音声言語情報処理 34－12
(2000. 12. 21)

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

†‡熊谷 建一　†中村 哲　‡鹿野 清宏

†ATR 音声言語通信研究所
† 〒619-0288 京都府相楽郡精華町光台 2-2-2

‡奈良先端科学技術大学院大学 情報科学研究科
‡ 〒630-0101 奈良県生駒市高山町 8916-5

E-Mail:{xkumata,nakamura}@slt.atr.co.jp, shikano@is.aist-nara.ac.jp

あらまし　近年，音声認識の性能は大きく改善されたが，さらに，音声の SNR が低い雑音環境での高い音声認識性能が求められている．そのような環境に適した音声認識システムとして，音声情報と唇周辺の動画像を用いたバイモーダル音声認識が注目されている．このようなシステムを構築するためには，音声情報と画像情報の統合が重要な問題となる．統合においては，(1) 音声を発話する前に発声の準備のために唇が動き，発話が終わった後に遅れて唇が閉じるといったような，音声と唇周辺の動きの非同期性，(2) 周辺環境に応じたシステムの適応化，といった問題がある．本稿では，まず (1) の問題に対し，音声と唇周辺の動きの非同期性を考慮する HMM 合成に基づいた統合を行う．次に (2) の問題に対しては，GPD アルゴリズムを用い，少数の環境適応用のデータ (以下適応データ) からストリーム重みを推定することを検討する．音響的な雑音がある場合について，単語認識実験を行った結果，認識性能が改善されることが示された．

キーワード　バイモーダル音声認識，隠れマルコフモデル，ストリーム重み，最小分類誤り，GPD アルゴリズム

An Adaptive Integration Method Based on Product HMM for Bi-modal Speech Recognition

†,‡Kenichi KUMATANI　†Satoshi NAKAMURA　‡Kiyohiro SHIKANO

†ATR Spoken Language Translation Research Laboratory
†2-2-2 Hikaridai, Seika, Kyoto, 619-0288 JAPAN

‡Nara Institute of Science and Technology
‡8916-5 Takayama, Ikoma, Nara, 630-0101 JAPAN

Abstract　In recent years, there has been higher demands for *Automatic Speech recognition system* operated robustly in the various noisy environments. Therefore, many researchers have interest in the bimodal speech recognition by using not only the audio but also the visual information extracted from the sequence of the speaker's lip images. To realize the bimodal speech recognition, it is important to integrate effectively the audio and visual information. In integrating them, "(1) Synchronization of the audio and visual information, (2) Adaptability of the system, adjusting to changes in environment" are important issues. In the problem of (1), each feature of the speech and lip movement has the time lag, and has the correlation. For such the problem, we introduce the integration method using HMM composition. In (2), we have examined that the stream weight can be adaptively estimated by GPD algorithm. The evaluation experiment shows that the proposed method improves recognition accuracy of noisy speech.

key words　Bi-modal speech recognition, HMM, Stream weight, MCE, GPD algorithm

1 はじめに

近年，話者の周辺が非常に騒がしいといったような音声の SNR が低い様々な場所において，音声認識システムが求められている．このような環境に適した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]-[6]．音声と画像は，全く異なった雑音や劣化が起こる為に，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．例えば，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，周りが騒がしかったり，話者の声の大きさが小さくても劣化は起こらないため，音声の SNR が低い状況では，音声より高い認識性能を示すということがあげられる．

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，音声情報と画像情報をどのように統合するかということと，周辺の環境に応じてどちらの情報を重視するかを決定することが重要な問題となる．

前者の問題に対しては，音声 HMM と画像 HMM を合成し，合成した HMM を再学習を行い音声と画像情報を統合する HMM 合成による統合方法 [4]（以下合成統合）を用いる．合成統合は，従来の統合方法よりも優れた認識性能をもつ．

後者の問題は，具体的には，合成統合により，統合を行った HMM の認識率がピークとなる音声と画像のストリーム重みを，ユーザが発話した適応データから，環境に応じ適応化することとなる．しかし，音声の SNR を推定するのは難しいので，ストリーム重みを推定するためには，他の基準が必要となる．通常，音声と画像の尤度のダイナミックレンジが大きく違うために，ML（尤度最大化）基準による学習では，良い性能が得られない [2] [7]．それに対し，MCE（最小分類誤り）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 [2] [3]．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 [3]，GPD アルゴリズムによる方法 [2] [5] 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．しかし，音声と画像のストリーム重みは，音素ごとに違い，従って，適応データ数に応じて，ストリーム重みの tying の単位は音素クラスごとに分割したほうが良いと考えられる．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，本稿では，木構造を用い，適応データ数に応じストリーム重みの tying の単位を分割し，GPD アルゴリズムにより，環境に応じストリーム重みを適応化することを検討する．

本稿では，まず次章で，合成統合を紹介し，第三章で，GPD アルゴリズムを用い，環境に応じストリーム重みを適応化する方法を述べ，最後に評価結果を報告する．

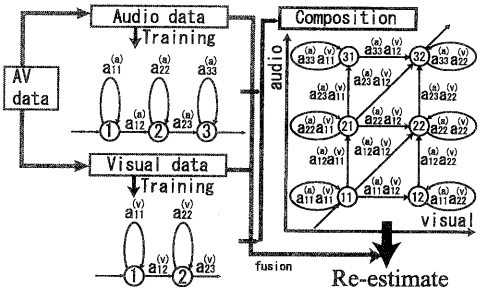

図 1: HMM 合成を用いた統合

2 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 [10]．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，$b_j^{(v)}(O_t^{(v)})$ 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，λa，λv 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．

音声 HMM と雑音 HMM を合成する方法 [8] 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 [6] 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．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，孤立学習と連結学習

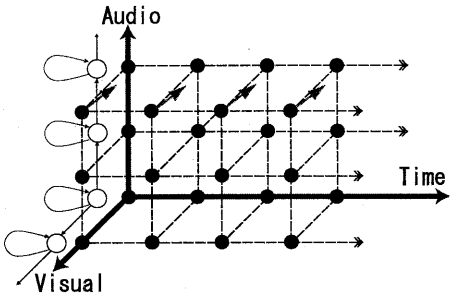

図 2: 合成統合の探索空間

を行う [10] . この合成 HMM の学習により，同期性を表現できる．

図 2 のように，認識の際に，合成統合による探索空間は，音声 HMM と画像 HMM の状態と時刻フレーム方向の 3 次元トレリスを探索することになり，音声と画像の状態を非同期に探索可能となる [4] .

3 ストリーム重みの環境適応

3.1 GPD アルゴリズムによるストリーム重み推定

GPD による推定 [2] [5] では，正しい分類と誤った分類との距離の情報を表す誤分類測程度を含む，滑らかな損失関数を最小化するように，HMM のストリーム重みを推定する．ここでは，GPD に基づくストリーム重みを推定する手続きを述べる．

まず，ある単語の発話 x の特徴ベクトル系列を $\mathbf{O} = [\mathbf{o}_x(1), ..., \mathbf{o}_x(t), ..., \mathbf{o}_x(T_x)]$ とする．ここで，t は時刻フレーム，$\mathbf{o}_x(t)$ は S 個のストリーム (モダリティ) をもったベクトルである．

次に，HMM の状態のある集合 $\mathbf{C}$ に対するストリーム重みセットを $\boldsymbol{\lambda}_c = [\lambda_{c1}, ..., \lambda_{cs}, ..., \lambda_{cS}]$ とし，全体のストリーム重みセットを $\Lambda = [\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_c, ..., \boldsymbol{\lambda}_C]$ とする．ただし，C は，ストリーム重みのクラス数である．

そのとき，ある単語の発話 x を，それに対応する単語 HMM で，Viterbi アルゴリズムで認識した時の，HMM の状態系列を $\mathbf{Q}_x = \{q_x(t); t = 1, ..., T_x\}$ とすると，そのときの対数尤度 L_x^R は，

$$L_x^R(\Lambda) = \sum_{j=1}^{J}\sum_{s=1}^{S} \lambda_{js} L_{x,js}^R \tag{3}$$

$$L_{x,js}^R = \frac{1}{T_x}\sum_{t=1}^{T_x} \delta_{q_x(t)}^j \log b_{js}[\mathbf{o}_{x,s}(t)] \tag{4}$$

のように，ストリーム重みのセット Λ の関数として表すことができる．ただし，式 (4) において，$b_{js}[\mathbf{o}_{x,s}(t)]$ は，状態 j において，ストリーム s の特徴ベクトル $\mathbf{o}_{x,s}(t)$ を観測する確率，$q_x(t) = j$ なら $\delta_{q_x(t)}^j = 1$，$q_x(t) \neq j$ なら $\delta_{q_x(t)}^j = 0$ である．

同様に，単語の発話 x に対して，誤った単語 HMM の中で，n 番目の候補により認識した場合の対数尤度 L_x^{Fn} は，

$$L_x^{Fn}(\Lambda) = \sum_{j=1}^{J}\sum_{s=1}^{S} \lambda_{js} L_{x,js}^{Fn} \tag{5}$$

と表すことができる．

次に，誤分類測度 d_x を

$$d_x(\Lambda) = -L_x^R(\Lambda) + \log\left[\frac{\sum_{n=1}^{N}\exp\left(L_x^{Fn}(\Lambda)\right)}{N}\right] \tag{6}$$

と定義する．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，式 (3), (5) は，最尤の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，

$$l_x(\Lambda) = \frac{1}{1+\exp[-\alpha d_x(\Lambda)]},\ \alpha > 0 \tag{7}$$

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

$$\mathrm{L}(\Lambda) = \sum_{x=1}^{X} l_x(\Lambda) \tag{8}$$

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 Λ は，GPD アルゴリズムにより

$$\Lambda_{k+1} = \Lambda_k - \epsilon_k \mathbf{E}_k \nabla \mathrm{L}(\Lambda)|_{\Lambda=\Lambda_k} \quad for\ \ k = 1, 2, ... \tag{9}$$

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 [9] .

3.2 ストリーム重みの更新式

ここでは，実際に，式 (9) を計算するための，式の展開を述べる．ただし，簡潔に記述するために (Λ) を省略する．

まず GPD アルゴリズムで，各々のストリーム重みのクラス c に

$$0 \leq \lambda_{cs} \leq 1\ \ and\ \ \sum_{s=1}^{S}\lambda_{cs} = 1 \tag{10}$$

の制限を加えるために，

$$\lambda_{cs} = \frac{\exp(\bar{\lambda}_{cs})}{\sum_{s'=1}^{S}\exp(\bar{\lambda}_{cs'})} \tag{11}$$

を満たす変換

$$\bar{\lambda}_{cs} = \log\lambda_{cs} \tag{12}$$

を行う．

そして，式 (9) によりストリーム重みを更新するために，式 (7) , (8) から，

$$\frac{\partial \mathrm{L}}{\partial \bar{\lambda}_{cs}} = \sum_{x=1}^{X} \alpha l_x(1-l_x)\frac{\partial d_x}{\partial \bar{\lambda}_{cs}} \tag{13}$$

を計算する．ここで

$$\frac{\partial d_x}{\partial \bar{\lambda}_{cs}} = -\frac{\partial L_x^R}{\partial \bar{\lambda}_{cs}} + \frac{\sum_{n=1}^{N}\frac{\partial L_x^{Fn}}{\partial \bar{\lambda}_{cs}}\exp(L_x^{Fn})}{\sum_{n=1}^{N}\exp(L_x^{Fn})} \tag{14}$$

$$\frac{\partial L_x^B}{\partial \lambda_{cs}} = \sum_{j \in \mathbf{C}} \lambda_{cs}[L_{x,js}^B - \sum_{s'=1}^{S} \lambda_{cs'} L_{x,js'}^B] \qquad (15)$$

となる．ただし，$B = R, Fn$ ，$\mathbf{C}$ はストリーム重みの値を tying する HMM の状態の集合である．式 (7)，(13)-(15) を計算し，式 (9) によりストリーム重みを更新する．

最後に，各ステップの更新後に式 (11) により変換する．

3.3 木構造を用いたストリーム重みの tying の単位の細分化

本稿では，音素 HMM のストリーム重みを基本単位とし，適応データ数に応じ，ストリーム重みの tying の単位をトップダウンに分割していく方法を検討する．

まず，HMM のクラスタリングを行う基準となる木構造を作る．木構造を作成する手順として，複数の質問を用意し，それらの質問に対して HMM のクラスタリングを行う．今回の実験で用いた質問は，HMM が母音か子音のどちらであるか？，有声音か無声音のどちらであるか？，調音位置が唇周辺であるかどうか？の三項目である．このように，クラスタリングを行うことで，音声の先見知識をストリーム重み推定に組み込むことができる．

そして，あらかじめ用意された複数の質問から，一つの質問を選択し，HMM をクラスタリングする．質問には，予備実験で最も認識性能の良かった”有声音か無声音であるか”の質問を選択した．適応時に，損失関数 (式 (7)) を最小化する質問を選択する方法が考えられるが，損失関数は認識性能に必ずしも一致せず，適応時の計算量の増加を招いてしまう．従って，このように，あらかじめ作成した木構造を用いて適応時に，ストリーム重みの tying の単位を分割していくことにした．処理の流れは以下の通りである．

1. 初期のストリーム重みは，全ての HMM について同じとする．つまり,全ての HMM について，ストリーム重みを tying する
2. 同じ木の階層にある各々のノードについて {
   2-1 同じクラスタに属する HMM のストリーム重みを tying
   2-2 各々のストリーム重み初期値＝上の階層で推定された値
   2-3 if (クラスの適応データ数＜閾値) then
   ストリーム重みを定数とする
   else
   ストリーム重みを変数とし，更新対象とする
   }
3. if (更新対象となるストリーム重みがない) then 処理終了
4. GPD アルゴリズムにより，n 回の更新を繰り返し，ストリーム重み更新し，2 の手続きへ戻る

この手法を用いる理由として，ルートノードから，順に，

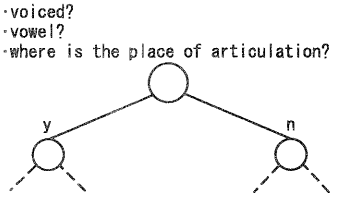

図 3: 木構造

表 1: 実験条件

音声	標本化周波数: 12 kHz 分析窓関数: ハミング窓 フレーム長: 32 msec フレームシフト: 8 msec パラメータ:MFCC16 次元 MFCC△16 次元
画像	フレームシフト: 33 msec 前処理 1: rgb → 256 階調の濃淡画像 前処理 2: ヒストグラム平坦化 前処理 3: 唇位置の正規化 パラメータ: 平滑化対数パワースペクトル 35 次元 平滑化対数パワースペクトル △35 次元
HMM 状態数	音声 3, 画像 2
確率密度関数 HMM	Gaussian: 2 Mixture 音素環境独立 55 音素モデル
学習データ	音声・画像同期データ 女性話者 1 名，4740 単語
テストデータ	200 単語 (OPEN) × 2 セット
適応データ	学習データとテストデータ以外の単語データ
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

ストリームを推定し，それを初期値として用いることで，安定した解に推定されるということと，適応データ数に応じ，精度の良い HMM の適応化が行われるということがあげられる．ただし，計算時間は木の深さが大きくなるにつれ増加し，膨大な量となってしまう．本稿では，分割の有効性を確認することを第一の目的にし，木の深さは最大 2，GPD の繰り返し回数 n は，最大 8 回とした．

4 実験

4.1 実験条件

評価実験として，200 単語 × 2 セットの認識実験を行った．評価として，2 セットの単語認識率の平均を用いた．表 1 に実験条件を示す．本研究では，音響実験室で，特定話者 (女性話者 1 人) が ATR 発声リストの 5240 単語を発話しているデータベース [3] を用いた．

音声と比べ画像のフレームシフトは長いため，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．また，収録した画像データは発話単語により，照明条件の違いや顔の傾きなどが見られる．そこで前処理として，ヒストグラム平坦化，基準フレームとの輝度の差分を最小化するように唇位置の正規化を行った．

音声 HMM の作成には，音響実験室で収録したクリー

ンな音声データから MFCC 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，画像 HMM は，前処理後の画像に 2 次元 FFT を行い，対数パワースペクトルを求める．そして，その周波数領域を 6×6 の領域分割を行い，直流成分を除いた領域の平滑化対数パワースペクトルを特徴ベクトルとしてモデル作成を行った [1]．本実験では，音声・画像 HMM は，各ストリーム重みを 1：1 と等しい重みで学習を行っている．

また，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．音声のみの実験は 3 状態の HMM，画像のみの実験は 2 状態の HMM そして初期統合は 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．

適応時の実験条件として，適応データは，学習データとテストデータ以外の単語発話データを用いた．従って，適応データは，テストデータと発話内容は異なっている．また，適応データ数を，15，25，50，75 及び 100 単語とした場合についてストリーム重み推定を行った．ただし，適応データ数が 15 単語の場合は，発話内容により推定されるストリーム重みが大きく異なる．そのため，適応データ数が 15 単語の場合は，適応データ 3 セットについての認識率の平均とする．適応時の辞書は，適応データの単語とテストデータの単語を含む 500 単語の辞書を用いた．

誤分類測定度の式 (6) において，誤りの候補数を $N = 1$，GPD アルゴリズムの式 (7) において，$\alpha = 0.1$ とした．また，式 (9) において，全てのストリーム重みが tying されているとき $\epsilon_k = 200/k$，ストリーム重みの tying の単位を分割した後は，$\epsilon_k = 100/k$ とし，全てのストリーム重みを tying したときよりも，緩やかに収束させている．

4.2 実験結果

まず，合成統合と他の統合方法の認識率を比較する．図 4, 5, 6 に，音声のストリーム重みと画像のストリーム重みを 式 (10) を満たすように，音声のストリーム重みを変化させたときの初期統合と合成統合の認識結果を示す．また，音声のみと画像のみの認識率もあわせて示す．図 4 は，SNR が 10 dB になるように音声に白色ガウス雑音を加えた場合の認識結果，図 5 は，同様に SNR が 20 dB のときの認識結果である．そして，図 6 は，収録データにノイズを加えていない場合の認識結果である．さらに，各々の図に，50 単語の適応データから GPD アルゴリズムで推定されたストリーム重みの値を示す．ただし，推定したストリーム重みは，再学習を行った合成統合の場合である．各々の図から，バイモーダル音声認識システムは，あるストリーム重みの値で認識率のピークをもつ傾向があり，このピークを推定することで単一モーダルの認識システムより高い認識性能が得られることが分かる．そして，GPD アルゴリズムによって，認識率のピークに近いストリーム重みの値が推定できることが分かる．また，合成した音声・画像 HMM を再学習する合

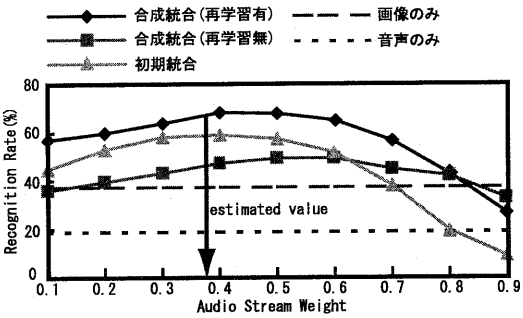

図 4: 音声の SNR 10 dB の場合の認識率

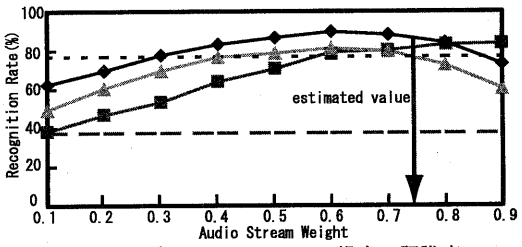

図 5: 音声の SNR 20 dB の場合の認識率

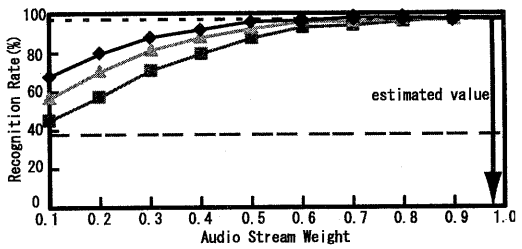

図 6: 音声のクリーンな音声の場合の認識率

成統合 (再学習有) は，初期統合と再学習しない合成統合 (再学習無) よりも高い認識性能が得られることが分かる．これは，初期統合は，音声と画像が同期していると仮定し，再学習しない合成統合は，同期性を学習していないが，再学習する合成統合は，音声と画像の同期関係を学習しているためであると考えられる．また，予備実験で音声と画像 HMM を合成せずに，単に HMM の状態数を増やし形状を変えて，音声・画像データで学習した場合は，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，高い性能は得られなかった。このことから，音声と画像 HMM を合成することで，良い初期モデルを与えることができると考えられる．さらに，学習できない場合も，合成モデルを初期モデルとしてそのまま使うことができる．

次に，ストリーム重みを分割せずに，音声と画像のストリーム重み値を GPD で推定した場合の実験結果を考察する．表 3 に，音声がクリーンな場合及び音声の SNR が 20 dB，10 dB となるように白色ガウス雑音を加えた場合に，音声と画像のストリーム重みを分割せずに環境適応したときの認識率を示す．なお，適応データ数の ( )

表 3: 音声・画像ストリーム重みを
環境適応した場合の認識率

適応データ数 \ SNR	clean	20dB	10dB
15 単語 (108)	96.86 %	77.15 %	56.35 %
25 単語 (193)	97.28 %	89.36 %	69.06 %
50 単語 (366)	97.28 %	87.38 %	68.57 %
75 単語 (521)	97.03 %	83.42 %	65.60 %
100 単語 (697)	97.03 %	87.38 %	68.81 %

表 4: 音声・画像ストリーム重み
を 2 分割した場合の認識率

適応データ数 \ SNR	clean	20dB	10dB
15 単語 (58,50)	97.03 %	77.39 %	61.96 %
25 単語 (104,89)	97.52 %	89.36 %	68.57 %
50 単語 (197,169)	97.28 %	87.87 %	66.34 %
75 単語 (266,255)	97.03 %	83.91 %	65.85 %
100 単語 (365,332)	97.28 %	87.38 %	68.57 %

は，(適応データに含まれる音素数) を表す．また，適応データは無作為に選んでいる．表 3 に示す通り，適応データ数が 15 単語であるとき，低い認識率になる．これは，少数の適応データ数から，音声と画像のストリーム重みを推定するとき，その値がテストセットに対して，最適なストリーム重み値から外れてしまうためである．そこで，適応データの内容により，どのぐらい認識率が変わるのかを調べるために，適応データが 15 単語の場合と 50 単語の場合について，3 回の認識実験を行い，認識率の分散を調べた．適応データ数が 15 単語の場合において，認識率の標準偏差は，10.18 となり適応データの発話内容で認識率がばらついていた．それに対し，50 単語の標準偏差は，0.57 となり，適応データの違いで認識率のばらつきはほとんどなかった．(ただし，標準偏差は，音声がクリーン，SNR 20 dB 及び 10 dB の平均値である．) 従って，少数の適応データから，音声と画像のストリーム重みを推定する場合は，適応データの発話内容を注意して選ばなければならないことがわかる．また，表 3 から，適応データ数が多いほど，適切なストリーム重みが推定されることがわかる．

最後に，ストリーム重みを 2 分割をした場合の実験結果を考察する．表 4 に，適応データ数を変化させて，ストリーム重みの tying の単位を 2 分割し，音声と画像のストリーム重みを推定した場合の認識率を示す．表 4 の適応データ数の (,) は，(有声音の音素数，無声音の音素数) を表す．なお，適応データは，表 3 と同様のものを選んでいる．実験では，更新する適応データ数の閾値は制限していない．従って，すべてのストリーム重みのクラスが更新されている．表 3 と 表 4 を比べると，適応データが 50 単語以上になると，少し認識率が高くなっている場合があるが，それほど差は見られない．適応データ数が 15 単語であるとき，ストリーム重みを分割しない場合より，認識率が高い．これは，ストリーム重みの tying の単位を分割することで，一方のストリーム重みのクラスがテストセットに対して最適な値に近い値が推定されたことと，単に GPD の繰り返し回数が増えたことがあげられる．

5 まとめ

本稿では，音声と画像情報を HMM を用い，合成統合を行った．そして，さらに，合成統合された HMM のストリーム重みを環境適応する手法を提案し，評価を行った．その結果，良い認識性能が得られることを確認できた．また，ストリーム重み推定に必要な適応データ数を考察した．さらに，ストリーム重みの tying の単位を分割した場合では優位な効果が得られなかった．このため，さらなる検討が必要と考えられる．

今後の課題としては，バイモーダル音声認識システムの不特定話者への発展を考え，画像の特徴ベクトル，本手法のストリーム重みの環境適応への効果だけでなく，話者適応への効果，木構造を作成するときの自動化 [11]，本手法の実行速度を向上するためにデコードの最適化を検討する予定である．特に，不特定話者への発展のために，バイモーダルデータをさらに収録する予定である．

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する。

参考文献

[1] Satoshi Nakamura, Ron Nagai, Kiyohiro Shikano, "Improved bimodal speech recognition using tied-mixture HMMs and 5000 word Audio-Visual Synchronous database", Proc. Eurospeech, Rhodes, pp.1623-1626, 1997.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 篠田 浩一, 渡辺 隆夫, "情報量基準を用いた状態クラスタリングによる音響モデルの作成", 信学技報, SP96-79, pp.9-15, Dec.1996