}$ 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．

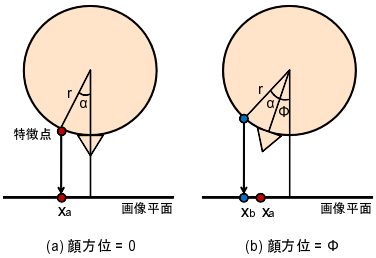

図 3　顔を頭上から見たときの概略図

$$\mathbf{e(c,p)} = \|\mathbf{g(c)} - \mathbf{I(W(p))}\|^2 \tag{7}$$

ただし，$\mathbf{p}$ はアフィン変換をするための拡大縮小，回転，平行移動に関するパラメータ，$\mathbf{W}$ はアフィン変換を実行する関数である．このようにして，入力画像から最適な AAM の $\mathbf{c}$ パラメータが決定される．$\mathbf{c}$ の次元数は，形状とテクスチャ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．

3.4　顔方位正規化のための回帰モデル

本研究では，Cootes らが提案している顔方位の正規化手法 [10] を用いる．Cootes らは，顔水平方向と，AAM の combined パラメータ c との間に相関があると考え，パラメータ c を回帰分析することで顔方位を推定し，目的とする方位へと顔方位を変換する手法を提案している．

図 3 は，顔を半径 r の球体とみなし，頭頂方向から見たときの概略図である．顔の中心から画像平面に垂線を引くと，垂線から角度 α にある顔特徴点は，図 3(a) のように，画像平面上の座標 x_a に射影される．また，顔を図 3(b) のように ϕ だけ動かしたとき，顔特徴点の位置は，画像平面上で x_b に射影される．このときの特徴点座標の変化を Δx とすると，

$$\begin{aligned} \Delta x &= x_b - x_a \\ &= r\sin(\phi+\alpha) - r\sin\alpha \\ &= r\sin\phi\cos\alpha + r\cos\phi\sin\alpha - r\sin\alpha \end{aligned} \tag{8}$$

となる．$\mathbf{c}$ パラメータはこの Δx と比例関係にあると考え，式 (8) の r，α を定数とみなせば，式 (9) のような回帰モデルを設定できる．

$$\mathbf{c} = \mathbf{c_0} + \mathbf{c_1}\cos\phi + \mathbf{c_2}\sin\phi \tag{9}$$

ここで，$\mathbf{c_0}$，$\mathbf{c_1}$，$\mathbf{c_2}$ は学習データから推定される回帰係数ベクトルである．

3.5　回帰係数の学習

前節で設定した式 (9) の回帰係数を学習する方法について述べる．学習画像の数を k 枚とし，それぞれの画像に対して AAM を適用して求めたパラメータ $\mathbf{c^i}$ と，人間

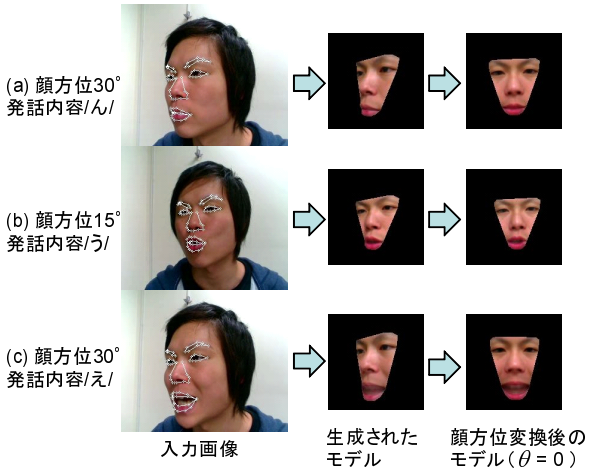

図 4　顔方位の変換例

が向いている方向の真の値 ϕ^i が与えられているとする (i≦k)．これらの学習用パラメータの組を用いて, 最小二乗法により適切な $\mathbf{c_0}$，$\mathbf{c_1}$，$\mathbf{c_2}$ を求める．

式 (9) を行列表現に変形すると，式 (10) になる．

$$\mathbf{c} = (\mathbf{c_0} \quad \mathbf{c_1} \quad \mathbf{c_2}) \begin{pmatrix} 1 \\ \cos\phi \\ \sin\phi \end{pmatrix} \tag{10}$$

式 (10) に対して，学習用パラメータとして与えられた $\mathbf{c^i}$ と ϕ^i を列に並べると，式 (11) のようになり，さらに計算すると式 (12) となる．このようにして得られる回帰係数は最小二乗法で得られるものと一致する．

$$(\mathbf{c^0} \quad \mathbf{c^1} \cdots \mathbf{c^k}) = (\mathbf{c_0} \quad \mathbf{c_1} \quad \mathbf{c_2}) \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \cos\phi^0 & \cos\phi^1 & \cdots & \cos\phi^k \\ \sin\phi^0 & \sin\phi^1 & \cdots & \sin\phi^k \end{pmatrix} \tag{11}$$

$$(\mathbf{c_0} \quad \mathbf{c_1} \quad \mathbf{c_2}) = (\mathbf{c^0} \quad \mathbf{c^1} \cdots \mathbf{c^k})\mathbf{A}^+ \tag{12}$$

ただし，$\mathbf{A}^+$ は行列

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \cos\phi^0 & \cos\phi^1 & \cdots & \cos\phi^k \\ \sin\phi^0 & \sin\phi^1 & \cdots & \sin\phi^k \end{pmatrix}$$

の疑似逆行列で $\mathbf{A}^+ = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$ として求めることができる．

3.6　顔方位推定

前節で求めた回帰係数と式 (9) から，顔方位を推定する方法について述べる．顔方位が未知の入力画像に対して AAM を適用し，特徴点探索によりパラメータ $\mathbf{c}'$ が得られたとき，顔の推定方向 ϕ は式 (9) を変形して，式 (13) のように求めることができる．

$$\begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix} = \mathbf{B}^+(\mathbf{c}' - \mathbf{c_0}) \tag{13}$$

ただし，$\mathbf{B}^+$ は $\mathbf{B} = (\mathbf{c_1} \quad \mathbf{c_2})$ の疑似逆行列で，$\mathbf{B}^+ = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ として計算できる．従って，式 (13) の $\cos\phi$，$\sin\phi$ から，推定角度 ϕ を式 (14) で求めることができる．

$$\phi = \tan^{-1}\left(\frac{\sin\phi}{\cos\phi}\right) \tag{14}$$

3.7　顔方位の正規化

前節で求めた顔方位の推定角度から，顔方位を正面顔に正規化する方法について述べる．顔方位が未知の入力画像に対して AAM を適用し，特徴点探索によりパラメータ $\mathbf{c}'$ が得られ，式 (14) により，顔方位 ϕ が得られたとき，式 (15) の残差ベクトル $\mathbf{c_{res}}$ が考えられる．

$$\mathbf{c_{res}} = \mathbf{c}' - (\mathbf{c_0} + \mathbf{c_1}\cos\phi + \mathbf{c_2}\sin\phi) \tag{15}$$

この $\mathbf{c_{res}}$ は，顔の方位以外の情報，例えば口の動きやまばたきなどの微細な動き情報を含んでいる．ここで，新しい角度 θ に顔方位を変換したい場合，式 (9) と残差ベクトル $\mathbf{c_{res}}$ より，式 (16) が得られる．

$$\mathbf{c_{new}} = \mathbf{c_0} + \mathbf{c_1}\cos\theta + \mathbf{c_2}\sin\theta + \mathbf{c_{res}} \tag{16}$$

式 (16) のパラメータ $\mathbf{c_{new}}$ を用いて，顔方位を正面（$\theta = 0$）に変換した結果を図 4 に示す．

3.8　回帰式の選択

本研究で用いる動画は，顔の変動以外にも発話による唇の大きな変動がある．前節で述べた顔方位変換手法をこのような動画に適用すると，式 (15) の残差ベクトル $\mathbf{c_{res}}$ には唇の動きといった，方位以外の顔の変動が含まれる．その変動を，式 (16) のように戻したい角度 θ における回帰モデル式の値に加えることで，角度 θ で発話している顔画像を表現できる．

しかし，2 次元画像上において，横方向を向いた顔の唇の変動は，正面顔の唇の変動とは見え方が異なっている．すなわち，横方向を向いた顔の唇の変動を，正面顔に加えても，唇が変動した正面顔を正確に表現することができない．横方向の唇の変動をより正確に正面方向の唇の変動に反映させるためには，発話内容によって回帰モデル式を選択できるように拡張する必要がある．こうすることで，正面顔と横顔との変動のミスマッチを抑えることができる．

学習時では，式 (9) の回帰モデル式を式 (17) のように複数個用意する．

$$\mathbf{c^j} = \mathbf{c_0^j} + \mathbf{c_1^j}\cos\phi + \mathbf{c_2^j}\sin\phi \tag{17}$$

j は回帰式の番号であり，各回帰式は，ある特定の音素を発話している画像のみで学習される．例えば，/あ/という音素を考えると，音素/あ/を表現する回帰式は，複数の方向で/あ/と発話している画像のみで学習される．その結果，回帰式は，/あ/と発話している画像を表現する

c パラメータで構成される．本研究では，音素の中でクラス分けとして最も大別される母音（/あ/,/い/,/う/,/え/,/お/）と，口を閉じた際の発音/ん/の 6 種類の回帰モデル式を用意した．

方位変換では，まず，発話動画に対して AAM を適用し，あるフレームにおいて，パラメータ $\mathbf{c_{in}}$ を求める．次に，パラメータ $\mathbf{c_{in}}$ より式 (14) によって顔方位 ϕ を推定する．顔方位の推定においては，3.6 で述べたように 1 つの回帰モデルを用いている．次に，異なる発話に対応する各回帰モデル式 (式 (17)) において，方位角度 ϕ における値と，$\mathbf{c_{in}}$ とを比較する．その中で最も距離が近い回帰式を，式 (18) のように選択する．

$$\hat{\mathbf{j}} \;=\; \underset{\mathbf{j}}{\operatorname{argmin}} ||\mathbf{c^j}(\phi) \;-\; \mathbf{c_{in}}|| \qquad (18)$$

その後，3.7 節で述べたように顔方位を変換する．すなわち，j が指定されている式 (17) の回帰モデルの係数を，式 (16) に代入して，正面顔に変換する．

3.9 特徴量抽出

認識を行うための画像特徴量は，従来，主成分スコア [11] や DCT [12] といった画素ベースの特徴量や，唇の幅や高さ [13] といった輪郭ベースの特徴量などが用いられてきた．また，画素情報と輪郭情報の両方を用いたもの [14] [15] なども，良好な精度が得られている．そのため，本研究では, [15] などで用いられている，AAM のパラメータ（c パラメータ）を画像特徴量として用いる．

本研究では，顔方位を正確に推定するため，顔領域の AAM を用いているが，顔領域の AAM を用いると，c パラメータには唇以外の顔情報も含まれてしまう．また，AAM は主成分分析を行なっているため，特定の次元にある程度の情報が集約される．そのため，方位正規化後，c パラメータの唇の情報が集約されている次元を抽出する．

次元の抽出法は，c パラメータの次元の全組み合わせを求め，発話認識を行って，最も認識精度が高かった次元の組み合わせを選択している．

3.10 音声と画像の統合法

音声と画像の統合法は，音声と画像の特徴ベクトルを連結する初期統合 [16] 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 [7] などがある．また，マルチストリーム HMM や，非同期性を考慮する Coupled HMM [17] などを用いている手法も多い．

唇の動きは通常，音声に先行することが知られており，音声と画像に時間的なずれが生じてしまう．そこで，本研究では，これらを考慮する必要のない結果統合を用いた．その計算式は式 (19) のようになる．

$$L_{A+V} \;=\; \alpha L_V \;+\; (1-\alpha)L_A\ , \quad 0 \leqq \alpha \leqq 1 \qquad (19)$$

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．

4. 実　　験

提案手法の有効性を確認するため，評価実験を行った．

4.1 実 験 条 件

4.1.1 AAM のモデル条件

本研究では，AAM のモデルの構築の際，モデルに与えた特徴点は両目，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．目や鼻など，発話中に大きな変化がなく，安定して抽出が行える領域にも特徴点を与えることで，唇領域を安定して抽出することや，顔方位を正確に推定することが可能となる．モデルの学習画像として，被験者 2 名でそれぞれ 108 枚 (正面，右 15 度，右 30 度でそれぞれ 36 枚) を用意し，PCA の累積寄与率を 95% としてモデルを構築した結果，被験者 1 の **c** パラメータの次元数は 5 次元，被験者 2 の **c** パラメータの次元数は 9 次元となった．回帰モデル式の回帰係数を学習する画像は，AAM の学習画像を全て用いた．

4.1.2 発話実験条件

本研究では，発話単語として，全音素をできるだけ均等に含み，かつ多様な発話内容を含む ATR 音素バランス単語 216 語と，ATR 音素バランス文よりランダムに選出した 100 単語を用いた．解像度は 320×240 画素で，フレームレートは 30fps である．

撮影条件として，不特定/特定話者，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，話者 2 名（男性 1 名，女性 1 名）に正面顔，右 15 度，右 30 度ではっきりとした口調で発話させた．左方向を考慮しないのは，人間の顔をある程度左右対称と考えると，右方向での認識精度と左方向での認識精度は同程度と考えたためである．雑音は音声抽出の後に，SN 比が 20dB，0dB となるよう雑音を加えた．

実験は，正面方向の 216 単語 ×9 セットを学習セットとし，正面方向の 216 単語 1 セットと，横 15 度，横 30 度における 216 単語の発話 1 セットをテストセットとして認識し，話者 2 名による認識率の平均を認識率とした（以下 closed 条件）．また，正面方向の 216 単語 ×10 セットを学習セットとし，学習セットには含まれていない未知データ 100 単語で正面，横 15 度，横 30 度における発話 1 セットをテストセットとして認識した（以下 open 条件）．これは，学習にない単語を認識できるか否かを評価するためである．HMM の状態数，混合数は実験的に最も良い値を選び，単語を学習単位としたワード型 HMM では 15 状態，2 混合，音素を学習単位としたサブワード型 HMM では 5 状態 16 混合とした．

画像特徴量は，各々の被験者から c パラメータ 2 次元

を抽出し，抽出した 2 次元とその時間的な回帰係数である Δ，ΔΔ 係数の計 6 次元を用いた．1 次回帰係数 Δ は，参照している時間の前後 2 フレームを含んだ計 5 フレーム成分より各次元ごとに算出し，2 次回帰係数 ΔΔ は，1 次回帰係数 Δ の結果に対して，再び回帰係数を求めることにより算出している．この 2 次元は，3.9 で述べた c パラメータの全次元の組み合わせの中から求めたものである．また比較として, [5] で提案されている，Minimum Cross-Pose Variance(MCPV) mask を用いた．MCPV mask は，リップリーディングにおいて最も広く利用されている特徴量であり，DCT 係数の中で顔方位が変動しても影響を受けにくい次元を抽出した特徴量である. [5] では，正面方向の発話のみを学習し，様々な横方向を向いて発話した内容を認識しており，顔方位を含む発話認識の従来手法の中で，最も実用的な手法の一つであると考えられる．

本研究では，MCPV mask20 次元と Δ，ΔΔ 係数の計 60 次元の特徴量を比較特徴量として用いた．画像特徴量は全て，フレーム間を 3 次スプライン関数で補間して内挿し，アップサンプリングを行った．音声特徴量は，MFCC12 次元とこれらの Δ，ΔΔ 成分，計 36 次元を用いた．

4.2 実験結果

表 1　画像特徴量のみでの実験結果 (%) (Closed1)

	0 °	15 °	30 °
C パラメータ (顔方位変換なし)	90.59	17.71	1.55
MCPV mask (従来手法)	89.44	40.31	5.25
C パラメータ (顔方位変換)	90.21	69.04	49.57
C パラメータ (回帰式選択)	89.19	69.63	60.32
MCPV mask (顔方位変換)	87.67	68.4	56.2

表 2　画像特徴量のみでの実験結果 (%) (Closed2)

	0 °	15 °	30 °
C パラメータ (顔方位変換なし)	80.67	13.39	1.3
MCPV mask (従来手法)	74.23	29.01	2.67
C パラメータ (顔方位変換)	78.67	54.32	42.35
C パラメータ (回帰式選択)	79.56	54.72	49.37
MCPV mask (顔方位変換)	74.84	52.34	47.61

表 1，表 2，表 3 に画像特徴量を用いて，単独で発話認識（リップリーディング）を行った結果を示す．実験は，正面顔からの発話のみを学習し，各方向の顔画像で発話認識を行っている．Closed1 はワード型 HMM による認識率，Closed2 は closed 条件でのサブワード型 HMM による認識率，Open は open 条件でのサブワード型 HMM による認識率を表す.

表 3　画像特徴量のみでの実験結果 (%) (Open)

	0 °	15 °	30 °
C パラメータ (顔方位変換なし)	60	15	3
MCPV mask (従来手法)	60	28	5
C パラメータ (顔方位変換)	59	51	43
C パラメータ (回帰式選択)	59	55	49
MCPV mask (顔方位変換)	59	56	49

表中，C パラメータ (顔方位変換なし) は，c パラメータを用いて，顔方位変換を行わずに認識を行った結果であり，C パラメータ (顔方位変換) は，3.7 で述べた顔方位の変換手法を用いて顔方位を変換し，発話認識を行った結果である．C パラメータ (回帰式選択) は，3.8 で述べた，顔方位の変換手法を拡張した手法で顔方位を変換し，発話認識を行った結果である．MCPV mask(従来手法) は，従来手法である MCPV mask を用いて認識を行った結果であり，MCPV mask(顔方位変換) は，MCPV mask を c パラメータに入れ替えて，3.8 で述べた手法により顔方位を変換し，発話認識を行った結果を示す.

4.2.1 正面方向の認識結果

まず，正面方向（0 °）のみを考える．特徴量を比較してみると，closed1 では c パラメータ（顔方位変換なし）が MCPV mask より 1.15 ポイント，closed2 では 6.44 ポイント高い認識率を得ており，本研究で用いる 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．

正面方向（0 °）を HMM のタイプごとに比較してみると，closed2 の音素サブワード HMM では約 80% の認識率であるのに対して，closed1 のワード型 HMM では約 90% の認識を示しており，ワード型 HMM の方が調音結合の影響を吸収していることが分かる．次に，サブワード型 HMM を用いた open 条件では，認識率が closed2 の音素サブワード型 HMM に比べて低下している．これは，学習で用いた単語に含まれる音素環境と，認識する単語に含まれている音素環境が異なっているためであると考えられる．

音素の特徴は，周囲の音素環境の影響を受けて大きく変化することが知られている．closed 条件ではテストデータの単語と学習データの単語が同じであることから，音素環境が一致している．しかし，open 条件では，テストデータは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環境の影響で，別の音素と認識されている可能性がある．例えば，同じ/い/という音素でも，音素の繋がりを考えると，音素/あ/と繋がった/い/（/

あい/）と，音素/え/と繋がった/い/（/えい/）では，/い/という口の形に到達するまでの口の形が大きく異なる．従って，closed 条件では，同じ単語を認識していることから，同じ音素の並びが学習されているが，open 条件では，異なる音素の並びを持つ単語を認識しているため，周囲の音素環境の影響で別の音素と認識され，精度が下がったものと考えられる．Open 条件における精度を向上させるためには，様々な音素の並びを triphone で学習させる必要があると考えられる．

4.2.2 15 °，30 °方向の認識結果

次に，15 度，30 度について考える．c パラメータ（顔方位変換なし）の認識率を見ると，横 15 度，横 30 度では，正面方向と比べ，著しく精度が下がっている．これは，正面方向の発話のみで学習しているため，横方向を向いて発話認識を行うと，カメラから見た時の唇の形が変わってしまうため，認識精度が下がっていると考えられる．しかし，横 15 度では，従来手法である MCPV mask の精度が比較的劣化が少ない．本研究では，この値を比較のベースとして考える．

C パラメータ (顔方位変換なし) と C パラメータ (顔方位変換) を比較すると，Closed1 では正面方向は 0.38 ポイント精度が下がったが，横 15 度で 51.33 ポイント，横 30 度で 48.02 ポイント精度が改善されている．Closed2 では正面方向は 2 ポイント精度が下がったが，横 15 度で 40.93 ポイント，横 30 度で 41.05 ポイント精度が改善されている．Open では正面方向は 1 ポイント精度が下がったが，横 15 度で 36 ポイント，横 30 度で 40 ポイント精度が改善されている．

C パラメータ (顔方位変換) と MCPV mask(従来手法) を比較すると，Closed1 では正面方向は 0.77 ポイント，横 15 度で 28.73 ポイント，横 30 度で 44.32 ポイント精度が改善されている．Closed2 では正面方向で 4.44 ポイント，横 15 度で 25.31 ポイント，横 30 度で 39.68 ポイント精度が改善されている．Open では，正面方向は 1 ポイント精度が下がったが，横 15 度で 23 ポイント，横 30 度で 38 ポイント精度が改善されている．以上のことから，顔方位を変換して発話認識を行う手法の有効性が確認できる．

さらに，C パラメータ (顔方位変換) と C パラメータ (回帰式選択) を比較すると，Closed1 では正面方向は 1.02 ポイント精度が下がったが，横 15 度で 0.59 ポイント，横 30 度で 10.75 ポイント精度が改善されている．Closed2 では正面方向で 0.89 ポイント，横 15 度で 0.4 ポイント，横 30 度で 7.02 ポイント精度が改善されている．Open では，横 15 度で 4 ポイント，横 30 度で 6 ポイント精度が改善されている．そのため，本研究で提案する C パラメータ (回帰式選択) 手法の有効性も確認できる．

また，顔方位を変換して MCPV mask を用いた結果，顔方位を変換しない場合と比べ，認識精度が飛躍的に改善されている．特に，open 条件での 15 度の認識率は，C パラメータ (回帰式選択) に比べ，1 ポイント精度が改善されている．

最も精度が高かった C パラメータ (回帰式選択) においても，横方向の認識精度は正面方向の精度に及ばない．これは，横方向を向いた際に，AAM の特徴点探索の精度が悪くなることなどが，精度劣化の理由として考えられる．この差を埋めるためには，横を向いた時に生じる，特徴点や情報量の欠如などに対処していく必要がある．ただし，各手法の比較結果から考えると，C パラメータ (回帰式選択) 手法の優位性が分かる．

4.2.3 特徴量の統合結果

前節の結果では，全体的に C パラメータ (回帰式選択) の認識率が高かったが，MCPV mask(顔方位変換) の精度の方が高い場合もあった．そのため，c パラメータと MCPV mask の特徴量を連結して新たなベクトルとし，HMM で結合して認識を行った (初期統合)．その結果を表 4 に示す．

表 4　C パラメータと MCPV mask の統合結果 (%)

	0 °	15 °	30 °
Closed1	90.21	69.78	60.75
Closed2	76.79	54.44	49.95
Open	61	58	49

前節の結果と比較すると，C パラメータ (回帰式選択) に比べて，わずかながら精度が改善されている．その他，CA や通常の DCT との特徴量の組み合わせ方も行ったが，表 4 より改善される組み合わせはなかった．

4.3　画像と音声の統合結果

次に，雑音状況下において音声との統合を行った．SN 比が 20dB，0dB となるよう音声に雑音を加え，Open 条件において，画像特徴量による HMM の出力尤度と音声による HMM の出力尤度を，式（19）により統合した．音声と画像の重み α を 0.1 単位で変化させたときの認識結果を，図 5，6 に示す．Closed 条件では画像特徴量のみである程度認識できるため，統合結果の図には載せていない．

図 5，6 はそれぞれ 20dB，0dB における統合結果を示しており，横軸は画像の重みを表わしている．重みが 0.0 のときは音声のみでの認識率，1.0 のときは画像のみでの認識率である．グラフ上の数字は，認識率が最も良かった値を表している．

実験結果を見ると，20dB では，音声は 70% 前後の認識率が得られている ($\alpha = 0.0$ の場合)．画像のみで認識した場合は，70% に満たない認識率であるが，音声と画像を統合し，重みの最適値をとることで，音声のみの場合よりも精度が飛躍的に改善されている．

一方，0dB では，音声では 20% 程度しか認識できていない．しかし，この条件下においても，重みの最適値をとることで精度が改善されている．

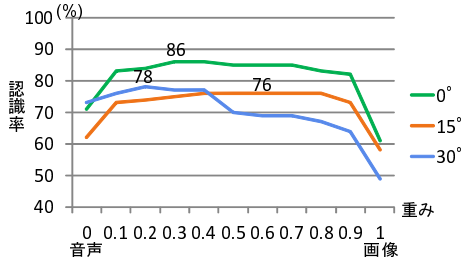

図 5　音声と画像の統合結果（20dB）

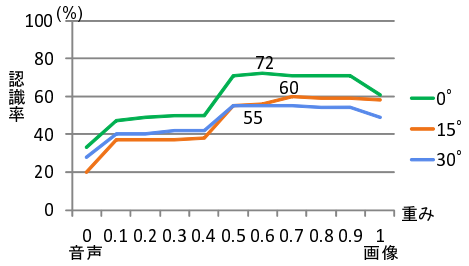

図 6　音声と画像の統合結果（0dB）

以上の結果より，提案手法では，雑音状況下で，顔方位が変動しても，精度劣化を低減できることが分かった．

5. ま と め

本研究では，Active Appearance Models を用いることで，様々な角度の顔方位を正面に変換してリップリーディングを行う手法を提案し，3 方向の角度において，その有効性を確認した．

提案手法では，顔方位に関する回帰モデル式を選択することで，正面顔と横顔との変動のミスマッチを抑えることができる．

実験では，正面方向の発話のみを学習し，正面，横 15 度，横 30 度の 3 方向の角度において認識を行った結果，3 方向全てにおいて，従来手法と比べ認識精度を改善することができた．また，音声と画像を統合することで，雑音環境下において顔方位が変動しても，発話内容を認識することができた．

本研究では，はっきりとした口調の発話を対象とした．今後の課題としては，垂直方向の顔方位変動への対応，不特定話者での認識，重み最適化手法の検討，自然な口調に対する認識，AAM による顔方位のより正確な正規化法，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関係から，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] 山口健，山本俊一，駒谷和範，尾形哲也，奥乃博，“ 多方向の唇画像を利用した音声認識 ”，人工知能学会全国大会 (JSAI2004)，1E2-02，pp.1-4，2004.

[2] P. Lucey, G. Potamianos, and S. Sridharan,“ A Unified Approach to Multi-Pose Audio-Visual ASR ”, IEEE International Conference on Acoustics, Speech, and Signal Processing 2007(ICASSP 2007), pp.650-653, 2007.

[3] P. Lucey, G. Potamianos, and S. Sridharan,“ Visual Speech Recognition Across Multiple Views ”, Visual Speech Recognition; Lip Segmentation and Mapping, 2009.

[4] Adrian Pass, Jianguo Zhang, Darryl Stewart,“ Feature Selection for Pose Invariant Lip Biometrics ”, IEEE International Conference on Acoustics, Speech, and Signal Processing 2010(ICASSP 2010), pp.1165-1168, 2010.

[5] “ An investigation into features for multi-view lipreading ”, International Conference on Image Processing 2010(ICIP2010), pp.2417-2420, 2010.

[6] 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.

[7] Ashish Verma, Tanveer Faruquie, Chalapathy Neti, Sankar Basu, Andrew Senior,“ Late Integration In Audio-Visual Continuous Speech Recognition ”, In Automatic Speech Recognition and Understanding, 1999.

[8] 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.

[9] T.F.Cootes,“ Active Appearance Models ”, Proc. European Conference on Computer Vision, Vol2, pp.484-498, 1998.

[10] T.F. Cootes, KWalker, and C.J. Taylor,“ View-based active appearance models, ”, in Image and Vision Computing 20(2002), pp.657-664, 2002.

[11] M.J. Tomlinson, M.J. Russell, and N.M. Brooke,“ Integrating audio and visual information to provide highly robust speech recognition ”, IEEE International Conference on Acoustics, Speech, and Signal Processing 1996(ICASSP 1996), pp.821-824, 1996.

[12] He Jun and Zhang Hua,“ Research on visual speech feature extraction, ”, International Conference on Computer Engineering and Technology 2009(ICCET 2009), pp.499-502, 2009.

[13] Takami Yoshida and Kazuhiro Nakadai, “ Audio-visual speech recognition system for a robot ”International Conference on Auditory-Visual Speech Processing 2010(AVSP 2010), pp.8-13, 2010.

[14] 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 2008(MIRU2008), IS-17, pp.434-439, 2008.

[15] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong, and Richard Bowden,“ Improving visual features for lipreading ”, International Conference on Auditory-Visual Speech Processing 2010(AVSP 2010), pp.142-147, 2010.

[16] Koji Iwano, Satoshi Tamura, and Sadaoki Furui, “ Bimodal Speech Recognition Using Lip Movement Measured by Optical-Flow Analysis ”, International Workshop on Hands-Free Speech Communication 2001(HSC2001), pp.187-190, 2001.

[17] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao and Kevin Murphy,“ A Coupled HMM for Audio-Visual Speech recognition ”, IEEE International Conference on Acoustics, Speech, and Signal Processing 2002(ICASSP 2002), pp.2013-2016, 2002.