論　文

隠れマルコフモデルを用いた視覚音声認識のための正規化学習

南角　吉彦[†]　　徳田　恵一[†]　　北村　正[†]　　小林　隆夫[††]

Normalized Training for HMM-Based Visual Speech Recognition

Yoshihiko NANKAKU[†], Keiichi TOKUDA[†], Tadashi KITAMURA[†], and Takao KOBAYASHI[††]

あらまし　本論文では，視覚音声認識のための連続密度 HMM（Hidden Markov Model）のパラメータ推定法について述べる．これまでの視覚音声認識の研究は，大きく画像ベース法とモデルベース法の二つに分類することができる．画像ベース法は，原画像の画素値にサブサンプリングや主成分分析などの何らかの前処理を施したものを特徴ベクトルとして用いる手法である．しかし，唇の位置や大きさ，照明条件などが認識率に直接的な影響を及ぼすため，これらの正規化が重要な要素技術となる．従来の正規化手法は，HMM とは独立に何らかの基準を設け，学習の前に正規化を行うのが一般的であった．本論文では，ML（Maximum Likelihood）基準による正規化を考え，唇の位置，大きさ，傾き，平均輝度，コントラストなどの正規化プロセスがモデルの学習と統合された正規化学習法を提案する．提案法は，EM（Expectation Maximization）アルゴリズムに基づいて定式化されており，正規化学習の繰返しにより学習データに関するゆう度が単調に増加することが保証されている．また，M2VTS データベースを用いた単語認識実験により提案法の有効性を示す．

キーワード　視覚音声認識，バイモーダル音声認識，隠れマルコフモデル，正規化学習，EM アルゴリズム

1. まえがき

音声は，人間にとって最も重要な意思伝達手段であることから，計算機による音声認識の研究が盛んに行われている．これらの研究では，周囲の様々な雑音による認識率の低下が問題となるが，その対処法の一つとして，聴覚情報に加えて視覚情報を利用するバイモーダル音声認識が注目を集めている．人間にとって聴覚情報と視覚情報は相互に大きな影響を与えており，McGurk 効果と呼ばれる現象がよく知られている [1], [2]．この現象は，異なる発声の音声と唇動画像を組み合わせ，被験者に視聴させた場合，被験者は音声と画像のどちらとも異なる発声をしているように知覚するという現象である．このように，人間は聴覚情報と視覚情報を統合することによりロバストな音声認識を可能にしていると考えられ，計算機による音声認識においても視覚情報が有用である可能性が高い．以上のような背景から，本論文では，バイモーダル音声認識を実現するための，視覚情報のみによる音声認識，自動リップリーディングについて取り上げ，その高性能化について述べる．

自動リップリーディングの研究では，いかに認識に適した特徴量を唇画像列から抽出するかが重要なポイントとなる．これまでの研究は，特徴抽出の手法として大きくモデルベース法 [3]〜 [7] と画像ベース法 [7]〜 [12] に分類することができる．モデルベース法は，唇の輪郭モデルを少ないパラメータ数で構成し，原画像に対してモデルをトラッキングさせることにより認識を行う手法である．得られたモデルパラメータが撮影環境の変動に影響を受けにくいという利点があるが，いかに効率良く特徴を表すモデルを構築するか，様々な撮影条件のもとでモデルをトラッキングさせるか，というところに難点がある．一方，画像ベース法は，画像の画素値にサブサンプリングや主成分分析などの何らかの前処理を施したものを特徴ベクトルとして学習・認識に用いる手法である．モデルの構築が簡単であるという利点があるが，唇の位置や大きさ，照明条件などが認識率に直接的な影響を及ぼすため，これらの正規化が重要な要素技術となる．以上の点を

[†] 名古屋工業大学知能情報システム学科，名古屋市
Department of Computer Science, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan
[††] 東京工業大学大学院総合理工学研究科，横浜市
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan

考慮し，本論文では，隠れマルコフモデル（Hidden Markov Model: HMM）を用いた画像ベース法の正規化について考える．

従来の正規化手法は，人間が見て正しいと思われる正規化を目標として，HMM とは独立に何らかの基準を設け，HMM を学習する前に正規化を行うのが一般的であった．本論文では，ML（Maximum Likelihood）基準による正規化を考え，唇の位置，大きさ，傾き，平均輝度，コントラストの正規化プロセスがモデルの学習に統合された正規化学習法を提案する．提案法の基本的なアイデアは，正規化パラメータと HMM のモデルパラメータを同時に決定するという意味では，[10] と同様であるが，提案法は，唇位置，平均輝度だけでなく，唇の大きさ，傾き，コントラストなど一般化された正規化を考慮しているのみならず，EM（Expectation Maximization）アルゴリズムに基づいて定式化されているため，学習の繰返しにより学習データに関するゆう度が必ず増加することが保証されているという点が異なっている．

以下，本論文では，**2.** で正規化変換を定義し，ML 基準による正規化学習の枠組みについて述べる．**3.** では，再推定アルゴリズムについて説明し，**4.** で単語認識実験により提案法の有効性を示す．また，**5.** では，正規化変換の定義と出力確率の計算について考察する．最後に **6.** で，本論文のまとめと今後の課題について述べる．

2. 正規化変換

提案法では，唇の位置，大きさ，傾き，平均輝度，コントラストの正規化を線形変換により表現する．

2.1 幾何学変換

顔画像から唇部分を抽出し，同時にサブサンプリングを行う変換を幾何学変換と呼び，以下の線形変換で定義する．

$$\hat{\mathbf{o}}^{(r)}(t) = \mathbf{A}^{(r)}\mathbf{o}^{(r)}(t) \tag{1}$$

ここで，$\mathbf{o}^{(r)}(t)$ は，発声データ r の時刻 t における顔画像ベクトル，$\hat{\mathbf{o}}^{(r)}(t)$ はサブサンプリングされた唇画像ベクトルであり，各画像ベクトルは，画素値をラスタスキャンした 1 次元の長いベクトルで表現される．幾何学変換 $\mathbf{A}^{(r)}$ は，顔画像ベクトルから特徴ベクトルを抽出する長方形の行列であり，$\mathbf{A}^{(r)}$ の各行は，サブサンプリングにおける 1 ブロックの計算に対応する．特徴ベクトルの i 番目の要素を計算するときのサブサンプリングのブロック内のピクセル数を s_i とすると，変換行列 $\mathbf{A}^{(r)}$ の i 行は s_i 個の $\frac{1}{s_i}$ と 0 から構成され，$\frac{1}{s_i}$ の配置によって抽出される唇領域が決定される．この幾何学変換により，唇の位置，大きさ，傾きなどの正規化を表現することができる．

2.2 輝度値変換

照明条件などの輝度値に対する正規化として，以下の線形変換を定義する．

$$\tilde{\mathbf{o}}^{(r)}(t) = \mathbf{C}^{(r)}\hat{\mathbf{o}}^{(r)}(t) + \mathbf{b}^{(r)} \tag{2}$$

この輝度値変換は，カラーの正規化などの様々な正規化が表現できるが，本論文では入力画像をモノクロ画像とし，平均輝度とコントラストの正規化を考える．$\mathbf{b}^{(r)}$ と $\mathbf{C}^{(r)}$ に，以下の拘束条件を与えることにより，平均輝度とコントラストの正規化を表現することができる．

$$\mathbf{b}^{(r)} = b^{(r)}\mathbf{1} \tag{3}$$

$$\mathbf{C}^{(r)} = c^{(r)}\mathbf{I}, \quad (c^{(r)} > 0) \tag{4}$$

ただし，$\mathbf{1} = [1 \ \ldots \ 1]^T$ とする．ここで，$\mathbf{b}^{(r)}$，$\mathbf{C}^{(r)}$ は，それぞれ，平均輝度及びコントラストの正規化に対応している．

本論文では，唇の位置やカメラからの距離の違いによる唇の大きさ，照明条件などが単語発声区間では変化しないと仮定し，これらの正規化変換を各単語発声データ r ごとに用意した．しかし，提案法の枠組みでは，複数の変換行列を用いることにより，音素単位やフレーム単位で正規化を行うことも可能である．

2.3 ML 基準による正規化

従来の正規化手法は，HMM とは独立に何らかの基準を設け，HMM のパラメータを推定する前に正規化を行うのが一般的であった．本論文では，正規化のための線形変換 $\mathcal{T}^{(r)} = (\mathbf{A}^{(r)}, \mathbf{b}^{(r)}, \mathbf{C}^{(r)})$ を ML 基準により決定することを考える．与えられた HMM $\mathcal{M}$ に対し，最適な正規化変換 $\mathcal{T}'^{(r)}$ は次式で与えられる．

$$\mathcal{T}'^{(r)} = \underset{\mathcal{T}^{(r)}}{\operatorname{argmax}} P(\tilde{\mathbf{O}}^{(r)} \mid \mathcal{T}^{(r)}, \mathcal{M}) \tag{5}$$

ただし，$\tilde{\mathbf{O}}^{(r)} = [\tilde{\mathbf{o}}^{(r)}(1), \tilde{\mathbf{o}}^{(r)}(2), \ldots, \tilde{\mathbf{o}}^{(r)}(T_r)]$ は発声データ r の唇画像系列であり，各フレーム $\tilde{\mathbf{o}}^{(r)}(t)$ は，正規化変換 $\mathcal{T}^{(r)}$ により正規化されている．ここで，式 (5) の推定は，HMM のモデルパラメータ $\mathcal{M}$ を必要とする．しかし，学習時に得られる HMM は，正規化されていないデータから学習された HMM であり，この HMM から最適な正規化変換を得ることは

できない．また，正規化された HMM を学習するためには，逆に正規化変換 $\mathcal{T}^{(r)}, r = 1, 2, \ldots, R$ が必要となる．つまり，正規化変換 $\mathcal{T}^{(r)}, r = 1, 2, \ldots, R$ と HMM $\mathcal{M}$ は，それぞれを推定するために互いを必要とし，独立に推定することはできない．そこで，提案法では，正規化変換と HMM を同時に決定することを考える．つまり，

$$\lambda' = \underset{\lambda}{\operatorname{argmax}}\, P(\tilde{\mathbf{O}} \mid \lambda) \tag{6}$$

ここで，$\tilde{\mathbf{O}} = \{\tilde{\mathbf{O}}^{(1)}, \tilde{\mathbf{O}}^{(2)}, \ldots, \tilde{\mathbf{O}}^{(R)}\}$ は，HMM $\mathcal{M}$ の学習に必要なすべての学習データを表しており，各学習データは正規化変換 $\mathcal{T}^{(r)}$ により正規化されている．λ は，モデルの学習に必要なすべて正規化変換 $\mathcal{T}$ と HMM のモデルパラメータ $\mathcal{M}$ からなる．

$$\lambda = \{\mathcal{T}, \mathcal{M}\} \tag{7}$$

$$\mathcal{T} = \{\mathcal{T}^{(r)} \mid r = 1, 2, \ldots, R\} \tag{8}$$

提案法の基本的な考え方は，各発声に対する正規化変換の集合 $\mathcal{T}$ と HMM のモデルパラメータ $\mathcal{M}$ を ML 基準により同時に決定することである．

3. 再推定アルゴリズム

3.1 出力確率の計算

提案法では，画像の正規化を特徴ベクトルに対する線形変換で定義した．この変換は，特徴空間の次元圧縮を含む線形変換とみなすことができ，各正規化変換 $\mathcal{T}^{(r)}, r = 1, \ldots, R$ によって変換された特徴空間を同一の確率空間として共有することを意味している．しかし，正規化変換によって確率空間のスケールが変化するため，異なる正規化変換から計算されたゆう度を正しく比較することができない．そこで，提案法では，ガウス分布の出力を正規化することにより，出力確率の計算を行う（詳しくは **5.** 参照）．

$$\begin{aligned} &P(\tilde{\mathbf{o}}^{(r)}(t) \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \mathcal{T}^{(r)}) \\ &\quad = |\mathbf{A}^{(r)}\mathbf{A}^{(r)T}|^{\frac{1}{2}} |\mathbf{C}^{(r)}| \mathcal{N}(\tilde{\mathbf{o}}^{(r)}(t) \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \end{aligned} \tag{9}$$

ここで，$\boldsymbol{\mu}_m$ と $\boldsymbol{\Sigma}_m$ は，それぞれ，HMM におけるガウス分布 m の平均ベクトルと共分散行列を表している．

3.2 Q 関数

先に述べた最適化問題を解くため，本論文では，EM アルゴリズムを用いる．EM アルゴリズムは，ML 推定を近似するためのアルゴリズムであり，Q 関数と呼ばれる補助関数の最大化の繰返しからなる．Q 関数は，関数値の増加が学習データに対するゆう度の増加を保証している関数であり，関数の最大化を繰り返すことにより結果的にゆう度を局所的な極大点に導くことができる．

$$\begin{aligned} &Q(\lambda, \lambda') \geqq Q(\lambda, \lambda) \\ &\quad \Rightarrow P(\tilde{\mathbf{O}} \mid \lambda') \geqq P(\tilde{\mathbf{O}} \mid \lambda) \end{aligned} \tag{10}$$

ここでは，HMM のモデルパラメータと正規化変換を考慮し，Q 関数を次式で定義する．

$$\begin{aligned} &Q(\lambda, \lambda') \\ &\quad = \frac{1}{P(\tilde{\mathbf{O}} \mid \lambda)} \sum_{\text{all } \mathbf{q}} P(\tilde{\mathbf{O}}, \mathbf{q} \mid \lambda) \log P(\tilde{\mathbf{O}}, \mathbf{q} \mid \lambda') \\ &\quad = K - \frac{1}{2} \sum_{r,m,t}^{R,M,T_r} \gamma_m^{(r)}(t) \Big[K_m + \log |\boldsymbol{\Sigma}_m| \\ &\qquad + (\tilde{\mathbf{o}}^{(r)}(t) - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\tilde{\mathbf{o}}^{(r)}(t) - \boldsymbol{\mu}_m) \\ &\qquad - \log |\mathbf{A}^{(r)}\mathbf{A}^{(r)T}| - \log |\mathbf{C}^{(r)}|^2 \Big] \end{aligned} \tag{11}$$

ここで，K は状態遷移確率のみに依存する定数，K_m は，ガウス分布 m を正規化するための定数である．また，$\gamma_m^{(r)}(t)$ は，時刻 t におけるガウス分布 m の存在確率であり，Forward-Backward アルゴリズムにより計算することができる．

$$\gamma_m^{(r)}(t) = P(q_t = m \mid \mathbf{O}^{(r)}, \mathcal{T}^{(r)}, \mathcal{M}) \tag{12}$$

EM アルゴリズムでは，各パラメータ（正規化変換 $\mathcal{T}$ と HMM $\mathcal{M}$）について，個別に Q 関数を最大化にすることで学習データに関するゆう度を増加させていくことができる．

3.3 Q 関数の最大化

まず，正規化変換 $\mathcal{T}$ について Q 関数の最大化を考える．幾何学変換 $\mathbf{A}^{(r)}$ は，拘束条件（唇部分の抽出とサブサンプリング）をもった行列であるため，解析的に定めることは困難である．しかし，$\mathbf{A}^{(r)}$ による幾何学的な正規化を座標のアフィン変換と仮定すれば，唇の位置，大きさ，回転などのいくつかのパラメータを与えることにより $\mathbf{A}^{(r)}$ を一意に決定することができる．よって，これらのパラメータについて直接的な探索を行うことにより，Q 関数を最大にする $\mathbf{A}^{(r)}$ を求めることができる．実験では，計算量の観点からこう配法による探索を行った．

次に，輝度値変換 $(\mathbf{b}^{(r)}, \mathbf{C}^{(r)})$ の推定を考える．輝度値変換は，二つの係数 $b^{(r)}, c^{(r)}$ からなるため，これらの係数について個別に Q 関数を最大化する．まず，平均輝度の正規化係数 $b^{(r)}$ について，

$$\frac{\partial Q(\lambda, \lambda')}{\partial b'^{(r)}} = 0, \quad r = 1, 2, \ldots, R, \tag{13}$$

とおくことにより，

$$b'^{(r)} = \frac{\displaystyle\sum_{m,t}^{M,T_r} \gamma_m^{(r)}(t)(\boldsymbol{\mu}_m - \mathbf{C}'^{(r)}\hat{\mathbf{o}}^{(r)}(t))^T \boldsymbol{\Sigma}_m^{-1}\mathbf{1}}{\displaystyle\sum_{m,t}^{M,T_r} \gamma_m^{(r)}(t)\mathbf{1}^T\boldsymbol{\Sigma}_m^{-1}\mathbf{1}} \tag{14}$$

同様に，

$$\begin{aligned} &c'^{(r)2}\left\{\sum_{m,t}^{M,T_r} \gamma_m^{(r)}(t)\hat{\mathbf{o}}^{(r)T}(t)\boldsymbol{\Sigma}_m^{-1}\hat{\mathbf{o}}^{(r)}(t)\right\} \\ &-c'^{(r)}\left\{\sum_{m,t}^{M,T_r} \gamma_m^{(r)}(t)(\boldsymbol{\mu}_m - \mathbf{b}'^{(r)})^T\boldsymbol{\Sigma}_m^{-1}\hat{\mathbf{o}}^{(r)}(t)\right\} \\ &-nT_r = 0 \end{aligned} \tag{15}$$

の方程式が得られる．ここで，n は特徴ベクトル $\tilde{\mathbf{o}}^{(r)}(t)$ の次元数である．式 (14)，(15) を $b'^{(r)}, c'^{(r)}$ について解くことにより最適な輝度値変換 $(\mathbf{b}^{(r)}, \mathbf{C}^{(r)})'$ を推定することができる．

更に，正規化変換 $\mathcal{T}$ を固定した状態で，HMM のモデルパラメータ $\mathcal{M}$ について Q 関数を最大化する．各ガウス分布の平均ベクトルと共分散行列の推定式を以下に示す．

$$\boldsymbol{\mu}'_m = \frac{\displaystyle\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t)\tilde{\mathbf{o}}^{(r)}(t)}{\displaystyle\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t)} \tag{16}$$

$$\boldsymbol{\Sigma}'_m = \frac{\displaystyle\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t)(\tilde{\mathbf{o}}^{(r)}(t) - \boldsymbol{\mu}'_m)(\tilde{\mathbf{o}}^{(r)}(t) - \boldsymbol{\mu}'_m)^T}{\displaystyle\sum_{r,t}^{R,T_r} \gamma_m^{(r)}(t)} \tag{17}$$

これらの式は，正規化変換 $\mathcal{T}$ により正規化された学習データ $\tilde{\mathbf{o}}^{(r)}(t)$ を入力データと考えれば，一般的な連続分布 HMM の再推定と同様の形をしている．

各推定式を用いた正規化学習の手順を以下に示す．

Step 0. 初期の正規化変換 $\mathcal{T}$ を与え，初期モデル $\mathcal{M}$ を構築．

Step 1. $\gamma_m^{(r)}(t)$ の計算と正規化変換 $\mathcal{T}$ の更新．

step 1-1. 輝度値変換 $(\mathbf{b}^{(r)}, \mathbf{C}^{(r)})$ の更新．
step 1-2. 幾何学変換 $\mathbf{A}^{(r)}$ の更新．

Step 2. $\gamma_m^{(r)}(t)$ の計算と HMM のモデルパラメータ $\mathcal{M}$ の更新を 5 回．

Step 3. $P(\tilde{\mathbf{O}} \mid \lambda)$ の変化が小さければ終了．それ以外は Step 1 へ．

この手順には，様々な繰返しのパターンが考えられるが，本論文では予備実験の結果から上記のとおり決定した．また，この手順の各パラメータ更新（Step 1, 2）により，Q 関数値が増加することは明らかである．

$$Q(\lambda, \lambda') \geqq Q(\lambda, \lambda) \tag{18}$$

よって，正規化学習の繰返しにより，学習データに関するゆう度が必ず増加することが保証される．

$$\begin{aligned} P(\tilde{\mathbf{O}} \mid \mathcal{T}, \mathcal{M}) &\leqq P(\tilde{\mathbf{O}} \mid \mathcal{T}', \mathcal{M}) \\ &\leqq P(\tilde{\mathbf{O}} \mid \mathcal{T}', \mathcal{M}') \end{aligned} \tag{19}$$

ここで，$\mathcal{T}'$ と $\mathcal{M}'$ は，それぞれ更新された正規化変換と HMM のモデルパラメータを表している．

3.4 テストデータの正規化

正規化学習で得られたモデルは，唇位置や照明条件が正規化されているため，テストデータについても同様に正規化を行う必要がある．本論文では，各モデル $\mathcal{M}_i, i = 1, 2, \ldots, I$ に対しゆう度が最大となる正規化変換を求めた上でテストデータ $\tilde{\mathbf{O}}$ のゆう度を計算する．

$$\mathcal{T}'_i = \underset{\mathcal{T}_i}{\operatorname{argmax}}\, P(\tilde{\mathbf{O}} \mid \mathcal{T}_i, \mathcal{M}_i) \tag{20}$$

この正規化変換は，学習時と同様の更新式を用いることにより推定することができる（Step 1）．各モデルについて推定された正規化変換を用いてゆう度を計算し，最もゆう度の高いモデルを認識結果とする．

$$\text{result} = \underset{i}{\operatorname{argmax}}\, P(\tilde{\mathbf{O}} \mid \mathcal{T}'_i, \mathcal{M}_i) \tag{21}$$

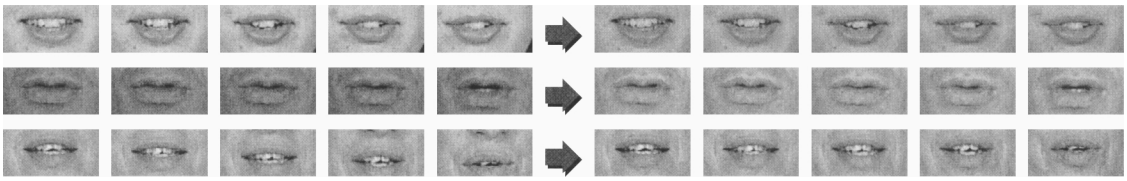


図 1　リップトラッキングと従来法による輝度値正規化
Fig. 1　Lip-tracking and conventional intensity normalization.

4.　単語認識実験

4.1　実 験 条 件

正規化学習の効果を確認するため，M2VTS データベース [13] を用いた単語認識実験を行った．このデータベースは，37 人の音声と顔動画像からなるバイモーダルデータベースである．各被験者につき，5 時期分の撮影が収録されており，フランス語の数字単語 “0” から “9” までを連続して発声している．撮影は，被験者の声や外見にばらつきをもたせるため，1 週間おきに撮影されている．また，5 時期目は他の時期に比べて認識条件を厳しくするため，撮影条件を変えてある．画像のフレームレートは 25 frame/s で，各画像は 24 bit カラーの 350×286 の画像である．

本実験では，画像をモノクロ画像に変換し，撮影データを単語区間に分割して実験に用いた．また，37 人の被験者のうち，唇の周りがほとんど髭で覆われている被験者と，正確にフランス語を発声していないと思われる被験者の 2 人を除いた 35 人（男性 23 名，女性 12 名）の 1〜4 時期の撮影を用いて認識実験を行った．実験方法は，leave-one-out 法を用いた話者独立の単語認識実験とした．この手法では，35 人の被験者のうち 1 人を認識のために残しておき，残りの 34 人で学習を行った．これをすべての被験者について行い，その平均の認識率を認識結果とした．HMM は left-to-right 型の単語 HMM とし，状態数は各単語のフレーム数の平均を考慮し，5〜8 とした．また，ガウス分布の分散は，対角共分散とした．特徴量は，画素値（サブサンプリング）と，その Δ，ΔΔ パラメータを連結したベクトルを用いた．ベクトルサイズは，各特徴量につき 16×8，全体で 384 とした．

4.2　従来法による正規化

提案法と比較するため，従来法による認識実験を行った．唇の位置については，単語データの第 1 フレームの唇位置を人手により決定した．認識に必要な唇領域は，80×40 ピクセルとした．本論文では，単語発声区間中は唇の位置が動かないと仮定して正規化学習の定式化を行ったが，実際には大きく動く話者もいるため，相関関数と sobel フィルタを用いたトラッキング [11] を用いた(注 1)．この手法では，連続したフレームに sobel フィルタをかけた上で，直前のフレームに対し，唇領域の相関値が最大となる位置を探索することにより，唇位置の補正を行う．また，従来法による輝度値変換として，サブサンプリングを行った単語データの画素値の平均や分散をデータ間で一定とする線形変換 [12] を行った．従来法による正規化を行った画像の変化を図 1 に示す．輝度値の正規化により，画像の明るさがデータ間でほぼ同じになっていることがわかる．また，3 列目のデータの原画像列では唇の位置がずれているが，トラッキングの効果により唇の位置が中心に保たれていることがわかる．これらの正規化手法による認識率を表 1 に示す．リップトラッキング，輝度値変換ともに高い認識率の改善を示しており，画像ベース法における正規化がいかに重要であるかがよくわかる．

表 1　従来法による認識率
Table 1　Recognition rate of conventional approach.

	without		tracking	
	mix1	mix2	mix1	mix2
without	66.4	70.8	72.8	73.7
mean	69.0	77.9	75.7	79.2
mean+variance	69.9	77.7	77.1	80.0

4.3　正規化学習による認識実験

提案法では，EM アルゴリズムによる局所的な最適解を求めるため，初期値を真の最適解に近い値に設定することが望ましい．そこで，本実験では，従来の正規化手法によって得られた正規化変換と HMM を初期値とした．また，予備実験の結果から，リップトラッキングと平均輝度の正規化により得られたモデルを初期モデルとした．混合数 2 のモデルを用いて学習を行っ

(注 1)：正規化変換をフレーム単位で割り当てることにより，提案法の枠組みの中で，トラッキングを行うことも可能である．

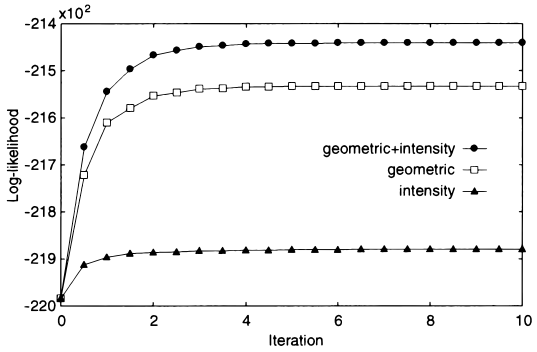


図 2　学習データに関するゆう度の変化
Fig. 2　Log likelihoods of HMMs for training data.

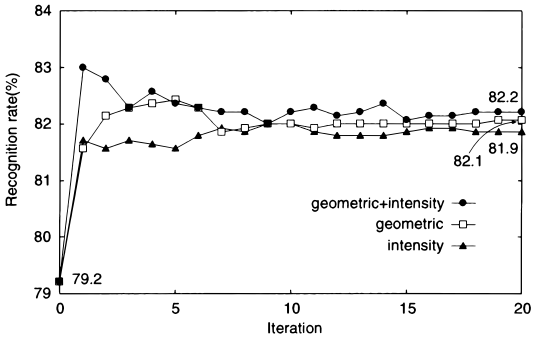


図 3　正規化学習による認識率
Fig. 3　Recognition rate of normalized training.

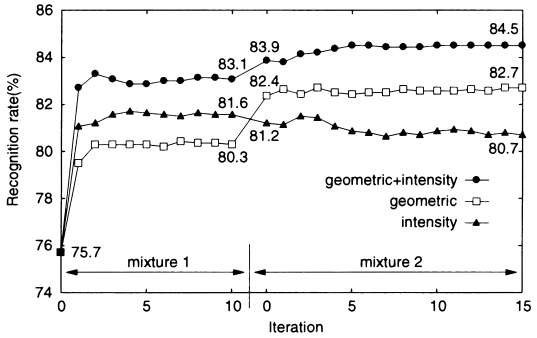


図 4　ミクスチャ分割による正規化学習の認識率
Fig. 4　Recognition rate of normalized training with mixture division.

たときの正規化学習の繰返しにおける学習データのゆう度変化を図 2 に示す．“geometric” は幾何学変換のみを用いた正規化学習を表しており，輝度値変換については初期値のまま固定している．同様に “intensity” は，輝度値変換のみを更新している．幾何学変換については，唇の位置，大きさ，傾きについて実装したが，データベースの画像に傾きがあまりなく，予備実験でも計算量の増加に対し認識率の改善があまり見られなかったため，本実験では，位置と大きさについてのみの正規化を行う幾何学変換を用いた．横軸は，**3.** で述べた正規化学習の手順に対応しており，正規化変換の更新（step 1）とモデルパラメータの更新 (step 2) で，繰返し回数を 1 とする．この図から，正規化学習の各ステップにおいて，学習データに関するゆう度が単調に増加していることがわかる．また，幾何学変換，輝度値変換を個別に見てもゆう度が増加してることが確認できる．

混合数 2 の初期モデルを用いて正規化学習を行った場合の認識結果を図 3 に示す．図より，正規化学習により，従来法と比較し，大幅に認識率が改善されることがわかる．しかし，輝度値変換と幾何学変換の個別の改善に比べ，両方とも正規化した場合の認識率の改善がほとんど見られなかった．また，後述する混合数 1 の正規化学習による認識率（“geometric+intensity” において 83.1%）より悪い結果となった．この原因として，マルチミクスチャHMM における初期値の影響が考えられる．この実験で用いた初期モデルは，唇位置と輝度値について従来法による正規化を行っているが，大きさについては正規化を行っていないため，学習データ間でかなりばらつきがある．このような状態で初期モデルを学習した場合，唇の大きさに依存したミクスチャが推定され，話者性や唇形状の違いをうまく表現できないと考えられる．また，大きさの正規化だけでなく，他の正規化パラメータについても同様のことがいえる．このような問題を回避する方法として，本実験では，あらかじめ混合数 1 で正規化学習を行い，その後，ミクスチャを分割する手法を用いた．実験の手順は以下のとおりである．

1. 混合数 1 で正規化学習による再推定を 10 回.
2. ガウス分布を二つに分割.
3. HMM のみの更新（step 2）を 20 回.
4. 混合数 2 で正規化学習による再推定を 15 回.

その認識結果を図 4 に示す．まず，混合数 1 において，幾何学変換，輝度値変換ともに認識率の改善が見られる．また，“geometric+intensiy” では 83.1%を示しており，従来法の 77.1%（表 1）に対し高い改善率が得られた．更に，混合数 2 で正規化学習を行った場合，輝度値変換のみでは，正規化学習の効果は見られなかっ

たが，幾何学変換を同時に行うことにより，84.5%の認識率が得られた．また，従来法の 80.0% （表 1）に対し 22.5%の誤り改善率を達成している．

各正規化による個別の効果を表 2 に示す．初期モデル “initial” は，人手による位置の正規化と従来法による平均輝度の正規化を行ったモデルであり，“mean”，“variance” は，それぞれ，提案法による平均輝度，コントラストの正規化を表す． 提案法による平均輝度の正規化のみを用いた場合，従来法に対し改善が得られなかった．また，従来法による平均輝度の正規化に提案法のコントラスト正規化を適用した場合についてもあまり改善が得られなかったが，提案法による平均輝度，コントラストの正規化を同時に行うことにより大幅な改善が見られた．幾何学変換についても，提案法による正規化を位置と大きさを個別に行った場合に比べ，同時に正規化を行うことにより効果が見られた．以上の結果から，従来法が各正規化に個別の基準を設けて正規化しているのに対し，提案法は ML 基準による統一された基準による正規化を行っているため，各正規化を同時に適用したときに大きな効果が得られたと考えられる．

表 2　各正規化による個別の効果
Table 2　Effects of individual normalization.

	mix1	mix2	mix2(1→2)
initial	75.7	79.2	–
mean	75.6	79.4	79.9
variance	76.0	79.6	80.9
mean+variance	81.6	81.9	80.7
location	76.3	80.3	80.7
scaling	78.1	79.3	80.3
location+scaling	80.3	82.1	82.7

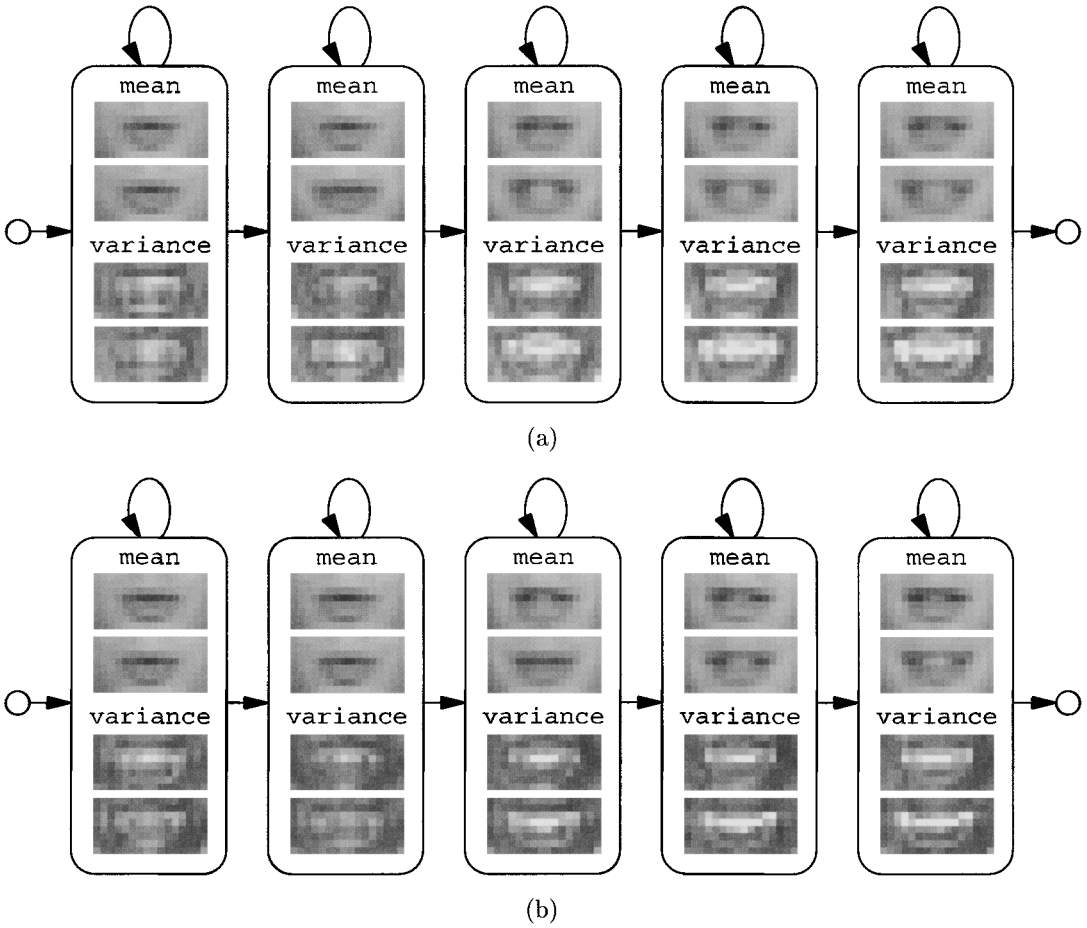


図 5　HMM /trois/ の比較．従来法モデル (a) と正規化学習モデル (b)
Fig. 5　HMM with conventional method (a) and with normalized training (b).

正規化学習によるモデルの変化を比較するため，静的特徴量の平均と分散を視覚的に表したものを図 5 に示す．分散については，明るい部分ほど値が大きいことを表す．ここで，(a) はリップトラッキングと平均輝度の正規化を行ったモデル，(b) は，ミクスチャ分割により最も高い認識率が得られたモデル（iteration 15）である．平均ベクトルを見ると，正規化学習により，唇の輪郭がはっきりしていることがわかる．また，従来法では，ミクスチャ間で唇の大きさが異なるが，正規化学習モデルでは，ほぼ同じ大きさになっている．分散についても，正規化学習モデルは，分散の大きい領域が少なくなっており，唇位置や照明条件に依存しない HMM が構築されたと考えられる．

5. 考　　察

MLLR（Maximum Likelihood Linear Regression）[14]〜[16] は，HMM を用いた音声認識において話者や発話環境の違いによる認識率の低下を防ぐことに成功している手法である．また，MLLR を認識だけでなく，学習にも用いた SAT (Speaker Adaptive Training) [16]〜[19] もその有効性が確認されている．これらの手法では，線形変換により話者性や発話環境を表現しており，線形変換は正規化を表現する有力な手法であると考えられる．提案法は，これらの音声認識に用いられる手法のアイデアを画像ベース法に基づいた自動リップリーディングの正規化に応用したものとみなすことができる．

線形変換による正規化には，特徴ベクトルを変換する特徴空間変換（feature space transformation）と HMM のガウス分布を変換するモデル空間変換（model space transformation）の 2 通りが考えられる．提案法では，画像の正規化を特徴空間変換で定義し，ガウス分布の出力を正規化することにより出力確率の計算を行った．以下では，提案法の正規化変換が平均ベクトルと共分散行列に同じ線形変換を用いる拘束付きモデル空間変換（constraint model space transformation）と等価であることを示す．

5.1 モデル空間変換による輝度値変換

提案法の輝度値変換は，MLLR と同様に，拘束付きモデル空間変換で表現することができる．モデル空間変換による輝度値変換を以下に定義する．ただし，r, m, t のインデックスは省略する．

$$\check{\boldsymbol{\mu}} = \check{\mathbf{C}}\boldsymbol{\mu} + \check{\mathbf{b}}, \quad \check{\boldsymbol{\Sigma}} = \check{\mathbf{C}}\boldsymbol{\Sigma}\check{\mathbf{C}}^T \tag{22}$$

ここで，$(\check{\mathbf{b}}, \check{\mathbf{C}})$ はモデル空間変換における輝度値変換を表しており，ガウス分布の平均ベクトル $\boldsymbol{\mu}$ と共分散行列 $\boldsymbol{\Sigma}$ を線形変換している．この変換を用いたガウス分布は以下のように変形できる．

$$\begin{aligned}&\mathcal{N}(\mathbf{o} \mid \check{\mathbf{C}}\boldsymbol{\mu} + \check{\mathbf{b}}, \check{\mathbf{C}}\boldsymbol{\Sigma}\check{\mathbf{C}}^T) \\ &\quad = |\mathbf{C}|\mathcal{N}(\mathbf{Co} + \mathbf{b} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\end{aligned} \tag{23}$$

ここで，$(\mathbf{b}, \mathbf{C})$ は特徴空間変換における輝度値変換を表している．ただし，

$$\mathbf{b} = -\check{\mathbf{C}}^{-1}\check{\mathbf{b}}, \quad \mathbf{C} = \check{\mathbf{C}}^{-1} \tag{24}$$

となる．ここで，確率の条件を考えると，モデル空間変換は，平均，分散をそれぞれ $\check{\boldsymbol{\mu}}, \check{\boldsymbol{\Sigma}}$ とするガウス分布であるため，

$$\begin{aligned}&\int \mathcal{N}(\mathbf{o} \mid \check{\mathbf{C}}\boldsymbol{\mu} + \check{\mathbf{b}}, \check{\mathbf{C}}\boldsymbol{\Sigma}\check{\mathbf{C}}^T)d\mathbf{o} \\ &\quad = |\mathbf{C}| \int \mathcal{N}(\mathbf{Co} + \mathbf{b} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{o} = 1\end{aligned} \tag{25}$$

となる．つまり，モデル空間変換の場合，ガウス分布の出力を確率とみなすことができるが，特徴空間変換では，ガウス分布を正規化する必要がある．よって，

$$\begin{aligned}&P(\mathbf{o} \mid \check{\mathbf{C}}\boldsymbol{\mu} + \check{\mathbf{b}}, \check{\mathbf{C}}\boldsymbol{\Sigma}\check{\mathbf{C}}^T) \\ &\quad = \mathcal{N}(\mathbf{o} \mid \check{\mathbf{C}}\boldsymbol{\mu} + \check{\mathbf{b}}, \check{\mathbf{C}}\boldsymbol{\Sigma}\check{\mathbf{C}}^T) \\ &\quad = |\mathbf{C}|\mathcal{N}(\mathbf{Co} + \mathbf{b} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\quad = P(\mathbf{Co} + \mathbf{b} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{b}, \mathbf{C})\end{aligned} \tag{26}$$

となり，提案法の輝度値変換は，モデル空間変換と等価である．

5.2 幾何学変換による確率計算

提案法の幾何学変換 $\mathbf{A}^{(r)}$ は非正方行列で表現されるため，モデル空間変換を用いることはできない．しかし，$\mathbf{A}^{(r)}$ についても輝度値変換と同様に，ガウス分布の正規化が必要である．予備実験では，ガウス分布の正規化なし（Q 関数の $\log|\mathbf{A}^{(r)}\mathbf{A}^{(r)T}|$ を無視する）で唇の大きさの正規化を行った場合，実際より広い領域が唇領域として推定され，認識率が低下する現象が確認されている．

ここでは，幾何学変換を次式のように正方行列に拡張することにより，幾何学変換によるガウス分布の正規化について考える．

$$\mathbf{A}_n = \begin{bmatrix} \mathbf{A}_p \\ \mathbf{A}_{n-p} \end{bmatrix} \tag{27}$$

ここで，$\mathbf{A}_n$ は，拡張された幾何学変換であり，正方行列で表される．$\mathbf{A}_p$ は，唇部分の特徴ベクトルを抽出する変換であり，提案法の幾何学変換に対応する．$\mathbf{A}_{n-p}$ は，唇領域以外の特徴量を抽出する変換とし，$\mathbf{A}_p$ と $\mathbf{A}_{n-p}$ は直交している（$\mathbf{A}_p$ の各行ベクトルに対し，$\mathbf{A}_{n-p}$ のすべての行ベクトルが直交している）と仮定する．また，$\mathbf{A}_n$ により変換された特徴ベクトルから学習されたガウス分布の平均ベクトルと共分散行列を以下に定義する．

$$\boldsymbol{\mu}_n = \begin{bmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_{n-p} \end{bmatrix}, \boldsymbol{\Sigma}_n = \begin{bmatrix} \boldsymbol{\Sigma}_p & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{n-p} \end{bmatrix} \quad (28)$$

ここで，$\mathbf{A}_p$ と $\mathbf{A}_{n-p}$ によって分割された部分空間の相関は考慮しないと仮定する．変換 $\mathbf{A}_p$ により，発声内容に関する情報が過不足なく抽出されているとするならば，$\mathbf{A}_p$ と $\mathbf{A}_{n-p}$ とは直交することになり，この仮定は妥当なものであることがわかる．拡張された幾何学変換 $\mathbf{A}_n$ は正方行列であるため，輝度値変換と同様，ガウス分布を正規化することにより，モデル空間変換と等価な表現が可能である．

$$\begin{aligned} &P(\mathbf{A}_n\mathbf{o} \mid \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \mathbf{A}_n) \\ &\quad = |\mathbf{A}_n| \mathcal{N}(\mathbf{A}_n\mathbf{o} \mid \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \end{aligned} \quad (29)$$

ただし，$\mathbf{o}$ は顔画像ベクトルを表す．$\mathbf{A}_p$ と $\mathbf{A}_{n-p}$ が直交することから，

$$\begin{aligned} |\mathbf{A}_n| &= |\mathbf{A}_n\mathbf{A}_n^T|^{\frac{1}{2}} \\ &= \left| \begin{bmatrix} \mathbf{A}_p\mathbf{A}_p^T & \mathbf{A}_p\mathbf{A}_{n-p}^T \\ \mathbf{A}_{n-p}\mathbf{A}_p^T & \mathbf{A}_{n-p}\mathbf{A}_{n-p}^T \end{bmatrix} \right|^{\frac{1}{2}} \\ &= |\mathbf{A}_p\mathbf{A}_p^T|^{\frac{1}{2}} |\mathbf{A}_{n-p}\mathbf{A}_{n-p}^T|^{\frac{1}{2}} \end{aligned} \quad (30)$$

となる．また，$\mathbf{A}_p$ と $\mathbf{A}_{n-p}$ によって分割された空間の相関は無視しているため，$\mathbf{A}_n$ による確率計算は，各部分空間で独立に計算できる．よって，$\mathbf{A}_n$ による出力確率は，次式で表される．

$$\begin{aligned} &P(\mathbf{A}_n\mathbf{o} \mid \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \mathbf{A}_n) \\ &\quad = P(\mathbf{A}_p\mathbf{o} \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p, \mathbf{A}_p) \\ &\qquad \times P(\mathbf{A}_{n-p}\mathbf{o} \mid \boldsymbol{\mu}_{n-p}, \boldsymbol{\Sigma}_{n-p}, \mathbf{A}_{n-p}) \\ &\quad = |\mathbf{A}_p\mathbf{A}_p^T|^{\frac{1}{2}} \mathcal{N}(\mathbf{A}_p\mathbf{o} \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \\ &\qquad \times |\mathbf{A}_{n-p}\mathbf{A}_{n-p}^T|^{\frac{1}{2}} \mathcal{N}(\mathbf{A}_{n-p}\mathbf{o} \mid \boldsymbol{\mu}_{n-p}, \boldsymbol{\Sigma}_{n-p}) \end{aligned} \quad (31)$$

ここで，$|\mathbf{A}_p\mathbf{A}_p^T|^{\frac{1}{2}}$ は，$\mathbf{A}_p$ によって得られた部分空間の確率を正規化していると考えられる．また，提案法では，唇領域の特徴量のみを考慮しているため，提案法の幾何学変換 $\mathbf{A}^{(r)}$ は，$\mathbf{A}_p$ に対応する．

$$\mathbf{A}^{(r)} = \mathbf{A}_p \quad (32)$$

$$\boldsymbol{\mu}_m = \boldsymbol{\mu}_p, \ \boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}_p \quad (33)$$

よって，提案法の幾何学変換による出力確率は次式で表される．

$$\begin{aligned} &P(\mathbf{A}_p\mathbf{o} \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p, \mathbf{A}_p) \\ &\quad = P(\hat{\mathbf{o}}^{(r)}(t) \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \mathbf{A}^{(r)}) \\ &\quad = |\mathbf{A}^{(r)}\mathbf{A}^{(r)T}|^{\frac{1}{2}} \mathcal{N}(\hat{\mathbf{o}}^{(r)}(t) \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \end{aligned} \quad (34)$$

更に，式 (34) に輝度値変換（式 (26)）を適用することにより，式 (9) が得られる．

6. むすび

本論文では，画像ベース法を用いた自動リップリーディングにおいて，唇位置，大きさ，傾き，平均輝度，コントラストの正規化プロセスが HMM の学習と統合された正規化学習法を提案した．M2VTS データベースによる単語認識実験により，最高で 84.5%の認識率，従来の正規化手法に対し 22.5%の誤り改善率を達成した．提案法は，EM アルゴリズムに基づき定式化されているため，正規化学習を繰り返すことにより学習データに関するゆう度が必ず増加することが保証されている．

今後の課題としては，より大規模なデータベースによる実験，連続音声認識への対応及び，聴覚情報を用いた音声認識部との統合などが挙げられる．また，提案法によるリップトラッキングの実現，カラーの正規化や他の特徴量の検討なども今後の課題である．

文　献

[1] J. MacDonald, S. Andersen, and T. Bachmann, “Hearing by eye: Visual spatial degradation and the McGurk effect,” Proc. EuroSpeech, pp.1283–1286, 1999.

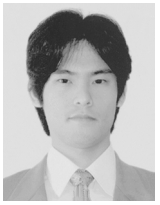
[2] M. Radeau and C. Colin, “The role of spatial separation on ventriloquism and McGurk illusions,” Proc. EuroSpeech, pp.1295–1298, 1999.

[3] J. Luettin, N. Thacker, and S. Beet, “Visual speech recognition using active shape models and hidden Markov models,” Proc. ICASSP, pp.817–820, 1996.

[4] J. Luettin, N. Thacker, and S. Beet, “Speechreading using shape and intensity information,” Proc. ICSLP,

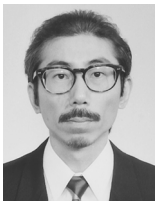
pp.58–61, 1996.
[5] J. Luettin, “Towards speaker independent continuous speechreading,” Proc. Eurospeech, pp.1991–1994, 1997.
[6] S. Dupont and J. Luettin, “Using the multi-stream approach for continuous audio-visual speech recognition: Experiments on the M2VTS database,” Proc. ICSLP, pp.1283–1286, 1998.
[7] G. Potamianos and A. Potamianos, “Speaker adaptation for audio-visual speech recognition” Proc. Eurospeech, pp.1291–1294, 1999.
[8] M.S. Gray, J.R. Movellan, and T.J. Sejnowski, “Dynamic features for visual speech-reading: A systematic comparison,” Advances in Neural Information Processing Systems, vol.9, pp.751–757, 1997.
[9] J.R. Movellan, “Visual speech recognition with stochastic networks,” in Advances in Neural Information Processing Systems 7, ed., G. Tesauro, D. Touretzky, and T. Leen, MIT Press Cambridge, 1995.
[10] O. Vanegas, K. Tokuda, and T. Kitamura, “Lip location normalized training for visual speech recognition,” IEICE Trans. Inf. & Syst., vol.E83-D, no.11, pp.1969–1977, Nov. 2000.
[11] O. Vanegas, K. Tokuda, and T. Kitamura, “Location normalization of HMM-based lip reading: Experiments for the M2VTS Database,” Proc. ICIP, pp.343–347, 1999.
[12] O. Vanegas, A. Tanaka, K. Tokuda, and T. Kitamura, “HMM-based visual speech recognition using intensity and location normalization,” Proc. ICSLP, pp.289–292, 1998.
[13] http://www.tele.ucl.ac.be/PROJECTS/M2VTS/
[14] M.J.F. Gales and P.C. Woodland, “Mean and variance adaptation within the MLLR framework,” Computer Speech and Language, pp.249–264, Oct. 1996.
[15] M.J.F. Gales and P.C. Woodland, “Maximum likelihood linear transformations for HMM-based speech recognition,” Computer Speech and Language, pp.75–98, Dec. 1998.
[16] M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” http://www-svr.eng.cam.ac.uk/reports/abstracts/speech/gales_tr291.html
[17] D. Pye and PC. Woodland, “Experiments in speaker normalisation and adaptation for large vocabulary speech recognition,” Proc. ICASSP, pp.1047–1050, 1997.
[18] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “Speaker adaptive training: A maximum likelihood approach to speaker normalization,” Proc. ICASSP, pp.1043–1046, 1997.
[19] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “A compact model for speaker-adaptive training,” Proc. ICSLP, pp.1137–1140, 1996.

（平成 14 年 4 月 4 日受付，7 月 9 日再受付）



南角　吉彦　（学生員）

平 11 名工大・工・知能情報システム卒．平 13 同大大学院博士前期課程了．現在，同大学院博士後期課程在学中．バイモーダル音声認識の研究に従事．日本音響学会会員．



徳田　恵一　（正員）

昭 59 名工大・工・電子卒．平元東工大大学院博士課程了．同年東工大電気電子工学科助手．平 8 名工大知能情報システム学科助教授．工博．音声分析，音声合成・符号化，音声認識，ディジタル信号処理，マルチモーダルインタフェースの研究に従事．平 13 電気通信普及財団賞，平 13 本会論文賞，猪瀬賞各受賞，日本音響学会，人工知能学会，情報処理学会，IEEE，ISCA 各会員．

北村　　正　（正員）

昭 48 名工大・工・電子卒．昭 53 東工大大学院博士課程了．同年東工大精密工学研究所助手．昭 58 名工大・工・電子工学科講師．昭 59 同助教授．平 7 名工大知能情報システム学科教授．工博．音声情報処理，マルチメディア情報処理の研究に従事．日本音響学会，情報処理学会，IEEE，ISCA 各会員．

小林　隆夫　（正員）

昭 52 東工大・工・電気卒．昭 57 同大大学院博士課程了．同年同大精密工学研究所助手．同助教授を経て平 10 東工大大学院総合理工学研究科物理情報システム創造専攻教授．工博．ディジタルフィルタ，音声の分析・合成・符号化・認識，マルチモーダルインタフェースの研究に従事．平 13 電気通信普及財団賞，平 13 本会論文賞，猪瀬賞各受賞，日本音響学会，情報処理学会，IEEE，ISCA 各会員．