Internal Use Only （非公開）

002

TR-S-0020

# バイモーダル音声認識のための音声と顔画像の統合モデルにとその適応化アルゴリズムに関する研究

# Study on adaptive integration of audio and visual information for bi-modal speech recognition

熊谷 建一 Kenichi Kumatani
村井 和昌 Kazumasa Murai
中村 哲 Satoshi Nakamura

2001. 3.30

本研究では，音声と口唇画像（バイモーダル）を用いた音声認識において二つの問題について述べる．(1) まず，小規模の音声画像同期データベースから，HMM 合成を用い，音声と口唇画像の特徴の同期／非同期性を表現する方法について述べる．(2) 次に，環境に応じて，その HMM を適応化する方法について述べる．本手法による統合方法で，単語認識実験を行った．その結果，従来の音声・画像の統合方法より良い認識性能が得られ，また，少数の単語データからストリーム重み推定を行うことで，音声のみしか用いない音声認識システムより良い性能が得られることが分かり，本手法が有効であることが確かめられた．

©2001 ＡＴＲ音声言語通信研究所

©2001 by ATR Spoken Language Translation Research Laboratories

# 1 abstract

本論文では，音声と口唇画像（バイモーダル）を用いた音声認識において二つの問題について述べる．(1) まず，小規模の音声画像同期データベースから，HMM 合成を用い，音声と口唇画像の特徴の同期／非同期性を表現する方法について述べる．(2) 次に，環境に応じて，その HMM を適応化する方法について述べる．

まず，音声のみのデータと口唇画像のみのデータを用い各々独立に学習した音声と画像 HMM を合成し，それから，音声画像同期データを用いて合成 HMM を再学習することで，音声と画像の同期／非同期性の学習を試みる．この方法により，音声のみ，あるいは画像のみのデータを利用することができ，比較的小規模の音声画像同期データで音声と口唇画像の同期／非同期性を良く学習できる．

さらに，その音声・画像 HMM をマルチストリーム HMM として形式化し，その音素 HMM を GMM として簡単化しておくことで，最小分類誤り基準 (MCE) により，少数の単語データから音声と画像のストリーム重みを推定する方法を検討する．本手法による統合方法で，単語認識実験を行った．その結果，従来の音声・画像の統合方法より良い認識性能が得られ，また，少数の単語データからストリーム重み推定を行うことで，音声のみしか用いない音声認識システムより良い性能が得られることが分かり，本手法が有効であることが確かめられた．

# 2 まえがき

近年，音声認識の性能は大きく改善されたが，未だ音声の SNR が低い雑音環境での高い認識性能には問題が残されている．このような環境に頑健な音声認識システムとして，音声だけでなく唇周辺の動画像を用いたバイモーダル音声認識システムが研究されている [1]-[10]．また，今日のマルチメディア機器の普及によりマルチメディアサービスが注目されており，音声と画像情報を用いたバイモーダル音声認識は，最も注目される技術の一つであると言える．

バイモーダル音声認識システムでは，音声の SNR が高い状況では，唇周辺は音声の調音器官の一部分でしかないため画像の情報による認識は音声に及ばないが，画像は音響的雑音による劣化が起こらないため，音声の SNR の低い状況では，音声より高い認識性能を示すということなどがバイモーダル音声認識システムとしてあげられる．従って，二つのモダリティが相互に音声認識に助けあうことが期待できる．

バイモーダル音声認識には，通常の音声認識システムと同様に HMM が用いられる．HMM は，統計モデルであり，学習データ数に応じて高い学習能力を持つことが知られ，現在の音声認識システムに広く用いられている [12]．バイモーダル音声認識に HMM を用いることには，音声と画像を確率・統計的に統合でき，既存の音声認識システムに組み込みやすいという利点もある．

はじめに，そのようなバイモーダル音声認識システムを構築する際に，発声する前に発話の準備のために口が開き，発声の後に口が閉じるといったような音声と画像のイベントが非同期に起きるといった問題がある．また，お互いのモーダリティは，完全に非同期ではなく同期したイベントも持つと考えられる．

HMM を用いて，音声と画像を統合する方法では，初期統合と結果統合が知られている [1]-[10]．しかし，両手法とも，学習データから効率よく同期・非同期性を学習することを十分に検討されているとはいえない．

まず本論文では，そのような問題を解決する方法として，音声 HMM と画像 HMM を合成し，非同期な HMM を作成し [4]，さらに，音声と画像の同期関係を学習するために，合成した音声・画像 HMM を再学習する，HMM 合成に基づいた統合方法 [5] ( 以下合成統合 ) を提案する．この方法では，学習用の音声・画像同期データが少なく，合成した音声・画像 HMM を再学習出来ない場合も，初期の合成モデルをそのまま代用することが可能である．従って，初期統合と結果統合と比べると，音声・画像同期データベースから同期／非同期性をよりよく表現できると考えられ，従来の統合方法よりも優れた認識性能をもつことが期待できる．そして，この方法による認識性能の評価を行なう．

次に，音声がクリーンな場合は，画像情報より音声情報の方が重要であり，音声が劣化しているなら，画像情報の方が重要であるというように，環境に応じてバイモーダル音声認識システムを適応化する問題がある．本論文では，音声と画像の HMM を マルチストリーム HMM [13] として合成し，各々の出力確率にかかるべき乗の重み（以下ストリーム重み）を操作することでその問題を解決する．通常，認識性能を最も良くするストリーム重みの値は，音声の SNR や画像の劣化などの環境要因によって変わる．しかし，様々な環境において，各々に最適なストリーム重み値を求めておくことは多大な労力を必要とする．従って，ストリーム重みを少数のデータから自動的に推定する必要がある．自動的に，ストリーム重みを推定するためには基準が重要となる．ストリーム重みを推定する基準としては，よく知られている方法として，尤度最大化 (ML) 基準と最小分類誤り (MCE) 基準がある [6] [8] [11]．ストリーム重みは情報の信頼度を表す変数であり確率変数でないため，一般に HMM の遷移確率と出力確率の学習の基準である ML 基準によるストリーム重み推定は不適切である [6] [10]．ストリーム

重みに，ヒューリスティックな制限をつけ ML 基準で推定する手法 [11] があるが，音声と画像では，尤度のダイナミックレンジが大きく違うため良い認識性能が得られない．それに対し，正しいクラスと誤ったクラスの距離を最大化する基準である MCE 基準による学習が認識率を最大化させるストリーム重みに一致することが報告されている [6] [8]．MCE 基準を達成するアルゴリズムとしては，直接探索による方法 [8]，GPD アルゴリズムによる方法 [6] [10] がある．直接探索による方法では，マシンパワーを必要としないという利点があるが，多変数のストリーム重み推定には適用できないという欠点がある．直接探索に対して，GPD アルゴリズムは，多変数にも適用可能で，応用性が高いアルゴリズムであるが，環境に応じ，少数のデータから，ストリーム重みを推定することは検討されていない．

そこで，本論文では，音素 HMM を GMM として簡単化しておき，その GMM の推定された音声と画像のストリーム重みを用いることで，バイモーダル音声認識システムを適応化することを検討する．そして，本手法によるストリーム重み推定精度と認識性能を評価を行なう．

以下の章は次のように構成されている．2 章では，音声と画像の同期・非同期性を学習する方法について述べ，3 章では，環境に応じ音声と画像のストリーム重みを推定する方法について述べ，4 章では，提案するバイモーダル音声認識システムの構成について説明する．そして 5 章で，単語認識実験による評価を行ない，最後に結びを述べる．

# 3 音声・画像情報の学習

## 3.1 合成統合

図 1 に，合成統合の概略を示す．まず，ある音素について，音声と画像の音素 HMM を合成するために，音声・画像同期データから音声データと画像データを抽出する．一般に，音声と画像データはフレームシフトが違うため，画像データを音声のフレームシフトに合うように調整する．そして，各々のパラメータのみで，EM アルゴリズムにより孤立学習と連結学習を行い，音声と画像の音素の HMM を各々作成する [12]．このように，音声と画像の各々のみのデータを用いて，学習することで，結果統合のように，データを有効に利用することができる．

次に，音声と画像の音素 HMM を合成する．このとき，合成した HMM の各状態の出力確率は，

$$b_{ij}(O_t) = b_i^{(a)}(O_t^{(a)})^{\lambda a} \times b_j^{(v)}(O_t^{(v)})^{\lambda v} \qquad (1)$$

のように，音声と画像の出力確率の積として合成される．ただし，$b_i^{(a)}(O_t^{(a)})$ は，時刻 t で，音声 HMM の状態 $i$ において特徴ベクトル $O_t^{(a)}$ を出力する確率，$b_j^{(v)}(O_t^{(v)})$

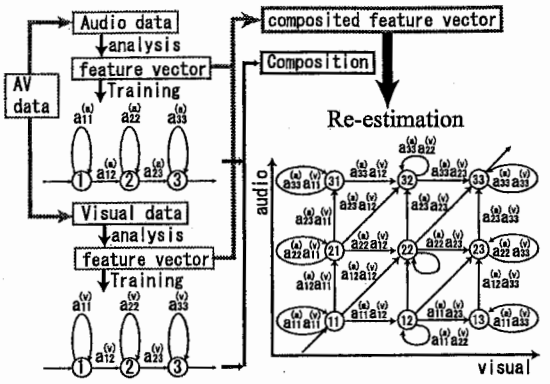



図 1: HMM 合成を用いた統合

は，画像 HMM の状態 $j$ で特徴ベクトル $O_t^{(v)}$ を出力する確率であり，$\lambda a$，$\lambda v$ は各々のストリーム重みである．

また，合成 HMM において，状態 $S_{ij}$ から状態 $S_{kl}$ への遷移確率 $a_{ij,kl}$ は，音声 HMM の状態 $S_i$ から状態 $S_k$ への遷移確率 $a_{ik}^{(a)}$ と画像 HMM の状態 $S_j$ から状態 $S_l$ への遷移確率 $a_{jl}^{(v)}$ を用いて，

$$a_{ij,kl} = a_{ik}^{(a)} \times a_{jl}^{(v)} \qquad (2)$$

となる．そして，この処理を全ての音素について行うことですべての音素 HMM を作成する．合成することにより，認識の際に，結果統合のように音声と画像 HMM の各々の最尤パスを求める必要がなく，音声 HMM と画像 HMM の状態と時刻フレーム方向の 3 次元トレリスを従来のワンパス (Viterbi) アルゴリズムで探索できる [4]．

音声 HMM と雑音 HMM を合成する方法 [14] と比較すると，音声・雑音 HMM 合成では，音声と雑音スペクトルの加法性が成り立つ線形スペクトル領域で出力確率分布を結合しているが，音声と画像では，加法性が成り立たないため，式 (1) のように出力確率分布の積として合成する．また，文献 [4] では，同じように音声と画像の合成しているが，式 (1) のようにマルチストリーム HMM の形にしていないため，環境に応じ音声認識システムを適応化が困難である．さらに，音声と画像で独立に学習を行っているため，音声と画像の同期性が考慮されていない．

そこで，合成 HMM を初期モデルとして，音声と画像の特徴ベクトルを合成した音声画像同期混合ベクトルを用い，EM アルゴリズムにより，再学習を行う [12]．この合成 HMM の学習により，同期性の学習を試みる．

# 4 ストリーム重み推定

ここでは，式 (1) のストリーム重み $\mathbf{\Lambda} = \{\lambda a,\ \lambda v\}$ を推定する方法を説明する．この章では，まず始めに GPD アルゴリズムによるストリーム重み推定について述べる．次に，環境に適応してストリーム重みを推定するための提案手法を述べる．

### 4.1 GPD によるストリーム重み推定

GPD による学習 [6] [10] では，正解のクラスと誤りのクラスの距離の情報を示す誤分類測度を定義する．誤分類測度は，滑らかな損失関数として定式化され，GPD アルゴリズムにより損失関数を最小化するストリーム重みの値が求められる．

ここで，適応のためデータ $x$ をそれに対応する正しい単語 HMM $c$ で，Viterbi アルゴリズムにより認識した時の対数尤度を $L_c^{(x)}(\Lambda)$ とおく．

同様に，適応データ $x$ を誤った単語 HMM で認識した時の対数尤度を $L_m^{(x)}(\Lambda)$ とおく．

そのとき，誤分類測度 $d^{(x)}$ は，

$$d^{(x)}(\Lambda) = -L_c^{(x)}(\Lambda) + \log\left[\exp\left(L_m^{(x)}(\Lambda)\right)\right] \qquad (3)$$

と定義される．この誤分類測度は，小さいほど分類誤り，つまり誤認識が少なくなることを表現する．しかし，$L_c^{(x)}(\Lambda)$ と $L_m^{(x)}(\Lambda)$ は，最尤状態系列での尤度を計算するため，滑らかでない関数になる場合がある．そこで，誤分類測度を用いて，

$$l^{(x)}(\Lambda) = \frac{1}{1+\exp[-\alpha d^{(x)}(\Lambda)]},\ \alpha > 0 \qquad (4)$$

としてシグモイド関数の形に変換し，滑らかな損失関数を定義する．また，勾配の方向を安定させるために，全体の適応データに対して損失関数

$$\mathrm{L}(\Lambda) = \sum_{x=1}^{X} l^{(x)}(\Lambda) \qquad (5)$$

とおく．ただし，$X$ は適応データの総数である．

全体のストリーム重み $\Lambda$ は，GPD アルゴリズムにより

$$\Lambda_{k+1} = \Lambda_k - \epsilon_k \mathbf{E}_k \nabla \mathrm{L}(\Lambda)|_{\Lambda=\Lambda_k} \quad for\ k = 1,.. \qquad (6)$$

と更新される．ただし，$\mathbf{E}$ は単位行列である．$\sum_{k=1}^{\infty}\epsilon_k = \infty$ と $\sum_{k=1}^{\infty}\epsilon_k^2 < \infty$ を満たすと，このアルゴリズムは収束することが証明されている [16] .

### 4.2 環境適応のためのストリーム重み推定

GPD による推定では，式 (6) の更新式の勾配 $\nabla \mathrm{L}(\Lambda)$ が収束に大きな影響を与える．合成統合における HMM は， 音声 HMM の状態数 × 画像 HMM の状態数であり，複雑な HMM 構造になるため，$\mathrm{L}(\Lambda)$ が複雑な関数になる．そして，勾配が不安定になり，最適なストリーム重み値を推定しにくい可能性がある．従って，本論文では，合成統合と同じ方法で，音素 GMM を作成し，その GMM のストリーム重みを GPD アルゴリズムで推定し，バイモーダル音声認識システムを環境適応することを提案する．処理の流れを以下に示す．

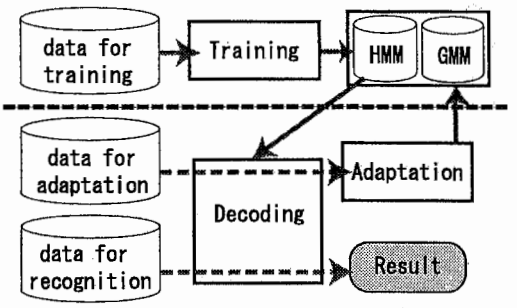



図 2: 提案バイモーダル音声認識システム

**step(1)** 音素 GMM の作成　前章で述べた合成統合と同じ方法で，あらかじめ音素 GMM を作成しておく．このとき HMM から GMM に縮退する方法が考えられるが，より精度の良い GMM を作成するために初期モデルから学習を行った．

**step(2)** ストリーム重みの推定　環境適応用の少数の単語データを用い，本章で述べた GPD アルゴリズムを用い，GMM のストリーム重みを推定する．

**step(3)** 音声認識システムの環境適応　GMM で推定されたストリーム重みを，合成統合で学習した HMM に代入する

音素 GMM を用いることで状態数が少なくなるので，モデル構造が簡単になる．そして勾配の計算量が削減でき，勾配を安定させることもできる．

## 5 バイモーダル音声認識システム

ここでは，提案するバイモーダル音声認識システムの全体的なアルゴリズムについて述べる．図 2 に，本提案システムの流れを示す．図 2 のように，学習データを用いて，あらかじめ合成統合により音素の音声・画像 HMM と GMM を学習しておく．そして，ある環境で発話した少数データをもとに，前章で述べたストリーム重み推定により，あらかじめ学習した HMM を適応化させる．

## 6 認識実験

### 6.1 音声・画像同期データベース

男性話者の音声・唇画像同期データベースを実験に用いた．データベースには，ATR 日本語発声リストの重要語 5240 単語が収録されている．被験者は椅子に座り，発話を行っている．また同時にカラーモニターで唇がカメラの中心位置に収まり，唇周辺領域のみ写るように調整している．収録時には顔に白熱灯をあて，唇を照らす．頭部は特に固定していないが，唇が中心にくるように撮

表 1: 実験条件

| | |
|---|---|
| 音声 | 標本化周波数: 12 kHz<br>分析窓関数: ハミング窓<br>フレーム長: 32 msec<br>フレームシフト: 8 msec<br>パラメータ:MFCC16 次元<br>MFCC$\Delta$16 次元 |
| 画像 | フレームシフト: 33 msec<br>平滑化対数パワースペクトル<br>係数 35 次元<br>ブロック単位の $\Delta$24 次元 |
| HMM 状態数 | 結果統合，合成統合:<br>音声 3, 画像 3<br>初期統合：3 |
| HMM の分布<br>HMM | 2 混合ガウス分布<br>音素環境独立 55 音素モデル<br>Matched モデル<br>(音声 SNR 15dB) |
| 学習データ | 音声・画像同期データ<br>男性話者 1 名，4740 単語 |
| テストデータ<br>(合成統合評価用) | 200 単語 (3 sets) (OPEN) |

表 2: ストリーム重み推定のための実験条件

| | |
|---|---|
| GMM の分布 | 12 混合ガウス分布 |
| 適応データ | 学習データとテストデータ以外の 100 単語データ |
| テストセット | 200 単語 (2 set) (OPEN) |
| 適応時の認識辞書 | テストセットの語彙を含む 500 単語辞書 |

影しているため特に位置ずれはない．また，1 発話の前後で口を閉じるよう指定している．収録は複数日にまたがって行ったため照明条件が発話単語によって異なっている．

## 6.2 音声・画像の特徴分析

音声 HMM の作成には，音声データから MFCC とその時間差分を求め，それを特徴ベクトルとしてモデル作成を行った．また，雑音処理のために，音声が SNR 15 dB となるように白色雑音を加えた音声データで，HMM を学習する．従って，音声は SNR 15 dB の matched モデルとなる．

本システムでは，ユーザーに唇領域を中心に撮影することを想定するので，位置の正規化は特に行なわない．しかし，輝度は照明により変わることが想定されるために，RGB 画像をモノクロ画像に変換し，そのヒストグラム平坦化を行うことで輝度を正規化する．

また，口唇領域画像の静的特徴として，2 次元 FFT を用いた方法 [8] [5] を用いる．まず各フレームに対して，画像を 128 × 128 の大きさに縮小を行う．このとき，スペクトルは歪まないように，縦横は等比で縮小を行い，情報がない画素は 0 で埋める．そして，輝度正規化処理後に，2 次元 FFT を行う．128 × 128 のパワースペクトル係数は，それぞれ対称の関係になっていることから，そのうちの 4 分の 1 の領域を扱う．この処理の後，周波数領域の値に対して，対数パワースペクトルを計算する．その値の数は，128 × 128 と数が多いため，対数スケールでスムージングを行い，6 × 6 の領域にまで削減し，この値を特徴量とする．ただし，直流成分は用いない．2 次元 FFT の手法は，口形状を直接モデル化した方法 [3] ではないが，位相成分を使わないので，唇画像の収録の際に，頭部の動きにより多少の唇の位置ずれが発生しても，画像の"分布"を抽出する手法のために，動きに強く非常に頑健な方法であり，高周波成分をスムージングすることで肌などの細かい特徴を除くことができる．

次に，口唇領域画像の動的特徴として，画像を 6 × 4 のブロックに等分割し，その輝度の平均値を求め，その 2 フレーム分の時間差分を動きの特徴とする [10]．一般に，ブロック単位を増加すると，計算量が膨大になる．さらに，頭部の動きによる位置ずれにより，口形状特徴がとれなくなる可能性がある．今回は，予備実験から経験的にこのブロック単位を設定した．

通常，音声と画像のフレームシフトは違い，音声の方がフレームシフトは小さい．このため，画像は，同じフレームを埋め込み，音声と画像のフレームシフトを調整を行う．

## 6.3 実験条件

表 1 に学習のための実験条件を示す．比較として音声のみ，画像のみ及び音声と画像を初期統合した場合の認識実験も行った．音声のみと画像のみの認識実験は 3 状態の HMM を用いた．そして初期統合も 3 状態の HMM を用いた．HMM の形状は，いずれも left-to-right 型である．

表 2 にストリーム重み推定のための実験条件を示す．ストリーム重み推定のための環境適応用のデータは，学習データとテストデータ以外の 100 単語を用いた．そして，適応用データの中から，15 単語，25 単語と 50 単語を選び，各々 3 セットについてストリーム重み推定実験を行った．また，適応時の辞書の語彙数は，適応データとテストセットを含む 500 単語である．

GPD アルゴリズムの式 (4) において，$\alpha = 1.0$ とし，式 (6) において，$\epsilon_k = 10/k$ とした．そして GPD の更新回数は最大 15 回で打ち切った．

## 6.4 実験結果 1（合成統合の評価）

まず，合成統合と従来の統合方法の認識率を比較する．図 3 に，音声のストリーム重みに対して SNR が 15 dB, 0 dB と -5 dB になるように音声に白色雑音を加えた場合

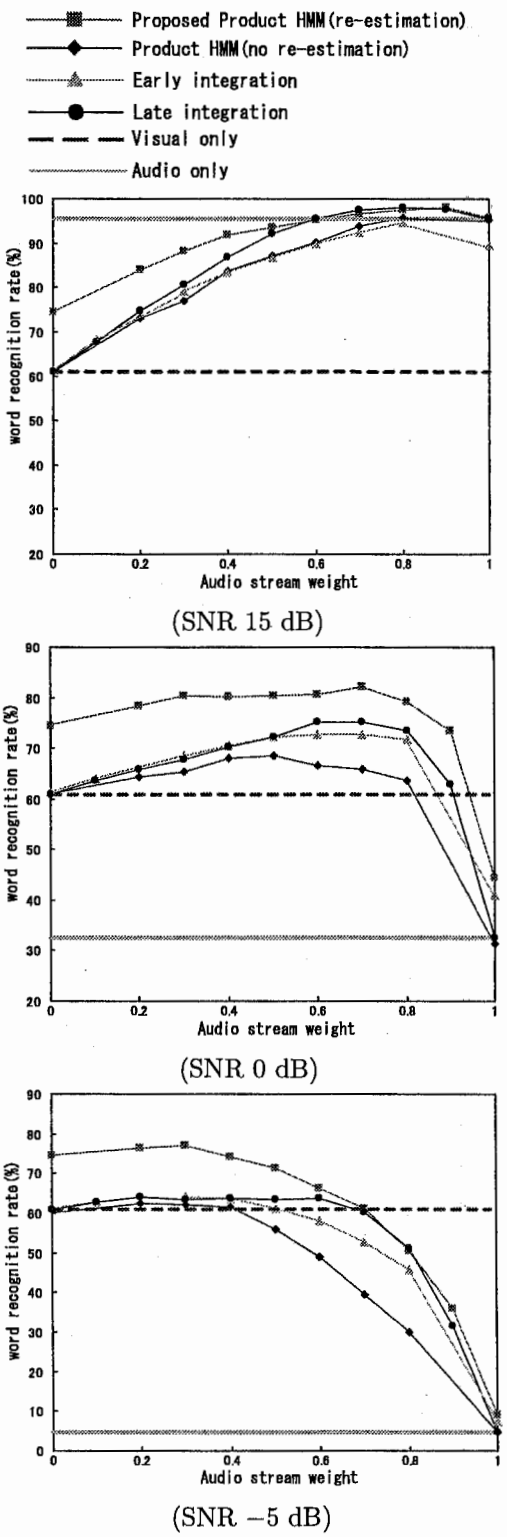
と音声がクリーンな場合の認識率を示す．また，音声のみと画像のみを用いた場合の認識率もあわせて示す．ただし，音声のストリーム重みは式 (1) を満たすようにしているので，音声のストリーム重み値が小さいほど，画像のストリーム重みが大きいことを示している．図 3 から提案手法は，従来手法よりどの音響環境においても（特に音声 SNR が悪くモデルとミスマッチが起っているとき）認識率が高いことが分かる．従来の統合方法と比べて認識率向上の主な理由として次のことが考えられる．

**(1)** 初期統合と比べると，初期統合では学習データから十分に非同期性を表すことが難しいが，提案法は，音声と画像の HMM トレリス空間を探索することによって，非同期性を良く表すことができるからである．

**(2)** 結果統合では，音声と画像を全く独立したモダリティとしているが，提案手法では，合成 HMM を再学習することで，音声と画像の同期性を学習しているからである．

さらに，統合 HMM において，状態数が増えたことが認識性能を改善しているかについて調べるために，単に状態数を増やし，初期統合により，HMM を推定する場合と比較を行った．その結果，HMM のパラメータ数が多くなるためにパラメータ推定がうまくいかず，合成 HMM をもとに学習したものより，高い性能が得られなかった．従って，音声と画像 HMM を合成することでよい初期モデルを与えることができると考えられる．

また，合成した HMM （再学習無）を用いたときの認識率は，結果統合に比べて低い性能となっている．(図 3 )．これは，実装上，合成した HMM は認識時に音素モデル境界で同期して探索してしまうためである．つまり，結果統合では，全く独立した過程で音声 HMM と画像 HMM 学習し音声と画像が完全に非同期な状態遷移を許し学習しているため，合成した HMM の音素境界同期の制約に対しミスマッチが起こっていると考えられる．しかし，音素境界で同期した合成 HMM を再学習することにより，音素境界での同期により認識率が低下する問題は避けることができる．




(SNR 15 dB)

(SNR 0 dB)

(SNR −5 dB)

図 3: 音声ストリーム重みに対する認識率

## 6.5 実験結果2（ストリーム重み推定）

ここでは，GPD アルゴリズムによるストリーム重み推定の精度を詳しく検討する．

図 3 のように，バイモーダル音声認識システムでは，音声と画像のストリーム重みの値により認識率のピークを持つ傾向があることが分かる．この認識率のピークに対するストリーム重みを推定することで，様々な環境でもバイモーダル音声認識システムは，単一モーダリティの音声認識システムより高い認識率が得られることが分かる．

提案手法では，環境適応時に計算量削減のために，GMM のストリーム重みを推定し，それを認識に用いる HMM（本実験では 3 × 3 状態）に代入する．これが問題ないかを確かめるために，GMM のストリーム重みに対する認識率ピークの位置と認識に用いる HMM のピークの位置が類似しているかどうかを確認する必要がある．図 4 に，音声の SNR が 15dB，0dB と -5dB の場合においての，GMM と認識に用いる HMM のストリーム重みに対する認識率を示す．ただし，図 4 の認識率は，図 3 と違い，適応データを除く 2 セットの認識結果の平均である．図 4 から，GMM と HMM のストリーム重みに対しての認識率の関数は似ていることが分かる．そして，ほぼ同じストリーム重み値で認識率のピークを持つため，GMM で推定したストリーム重み値を HMM に代入しても問題ないと言える．

また，図 4 の各々に，GPD アルゴリズムにより推定されたストリーム重みの 3 セットの平均値を示す．ただし，実線の矢印は，25 単語で適応したときのストリーム重みの平均値であり，破線の矢印は，50 単語で適応したときのストリーム重みの平均値である．図から，最も最適なストリーム重みの値を推定しているとは言い難いが，テストセットの認識率のピークに近いストリーム重み値が推定されることが分かる．最も最適な値を推定していない理由は，適応データ数が少なく，適応データとテストセットの発話内容が違うためである．

また，MCE 基準による推定では，適応データの発話内容により推定されるストリーム重みの値が異なる．従って，ストリーム重み推定の値が，発話内容とデータ数によりどの程度ばらつくのかを詳細に調べる必要がある．このための実験として，3 セットの適応データセットにより推定されるストリーム重みとその認識率の標準偏差を調べる．表 3 は，適応データ数が 15，25 と 50 単語の場合において，推定されるストリーム重み値とその認識率の標準偏差を示す．ただし，各々の値は，音声が SNR 15 dB，10 dB，5 dB，0 dB と −5 dB の標準偏差の平均値である．

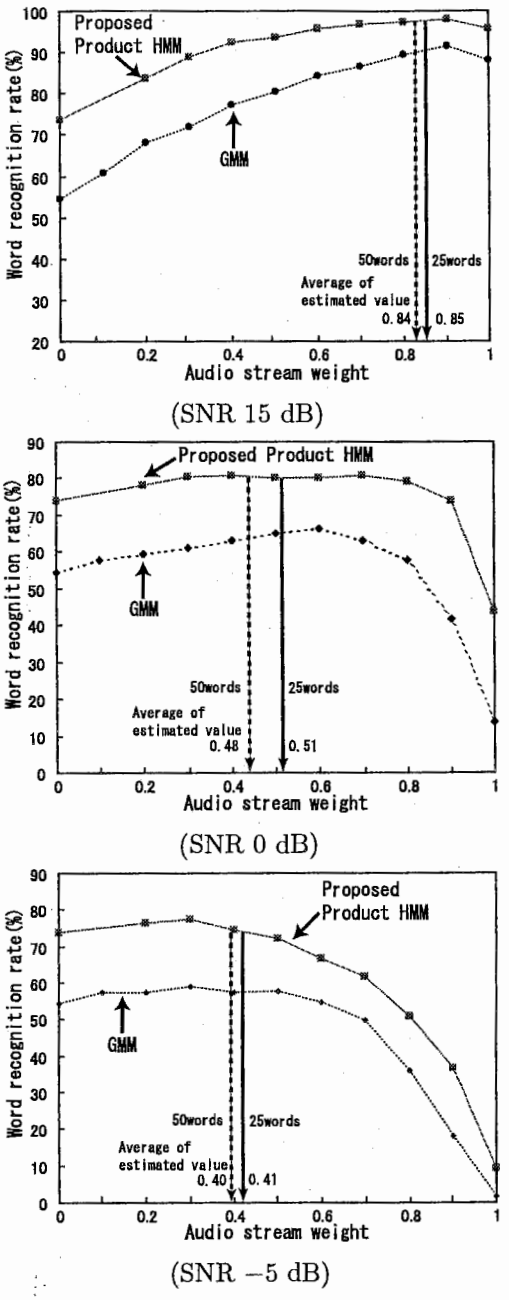



図 4: GMM と HMM のストリーム重みに対する認識率

適応データ数が多くなるほど，推定されるストリーム重みの値はばらつきが少なくなることがわかる．同様に，適応データ数が多くなるにつれ，認識率のばらつきもなくなることがわかる．適応データの量が 15 単語では，まだ安定した推定がされないが，25 単語以上程度であれば，安定した値が得られることが分かる．

## 6.6 システム全体の評価

図 5 に，音声の SNR に対しての認識率を示す．図 5 において，縦軸の認識率は，音声と画像 HMM を合成し，それを再学習して，さらに，25 単語の異なる適応データセットでストリーム重みを推定した場合の認識率の平均

表 3: 推定値と認識率の標準偏差

| The amount of adaptation data | STDV of stream weight | STDV of recognition rate |
|---|---|---|
| 15 words | $1.47 \times 10^{-1}$ | 2.18 |
| 25 words | $8.16 \times 10^{-2}$ | $9.95 \times 10^{-1}$ |
| 50 words | $4.37 \times 10^{-2}$ | $9.08 \times 10^{-1}$ |

STDV = Standard deviation

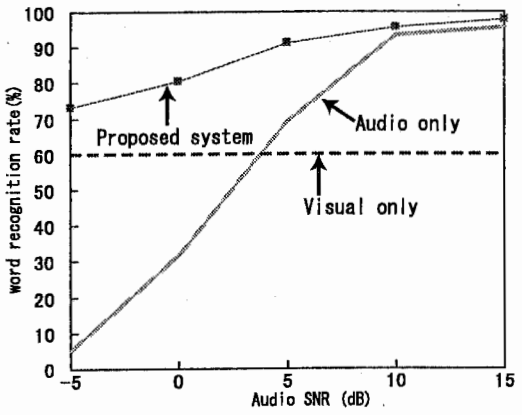



図 5: 音声の SNR に対する認識率

である．各々の状況で，ストリーム重みを推定することより，単一モーダルの音声認識システムより，常に高い認識性能が得られることがわかる．従来，音声がクリーンな場合には，画像は，音声より音素の識別性能が悪いために，バイモーダル音声認識システムは，音声のみの認識率より認識性能が悪くなるが，提案手法により，ストリーム重みを推定することで音声がクリーンな場合でも良い認識性能が得られることがわかる．

## 7 むすび

本論文では，音声と口唇画像を用いたバイモーダル音声認識において，小規模の音声画像同期データベースから，HMM 合成を用い，音声と口唇画像の特徴の同期／非同期性を表現する方法を提案した．評価実験により，従来より高い認識率が得られ，本手法の有効性を確認した．また，GMM と GPD アルゴリズムを用いて，少数のデータから音声と画像のストリーム重みを推定する方法を提案した．評価実験により，ほぼ最適なストリーム重みが推定でき，このことにより，単一のモダリティしか用いない音声認識システムより高い認識性能が得られることが分かった．

今後の検討課題としては，ユーザーに唇を中心に撮影させるという制約をなくすために，顔画像からの唇位置の自動抽出の検討，画像劣化したときの実験，不特定話者タスクへの拡張がまずあげられる．また，音素内だけでなく音素境界でも非同期性を表現するシステムと音素レベルでのストリーム重みの推定なども検討したい．

## 謝辞

本研究の機会を与えてくださった、ATR 音声言語通信研究所 山本誠一社長に感謝します．また，本研究のデータベース作成を行った奈良先端科学技術大学院修士の伊藤秀俊さんに感謝します．

## 参考文献

[1] D. G. Stork and M. E. Hennecke eds., "Speechreading by Humans and Machines", Springer-Verlag, Berlin, 1996.

[2] C. Bregler, H. Hild, S.Manke, and A. Waibel, "Improving connected letter recognition by lipreading",Proc. ICASSP-93, vol.1, pp.557-560, Apr 1993.

[3] Juergen Luettin, Neil A. Thacker and Steve W. Beet, "Visual Speech Recognition using Active Shape Models and Hidden Markov Models", Proc. ICASSP-96, vol.2, pp.817-820, May 1996.

[4] M. J. Tomlinson, M. J. Russell, N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition", Proc. ICASSP-96, vol.2, pp.821-824, May 1996.

[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"，音講論，2-Q-11，Sept. 2000.

[6] Gerasimos Potamianos, Hans Peter Graf, "Discriminative training of HMM stream exponents for Audio-Visual speech recognition", Proc. ICASSP-98, vol.6, pp.3733-3736, May 1998.

[7] Alexandrina Rogozan, Paul Deleglise and Mamoun Alissali, "Adaptive determination of audio and visual weights for Automatic speech recognition", AVSP'97, Rhodes(Greece), pp61-64, 26-27 Sept. 1997

[8] Satoshi Nakamura, Hidetoshi Ito, Kiyohiro Shikano, "Stream weight optimization of speech and lip image sequence for Audio-Visual speech recognition ", Proc. ICSLP2000, vol.3, pp.20-23, 2000.

[9] 熊谷 建一，中村 哲，鹿野 清宏，"バイモーダル音声認識のためのモデル合成に基づく統合法と適応化"，信学技報，SP2000-86，pp67-72，Dec. 2000.

[10] Chiyomi Miyajima, Keiichi Tokuda, Tadashi Kitamura, " Audio -Visual speech recognition using MCE-based HMMs and model-dependent stream weights " , Proc. ICSLP2000, vol.2, pp.1023-1026, 2000.

[11] J.Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition " , Proc. ICASSP-97, vol.2, pp.1267-1270, Apr.1997.

[12] X.D.Huang, Y.Ariki, N.A.Jack, " HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION " , Edinburgh Information Technology Series, EDINBURAGH

[13] 武田 一哉，" 頑健な音声処理手法－多元信号の統合に基づく音声処理 " 信学技報，SP2000-86，pp1-6，Dec. 2000.

[14] S.Nakamura, T.Takiguchi, K.shikano, "Noise and room acoustics distorted speech recognition by HMM composition",Proc. ICASSP-96, vol.1, pp.69-72, May 1996.

[15] S.S.Stevens, "Psychophysics",John Wiley & Sons, New York, 1975

[16] W.Chou, B.-H. Junang, C.-H. Lee, and F.K.Soong, "A minimum error rate pattern recognition approach to speech recognition", J.Pattern Recog.Art.Intell., Col.VIII, pp.5-31, 1994.

図 6: 「あいかわらず」発話時における 1,10,20,25,30,40 フレームの画像

# A 付録

図 6 に実験のために収録した唇動画像のサンプルを示す.