TR-S-0021

マルチモーダル音声認識のための音声と発話顔画像の同期のモデリング

On Audio and Video Modality Synchronizing Model for Multimodal Speech Recognition

田村哲嗣　熊谷健一
Tetsji Tamura　Ken-ichi Kumatani
村井和昌　中村哲
Kazumasa Murai　Toru Nakamura

2001.3.26

近年、雑音下における頑強な音声認識システムとして、音声雑音から影響を受けない画像情報を用いたバイモーダル音声認識の研究が行われている。このバイモーダル音声認識を行うためのモデルの構成法のひとつとしてHMM合成法があり、効率よく認識制度の高いモデルを生成することができる反面、音声情報と画像情報の同期のミスマッチにより性能が低下しているという問題がある。本研究では、このミスマッチの問題を解決するための新たなモデルの提案を行い、実験によってこのモデルの有効性を検討する。

1 はじめに — バイモーダル音声認識と本研究の背景

今日、雑音下におけるより頑健な音声認識システムとして、唇動画像を用いたバイモーダル音声認識が注目され研究が行われている。現在の音声のみによる認識技術では SNR が小さな環境においては認識率が低く不十分であるという問題が指摘されているが、バイモーダル音声認識は、音声雑音からの影響を受けない画像情報、とりわけ発声調音器官の一部である口唇の画像を利用することにより、両者が相互に協調し合うことで高い認識性能を実現できるシステムとして考えられている。電車内や街頭といった周辺環境が騒がしい状況下においても十分な性能を発揮することが期待できるバイモーダル音声認識は、実用的側面から見ても非常に重要な技術であり、近年の音声認識を利用した電化製品や OA 機器の普及もあいまって、いま最も注目されている技術のひとつと言える。

このような背景のもと、バイモーダル音声認識システムとしてはさまざま手法が提案されているが、現在では HMM を用いた手法が主として研究されるようになっている。その理由としては、HMM は既存の音声認識システムでも広く用いられているために組み込みが容易であること、統計・確率的モデルであり音声と画像というふたつの異なるモダリティを独立性を保ったまま融合できること、統計量を用いるので学習データ量に応じて性能を向上させることができること、などが挙げられる。このような HMM を用いた手法のひとつとして、音声と画像の HMM を作成しそれを 2 次元的に合成する HMM 合成法がある。HMM 合成法は音声 HMM と画像 HMM を個別に作成するので個々のモダリティにおいて有効に学習できるだけでなく、HMM を合成した後にも音声・画像統合データにより、再度学習を行うことで両者の同期性を表現できるというメリットがある [1]。しかしながら、このようにして得られたモデルは音素単位の HMM により構成されているために音素境界において遷移パスが制限されてしまい、結果として音声と画像が同期を強制されることになってしまうという問題が明らかになってきた（図 1 参照）。

そこで本研究では、HMM 合成法によるバイモーダル音声認識システムの改良のひとつとして、まず音素境界付近でおこる音声と画像のアライメントのずれに関して、これを確認するとともに試験的なモデルを作成してその効果を検証する。しかる後に、この予備実験の結果を考慮して新たな合成モデルの提案を行い、その認識性能を調べ、考察を行う。

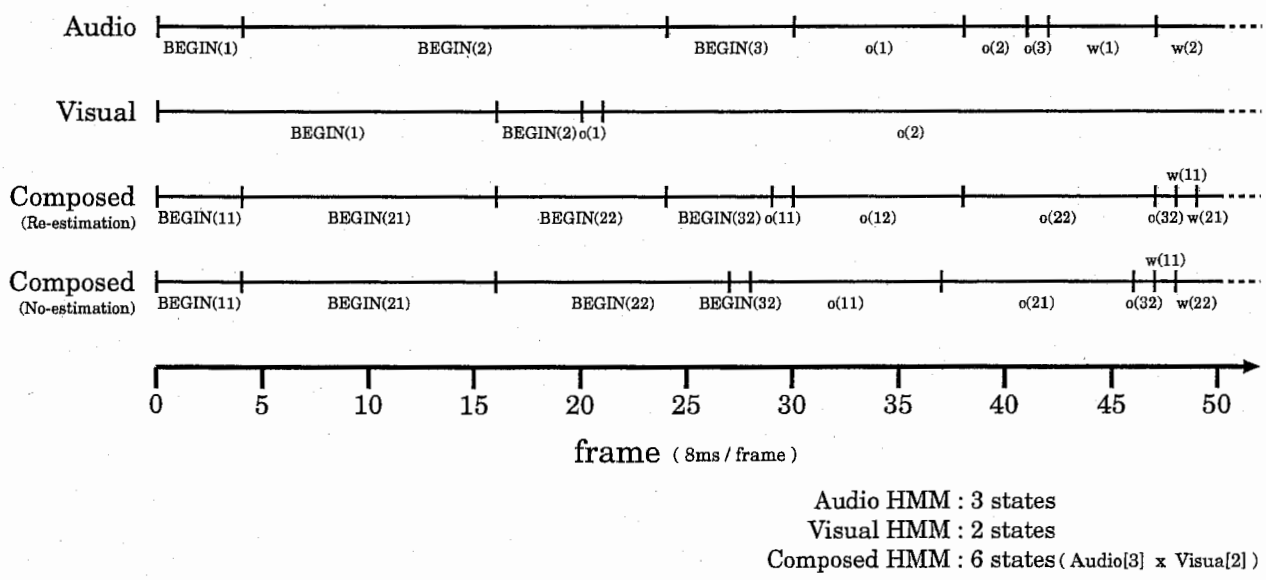


図 1: 音声と画像の同期のずれの例

2 HMM 合成法 — その手法と利点および課題

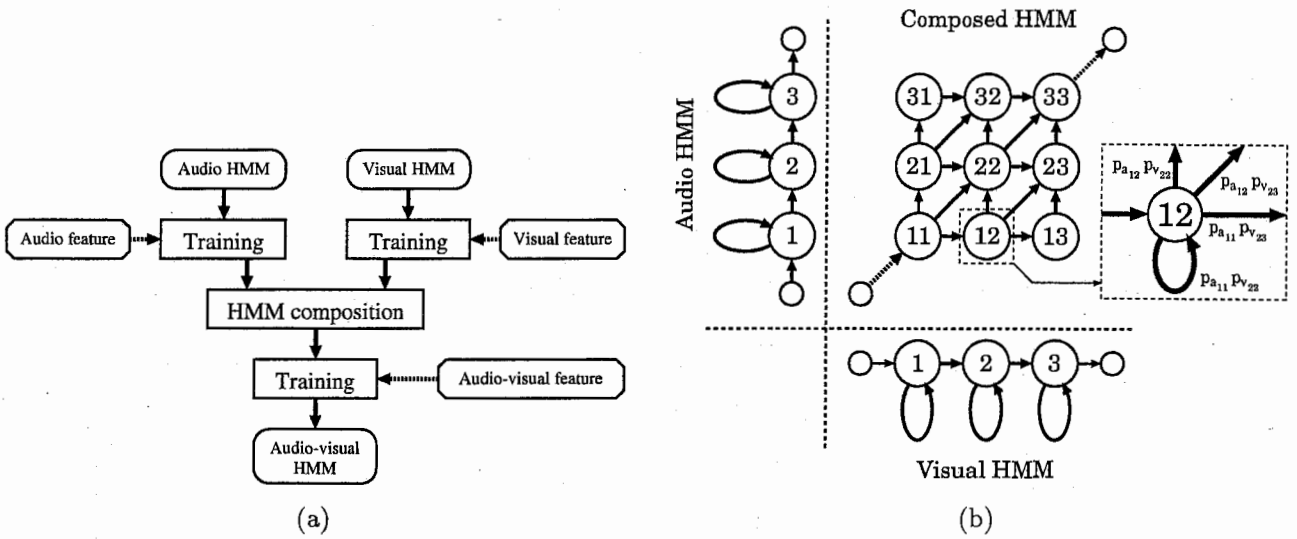


図 2: HMM 合成法

はじめに、本研究で用いている HMM 合成法について、バイモーダル音声認識に特化して簡潔に説明する（図 2 (a) 参照）。

まず第一に、合成するもととなる音声 HMM および画像 HMM を生成する。本研究では、収録した音声・画像同期データから音声データと画像データをそれぞれ分離し、画像データに関しては音声データに合うようにフレーム周期を調整した上で、ラベル付き学習と連結学習によって音声 HMM と画像 HMM を作成した。以上で得られた音声 HMM、画像 HMM を合成して、新たに音声・画像 HMM を生成する。このときの HMM 合成の例（音声 HMM、画像 HMM ともに状態数 3 ）を図 2 (b) に示す。音声・画像 HMM は音声と画像の 2 つのストリームを持つマルチストリーム HMM として構成される。具体的には、音声 HMM の状態 S_i と画像 HMM の状態 S_j から音声・画像 HMM の状態 S_{ij} を合成するとき、音声 HMM の状態 S_i がもつ平均、分散、混合重みなどの情報を音声ストリームのパラメータ、同様に画像 HMM の状態 S_j がもつ情報を画像ストリームのパラメータとする。一方、状態間の遷移については、音声・画像 HMM において S_{ij} から S_{kl} に遷移する確率 $p_{ij,kl}$ は、

$$p_{ij,kl} = p_{a_{i,k}} \times p_{v_{j,l}}$$

により与えられる。ここで $p_{a_{i,k}}$ は音声 HMM における S_i から S_k への、$p_{v_{j,l}}$ は画像 HMM における S_j から S_l への遷移確率である。ただし遷移不可能な状態の組み合わせについては考えず、このために $\sum^{kl} p_{ij,kl} < 1$ となる S_{ij} が出てくる場合があるが、このときは次のようにして確率を正規化する処理を行う。

$$p_{ij,kl} = \frac{p_{ij,kl}}{\sum^{mn} p_{ij,mn}}$$

また、ここで得られた音声・画像 HMM における状態 S_{ij} で音声・画像特徴量 O_t を観測する確率 $b_{ij}(O_t)$ は次式のように表される。

$$b_{ij}(O_t) = b_{a_i}(O_{a_t})^{\lambda_a} \times b_{v_j}(O_{v_t})^{\lambda_v}$$

ここで $b_{a_i}(O_{a_t})$ は時刻 t における音声 HMM の状態 S_i で特徴ベクトル O_{a_t} を観測する確率、$b_{v_j}(O_{v_t})$ は画像 HMM の状態 S_j で特徴ベクトル O_{v_t} を観測する確率、λ_a および λ_v はそれぞれ音声ストリーム重み、画像ストリーム重みである。ここで合成した音声・画像 HMM は、先述したようにもとの音声 HMM と画像 HMM が別々に学習されているために、音声と画像の同期については考慮されていないモデルとなっている。そこで文献 [1] のように、合成した HMM を統合した音声・画像特徴量を用いて再学習するなどの処理によって、同期性を表現し、より良いモデルを構築することができる。

この HMM 合成法の利点としては、音声 HMM と画像 HMM を別々に作成できるので、音声・画像同期データがなくてもこれらを学習し合成することが可能であるという点がひとつ挙げられる。もちろん音声・画像同期データベースから分離されるデータに加えて、さらに別のモノモーダルデータベースを使って学習することも可能である。また HMM 合成して得られる音声・画像 HMM はこの段階で既にある程度の認識性能を有しており、音声・画像同期データを用いた再学習ができない、またはデータが少なく再学習による改善があまり望めないといった状況であっても有効に機能するという特徴がある。

その一方で、例えば発声前に口が動くといったように、実は音声と画像のイベントには時間的なずれが生じているのだが、このずれは音素内だけでなく音素（発声開始／終了を含む）間にも存在すると考えられ、前者が合成した音素 HMM の中で吸収できるのに対し、後者は monophone レベルの音声・画像モデルを用いている場合にはこの同期のずれに対処することができないという問題がわかってきた。triphone レベルのモデルを用いる方法もあるが、これには大量の音声・画像同期データが必要となり、現時点では利用できるデータが少ないなどの問題があるため適用することができない。そこで、少量のデータであってもこのミスマッチングの問題を解決し、より頑健な認識を行うことのできるモデルが必要となってくる。

次章ではこの点について考慮したモデルを生成して予備実験を行い、音声と画像のミスマッチングを確認するとともに、その解決手段について考えていく。

3 予備実験 — 音声・画像の音素境界での非同期性を考慮したモデルの提案

従来の HMM 合成法では、音素間を遷移する際に音声と画像がシンクロナイズされてしまうため、音素境界のアライメントがずれてしまい、これが認識性能に悪影響を及ぼしていた。本章ではこの問題を解決する方法の見当をつけるため、予備実験として新たなモデルの提案を行い、その性能を評価した。

3.1 モデルの提案

3.1.1 Additional 1-state モデル

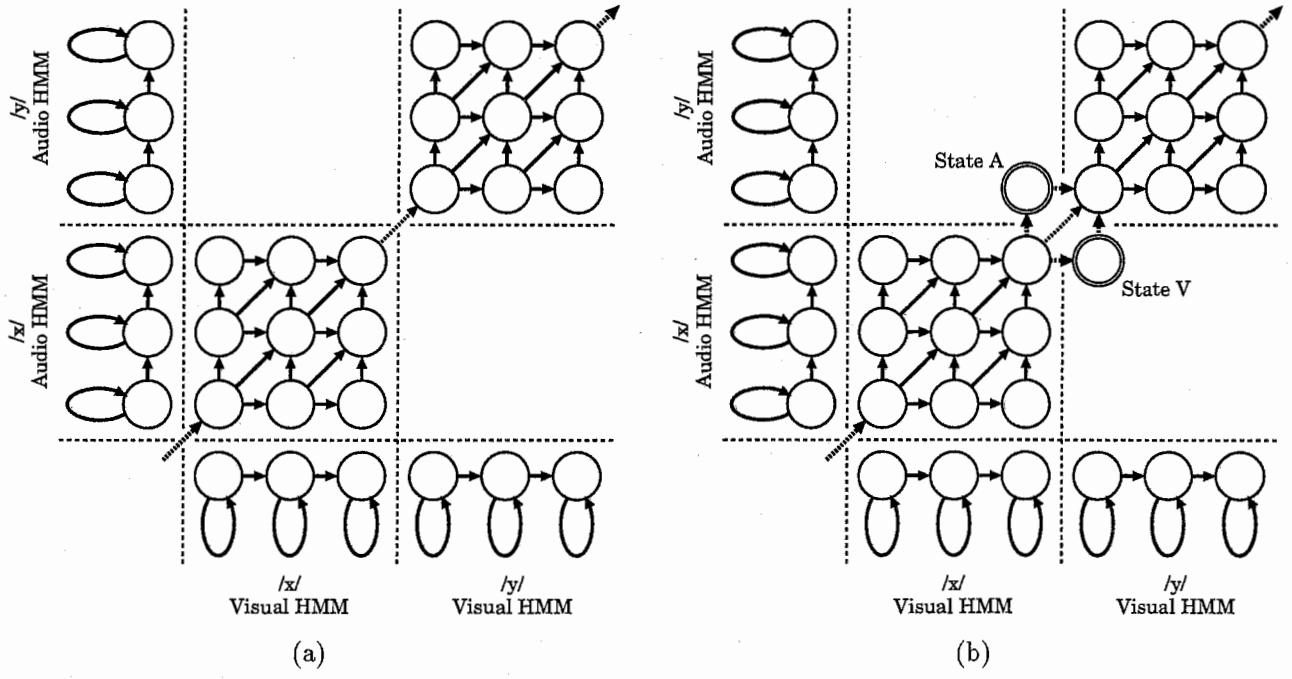


図 3: Additional 1-state モデル

ある 2 つの音素が連続して発声されたとき、前章で説明した HMM 合成法を用いると、図 3 (a) に示すように、各音素に対して音声・画像 HMM が合成され、音素間は先に発声された音素の最終状態から後に発声された音素の初期状態に至るパスによって遷移することになる。ところがこのモデルでは、音素間を結ぶパスがひとつしかないため、本来は音素境界の異なる音声と画像が強制的に同期をとられることになってしまい、ここでミスマッチングが生じ、認識性能を低下させている要因となっている。

そこで、先の音素の音声 HMM の最後の状態と後の音素の画像 HMM の最初の状態を合成した 1 状態の HMM と、先の音素の画像 HMM の最後の状態と後の音素の音声 HMM の最初の状態を合成した 1 状態の HMM を作り、音素間を遷移するときには、従来のような直接遷移に加えてこれらを経由しての遷移も許すモデル（以下、「Additional 1-state モデル」と呼ぶ）を考案した。この Additional 1-state モデルの一例を、図 3 (b) に示す。ここで、新たに付与された状態を◎で示してある（以下、図 3 (b) において左上のものを State A、右下のものを State V と呼ぶ)。

このようなモデルにすることで、ある音素 HMM の最終状態から次の音素の初期状態へのパスが複数存在することになり、音声情報と画像情報の同期のずれによるミスマッチングをある程度抑制することが期待できる。

3.1.2 Word モデル

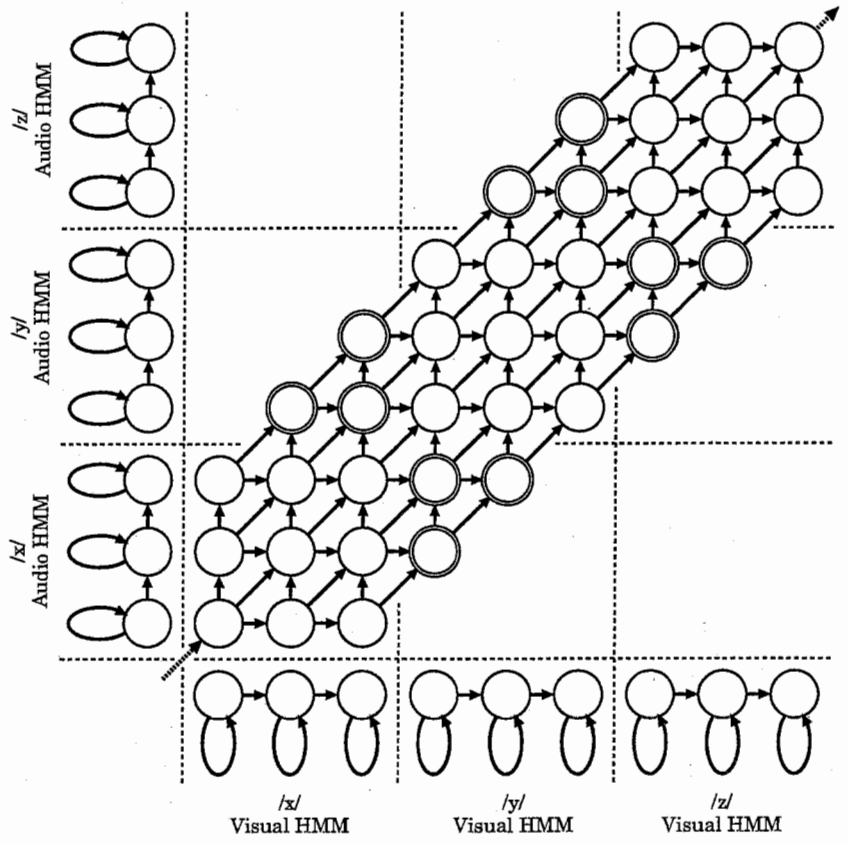


図 4: Word モデル

Additional 1-state モデルにおいても、音素間の遷移に関しては、依然として先の音素の最終状態と後の音素の初期状態を通らなければならないという強い制約が存在する。

この問題を解消するため、各単語に対してその構成音素の音声 HMM をつなげて 1 つにまとめ、同様にして得られる画像 HMM と合成して、ひとつの HMM を割り当てるモデル（以下、Word モデルと呼ぶ）を考案した。ただしこのとき、音素単位での音声（画像） HMM の状態数を n、単語の構成音素数を m とすると、合成した HMM の状態数は $(mn)^2$ となり、m^2 に比例するので m が大きくなると状態数が膨大になってしまう。また、音声情報と画像情報の同期のずれはある一定の範囲内に収まると考えることができ、全ての音声 HMM と画像 HMM の状態の組み合わせについて合成を行うことはあまりにも冗長である。そこで本研究では、音声と画像のずれを 1 音素未満（状態数で表すと $n-1$ 以下）と仮定して Word モデルを作成した。このときの合成 HMM の状態数は、次式で表されるとおり m の 1 次関数として表される。

$$\begin{aligned} State &= m \times n^2 + (m-1) \times (n^2 - n) \\ &= m(2n^2 - n) - (n^2 - n) \end{aligned}$$

図 4 に 3 音素からなる単語の Word モデルの例を示す。

この Word モデルでは、音声または画像の音素境界間を超えるパスが多数存在しており、また、音声・画像の間でも音素境界の同期性が認められる発声開始前および発声終了後にのみ、必然的に同期を要する初期状態・最終状態が割り当てられるので、先述の Additional 1-state モデルよりも音声と画像の同期について考慮したモデルとなっている。

データベース	ATR 発声リスト 5240 単語 音声・画像同期データ 男性話者 1 名
音声	サンプリング周波数 ： 12kHz フレーム長 ： 32ms フレーム周期 ： 8ms 窓関数 ： ハミング窓 パラメータ ： MFCC （ 16 次元） ： MFCCΔ 成分 （ 16 次元）
画像	フレーム周期 ： 33ms パラメータ ： 2 次元 FFT 交流成分係数 （ 35 次元） ： 領域内輝度平均時間差分 （ 24 次元）
音声 HMM、画像 HMM	left-to-right・混合正規型 HMM 状態数 ： 3 混合数 ： 2 HMM 個数 ： 56
音声・画像 HMM	混合正規型 HMM 詳細は別表（表 2 ）参照
学習セット	データベースより抽出した 4740 単語 音声はクリーン
テストセット	データベースより抽出した 200 単語 × 3 セット 学習セットとは排他的に作成 音声はクリーンと 15 dB 白色雑音畳乗の 2 種類

表 1: 実験条件

3.2 実験条件

表 1 に、実験条件の概要を示す。実験では、データベースとして ATR 発声リスト 5240 単語を発声している男性話者 1 名のものを利用した。音響特徴量には、12kHz サンプリングされたクリーンな音声データに対して、フレーム長 32ms、フレーム周期 8ms でハミング窓をかけ、そこから得られる MFCC 係数とその Δ 成分、計 32 次元を用いた。画像特徴量には、30 フレーム／秒のビデオストリームから抽出した 720 × 480 の原画像に対して、重複のないよう 120 × 120 ごとに 6 × 4 に分割した各領域内の輝度平均の時間差分成分 24 次元および、128 × 128 に縮小し 2 次元 FFT を施して得られた 6 × 6 の周波数成分のうち DC 成分を除いたもの 35 次元の、計 59 次元のパラメータを用いた。これらのパラメータを用いて、音声 HMM と画像 HMM をラベル付き学習および連結学習によって作成した。なおこれらの HMM はいずれも、left-to-right の混合正規型 HMM で、その状態数は 3、混合数は 2 である。そして得られた音声 HMM と画像 HMM から、

(1) HMM 合成モデル（従来のもの）

(2) Additional 1-stste モデル

(3) Word モデル

の 3 つのモデルを合成した。(1) については、(2) および (3) との比較対照を行うために用意した。また 3 つのモデルとも、音声ストリームと画像ストリームの重みはいずれも 1.0 : 1.0 と等しく重み付けを行った。これら 3 つの HMM に関する比較を、表 2 に示す。

各モデルの性能評価については、音声 HMM および画像 HMM の学習時において使用していないデータよりなる 200 単語のテストセットを 3 セット作成して認識を行い、次式で計算される単語正解精度の平均値を利用した。

$$Accuracy = \frac{W - E}{W}$$

ここで、W は単語総数 ($= 200$)、E は認識誤り数である。認識時には音響特徴量と画像特徴量を融合した 91 次元のパラメータを用いるが、両者のフレーム周期が一致していないので、画像パラメータは 3 ～ 4 フレーム分同じものを埋め込んで調整した。また各テストセットそれぞれに、次に挙げる 2 つの異なる SNR のものを用意した。

	モデル (1)	モデル (2)	モデル (3)
HMM 個数	56	[1] 6328	[2] 200
HMM 状態数	9	1 or 9	39 ～ 204
認識辞書登録数	200	600	200

[1]Context Independent、[2]Context Dependent

表 2: 3 つのモデルの比較

(a) 音声がクリーンのデータ

(b) 音声に SNR=15 dB となるよう白色雑音を加えたデータ

なおモデル (2) において、全ての音素間について State A と State V の有無の組み合わせを考慮した辞書を作成すると計算に膨大な時間を要するため、各単語に対して、音素間にこれらを全く持たないもの、常に State A を持つもの、常に State V を持つもの、の 3 種類に限定して辞書に登録した。

3.3 実験結果・考察

まず Additional 1-state モデルにおいて、(a) を評価データとしたときに音素列レベルでのデコーダの出力に State A または State V を含むか含まないかによって、その認識結果を分類したものを、表 3 に示す。これから、State A もしくは State V を音素間に含んで認識されたケースの方が認識率が高くなっており、このような音素間遷移を改善する手法が有効であることが示された。なお認識率の改善の程度が小さいのは、State A および State V を用いる場合には辞書において全ての音素間にこれらを含む場合しか許容しておらず、これによる制約が影響しているものと考えられる。

次に Word モデルにおいて、単語内の音素数が 5 のもの、および音素数 7 のもの（いずれも発声開始前／終了後の無発声部分を各々一音素とみなして含む）について、HMM 内の状態遷移の様子を表したものを図 5、図 6 に示す。ここで (a) は評価データ (a) を、(b) は評価データ (b) を用いたときの遷移状況である。この結果を見ると、Word モデルで新たに付与した状態や遷移の使用頻度が高く、強制的な同期を要するパスをほぼ完全に除去したことによる音声・画像の単語境界に関するミスマッチングの抑制の効果が大きいことがわかる。

	正　解	不正解	計	認識率
State A/V を含まない	388	50	438	88.58%
State A/V を含む	145	17	162	89.51%
計	533	67	600	88.83%

表 3: Additional 1-state モデルにおける認識結果

	モデル (1)	モデル (2)	モデル (3)
データ (a)	88.50%	88.83%	92.67%
データ (b)	71.67%	73.00%	76.67%

表 4: 各種音声データに対する各モデルの認識率

最後に、(1) ～ (3) のモデルについて (a)、(b) の 2 種類のデータに対する実験結果（平均認識率）を表 4 に示す。

以上の結果より、従来の方法で生成した (1) よりもクリーンな場合、雑音を加えた場合ともに (2) や (3) の方が認識率が高くなっており、音声と画像の音素境界における同期のずれを、互いに異なる音素の音声 HMM と画像 HMM（図 3 (b)、図 4 の◎の状態）を用いて合成した HMM を組み込むことで吸収できていることがわかる。また、(2) と (3) を比べると (3) の方が高い性能を示している。これは、単に異なる音素から HMM を合成し追加するだけでなく、音声における音素境界、画像における音素境界を越える遷移を許容し、同期を要するパスを除去することの有効性を示している。

3.4 まとめ

本章では、音素境界において音声と画像のアライメントが一致しないという問題を解決するため、音素境界を超えて異なる音素の音声 HMM と画像 HMM から新しい HMM を合成するといったより柔軟なモデルの作成を行い、これを用いることで、従来起こっていた音声と画像のミスマッチングを抑制することに成功した。

しかしながら、ここで提案したモデルには合成後に学習を行うことができないという問題がある。そこで次章では、本章で得られた結果をもとに、学習可能なモデルを構築することについて考えていく。

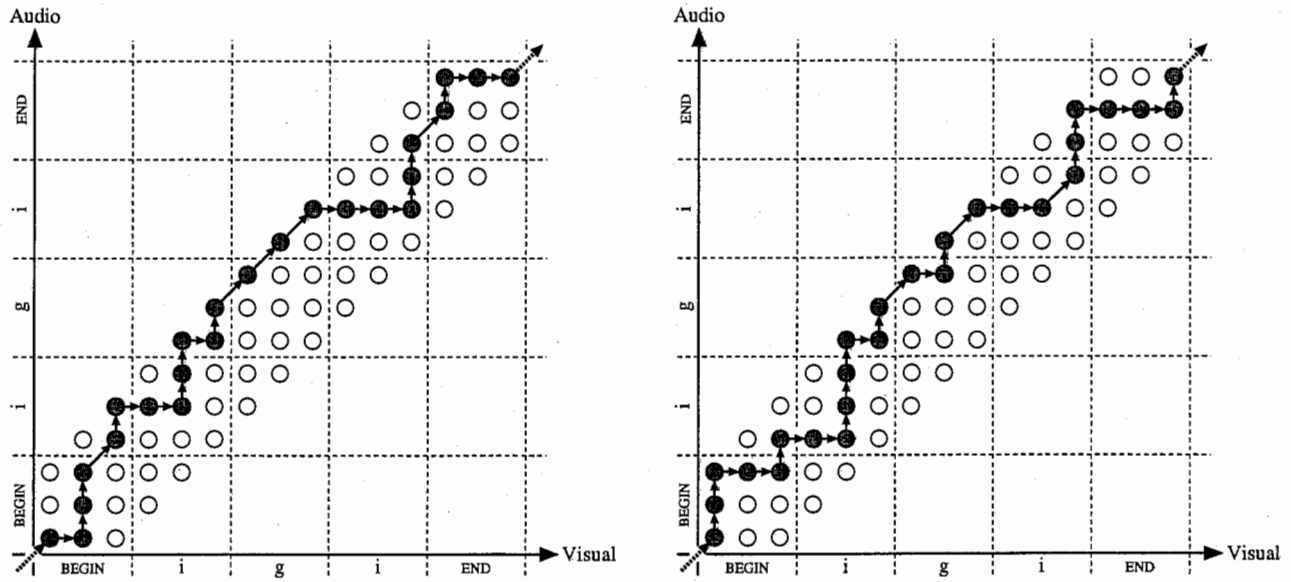
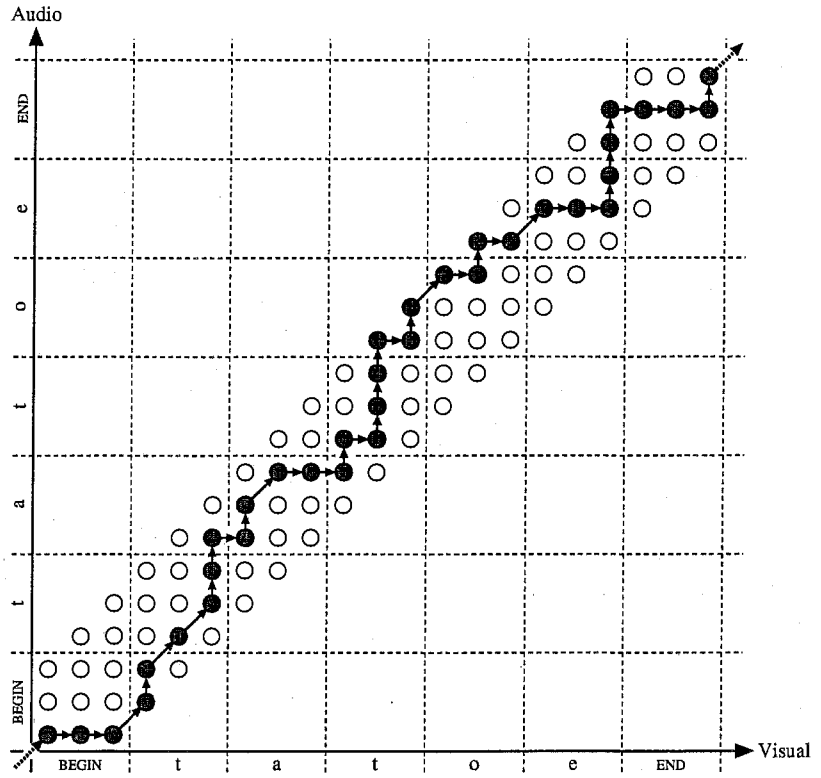
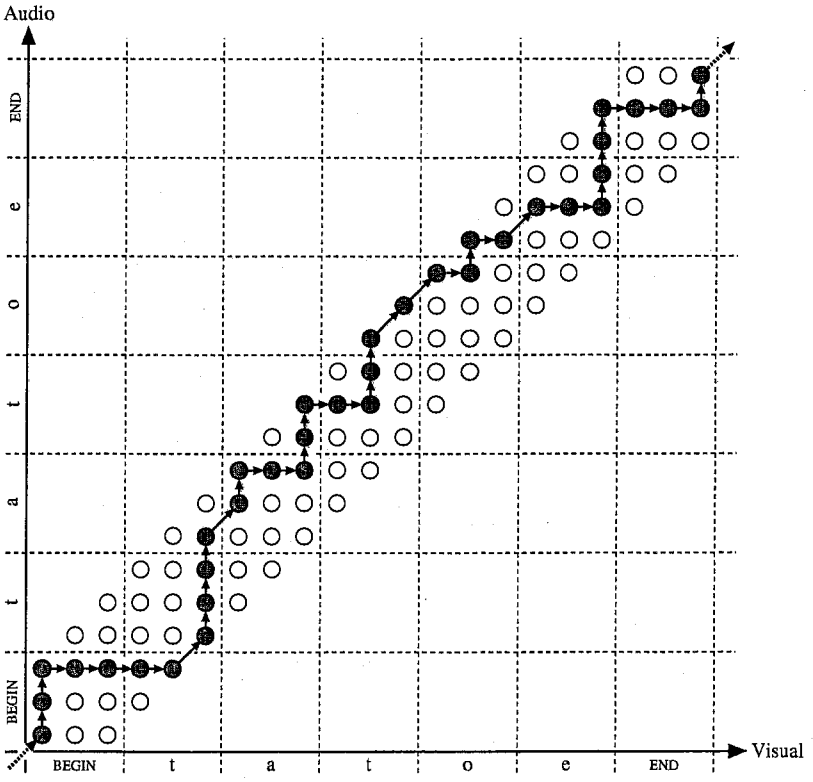


図 5: Word モデルの状態遷移の様子（音素数 5、単語「異議 (BEGIN-i-g-i-END)」）



(a)



(b)

図 6: Word モデルの状態遷移の様子（音素数 7、単語「例え (BEGIN-t-a-t-o-e-END)」）

4 実験 — 学習可能な非同期性モデルの提案

予備実験により音声・画像の音素境界における同期の問題を、異なる音素の音声 HMM と画像 HMM から合成したモデルを用いることで解消できることがわかった。本章ではこの点を踏まえ、学習可能な新たなモデルを提案して音声・画像データにより再学習を行い、認識実験によってその結果を考察した。

4.1 モデルの提案

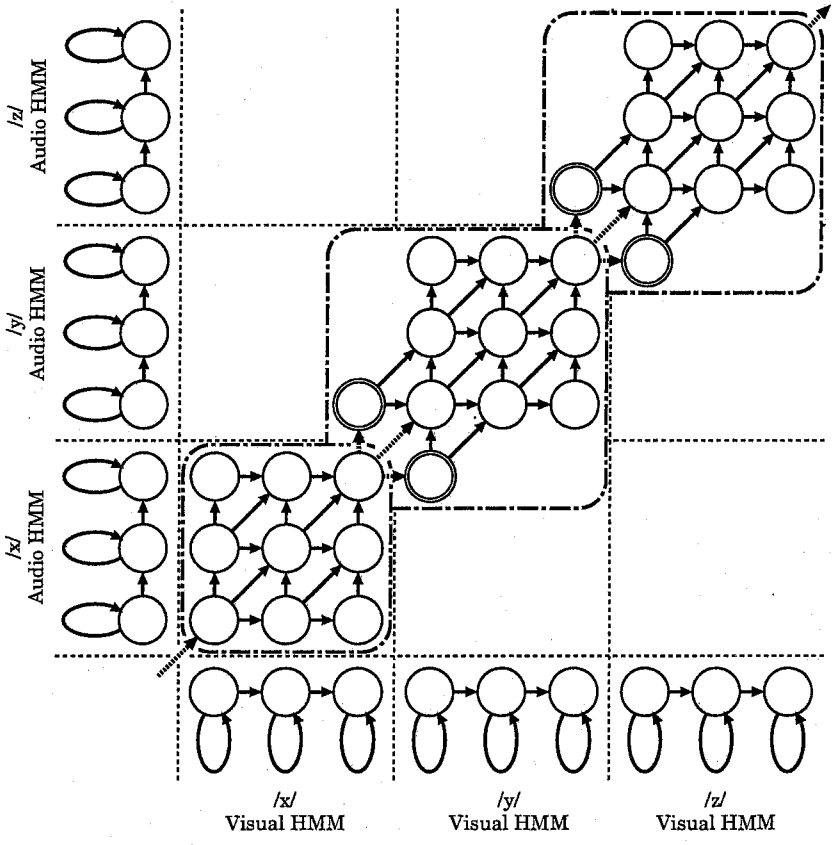


図 7: Pseudo-biphone モデル

予備実験で提案したモデルにおいては、Additional 1-state モデルでは音素間の State A と State V の有無の組み合わせのために、一単語に対して辞書に登録する読みが構成音素数によって指数的に増加すること、Word モデルでは単語発話単位でのデータがある程度ないと学習できないだけでなく、そもそも状態数が非常に多いので HMM 全体としての学習による改善が難しいこと、などの問題があり、この 2 つのモデルでは、合成した HMM を認識に用いることはできても、さらに音声・画像データを用いて学習することができなかった。しかしながら、合成した HMM の音声・画像データによる再学習ができれば、音声の音素境界と画像の音素境界のアライメントの改善（両者のずれの補正）を行うことができ、このような個々のモーダルでの学習では得られない HMM の改良により認識性能が向上すると考えられる。ゆえに、Additional 1-state モデルや Word モデルのような音声と画像の音素境界による非同期性を吸収し、かつ音声・画像データによって学習可能なモデルを構築することが重要になってくる。

そこで本研究では biphone モデルの考え方を参考に、合成した音素の HMM に対して Additional 1-state モデルにおける State A と State V の 2 つの状態を組み込み、これらと中心となる音素 HMM（以下、核と呼ぶ）との遷移を適切に付与して、新しいひとつの HMM とするモデル（以下、Pseudo-biphone モデルと呼ぶ）を考案した。このとき、State A および State V から核のそれぞれの状態への遷移確率は HMM 合成法と同様にして計算し、また直前の HMM からの遷移に際しては、核の初期状態、State A、State V への遷移に等しく確率を振り分

データベース	ATR 発声リスト 5240 単語 音声・画像同期データ 男性話者 1 名
音声	サンプリング周波数 ： 12kHz フレーム長 ： 32ms フレーム周期 ： 8ms 窓関数 ： ハミング窓 パラメータ ： MFCC （ 16 次元） ： MFCCΔ 成分 （ 16 次元）
画像	フレーム周期 ： 33ms パラメータ ： 2 次元 FFT 交流成分係数 （ 35 次元） ： 領域内輝度平均時間差分 （ 24 次元）
音声 HMM、画像 HMM	left-to-right・混合正規型 HMM 状態数 ： 3 混合数 ： 2 HMM 個数 ： 56
音声・画像 HMM	混合正規型 HMM 状態数 ： 11 、混合数 ： 2 音声重み：画像重み = 1.0 : 0.0 ～ 0.0 : 1.0 （0.1 刻み） 学習方法についての詳細は別表（表 6 ）参照
学習セット	データベースより抽出した 4740 単語 音声は 15 dB 白色雑音重乗
テストセット	データベースより抽出した 200 単語 × 3 セット 学習セットとは排他的に作成 音声は 15 dB、0 dB、−5 dB 白色雑音重乗の 3 種類

表 5: 実験条件

けた。この Pseudo-biphone モデルの例を、図 7 に示す。図では 3 音素の例示となっているが、これからもわかるように、このモデルは必然的に context-dependent になり、単語内の最初の音素については HMM 合成して得られる monophone レベルのモデルをそのまま用いている。なお少ないデータで有効に学習するため、核の部分は音素ごとに全ての HMM で状態を共有している。

以上、提案したこの Pseudo-biphone モデルは Additional 1-state モデルの発展形と考えることもできるが、State A、State V まわりの遷移が Additional 1-state モデルよりも多様化していること、加えて biphone モデルのように再学習可能なモデルとなっていることが特徴であり、より音声と画像の音素境界の非同期性を吸収して認識性能が向上することが期待できる。

4.2 実験条件

表 5 に、実験条件の概要を示す。データベース、音響特徴量、画像特徴量、および音声 HMM と画像 HMM の生成までの過程については予備実験と同様であるが、音声 HMM の作成にあたっては、15 dB の白色雑音が加算的に付与されているデータを用いた。音声・画像 HMM については、音声 HMM、画像 HMM から合成した後、学習方法やそのタイミングを変えて次の 5 つのモデルを用意した。

(1) そのまま Pseudo-biphone 化

(2) 連結学習し、Pseudo-biphone 化

(3) そのまま Pseudo-biphone 化し、その後連結学習

(4) 連結学習し、Pseudo-biphone 化したあと再度連結学習

(5) 音声のラベルによってラベル付き学習し、Pseudo-biphone 化

図 8 にこの処理の流れ図を、表 6 にこれらのモデルの学習方法の違いをまとめておいたので参照されたい。なおいずれの学習においても、データとしては音声 HMM と画像 HMM の学習時に使ったもの（ただし融合した 91 次元のもの）を用いている。

	モデル (1)	モデル (2)	モデル (3)	モデル (4)	モデル (5)
HMM 合成後	学習なし	連結学習	学習なし	連結学習	ラベル学習
Pseudo-biphone 変換後	学習なし	学習なし	連結学習	連結学習	学習なし

表 6: 5 つのモデルの学習方法における比較

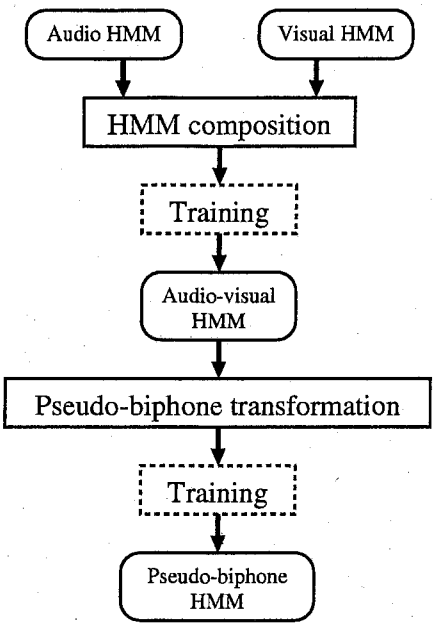


図 8: Pseudo-biphone モデル生成までの流れ図

学習時には音声ストリーム（ λ_a ）と画像ストリームの重み（ λ_v ）は 1.0 : 1.0 であるが、認識の際には次の制約の下でこれらを 0.0 から 1.0 まで 0.1 刻みで変化させ、11 通りのパターンをテストした。

$$\lambda_a + \lambda_v = 1$$

テストセットについては予備実験と同様に作成し、それぞれに対して音声データに次に挙げるレベルの白色雑音を加えたものを用意した。

(a) SNR=15 dB

(b) SNR=0 dB

(c) SNR=−5 dB

学習時には SNR=15 dB のデータを用いているので、実験に用いた HMM は (a) に対する雑音適応モデルとなることに注意されたい。

4.3 実験結果・考察

作成した (1) ～ (5) の各モデルにおいてそれぞれ、SNR=15 dB、0 dB、−5 dB のデータに対するストリーム重みを変化させた実験の結果のグラフを図 9 に示す。またこの結果を各 SNR 毎にまとめ 5 つのモデルの比較を行ったグラフを図 10 に示す。

これらの結果について検証してみると、まず (1) のモデル（学習なし）について、図 9 から SNR=15 dB における $\lambda_a = 0.5$ の場合の認識率は 92% とわかり、予備実験とは条件や計算式が違ってくるので単純な比較はできないものの、音声雑音適応モデルにおける音声重みと画像重みが等しい場合での認識率は Word モデルに近い値

を示しており、このモデルにおいても、音声と画像のミスマッチングを抑止し認識性能を向上させることが可能なことが推測できる。一方で図 10 から、音声情報だけでも十分な性能が得られる雑音適応時（ SNR=15 dB ）の音声ストリーム重みが 1 に近いところを除き、それ以外の場合ではモデル (1) は他の何らかの学習を行ったモデルよりも認識率が下回っている。これはモデル合成後・Pseudo-biphone 化後の学習が有効であることを示している。

次に (2) のモデル（合成後に連結学習）と (3) のモデル（変換後に連結学習）を比べてみると、いずれの SNR においても $\lambda_a \geq 0.8$ ではモデル (2) の方が、$\lambda_a \leq 0.7$ ではモデル (3) の方が認識率が高くなっている。モデル (2) は合成した音素 HMM の段階で学習しているので、音声・画像のミスマッチングの影響が少ない音声ストリームの大きなところではモデル (3) を上回り、反面 Pseudo-biphone の変換を行った後に学習したモデル (3) の方が、音声と画像の非同期性がより考慮されたモデルとなっているため、画像ストリーム重みが大きくなるにつれてモデル (2) よりも高い性能を示したものと考えられる。

そして (4) のモデル（合成後と変換後に連結学習）では、音声ストリーム重みが大きいところではモデル (2) とほぼ同等かそれ以上の性能を示し、画像ストリーム重みが大きいところになるとモデル (3) にかなり近い認識率を示すようになり、学習方法から予想できるようにちょうどモデル (2) とモデル (3) 双方の長所を持ち合わせたモデルとなっている。ただし雑音状況に応じて最適なストリーム重みを推定できると仮定すれば（すなわち認識率の最高値で比較すれば）、図 10 からモデル (3) とモデル (4) のモデルはほとんど同じ認識性能であることがわかる。モデル (3) は HMM 合成直後には学習を行っていないにもかかわらずモデル (4) と同程度の結果が得られたことは、Pseudo-biphone にした後の学習であっても核の部分が状態共有しているために学習できたからと結論づけられる。

(5) のモデル（合成後にラベル学習）は、図 10 より雑音適応の SNR=15 dB においては λ_a が 0.4 くらいまで、それ以外の SNR では $\lambda_a = 1.0$ のときにのみ他の学習ありモデルと比較的遜色のない性能を示したものの、全体的には学習なしモデルからわずかな認識率の向上にとどまった。この原因は使用したラベルが音声情報に基づくものであったために、音声と音素境界が異なる画像情報のアライメントが適正に行われなかったからである。逆に音声情報にとって有利な雑音適応時および画像情報を使用していないような上記の状況では、ラベル付きで学習された音声ストリームのために認識率が改善したものと考えられる。

以上の考察から、今回作成したこの 5 つのモデルの中ではモデル (3) およびモデル (4) が最も高い認識性能を有するものと考えられる。両者ともに Pseudo-biphone に変換しその後に連結学習を行っていることから、今回提案したこの Pseudo-biphone 化の手法が従来の HMM 合成法のみによる手法よりも認識性能の向上に有効であることがわかる。

最後にモデル (1)、モデル (3) およびモデル (4) と、従来の HMM 合成法で生成したモデル（学習なし、連結学習のみ=学習あり [A]、ラベル付き+連結学習を行ったもの=学習あり [B]）との比較を図 11 に示す。

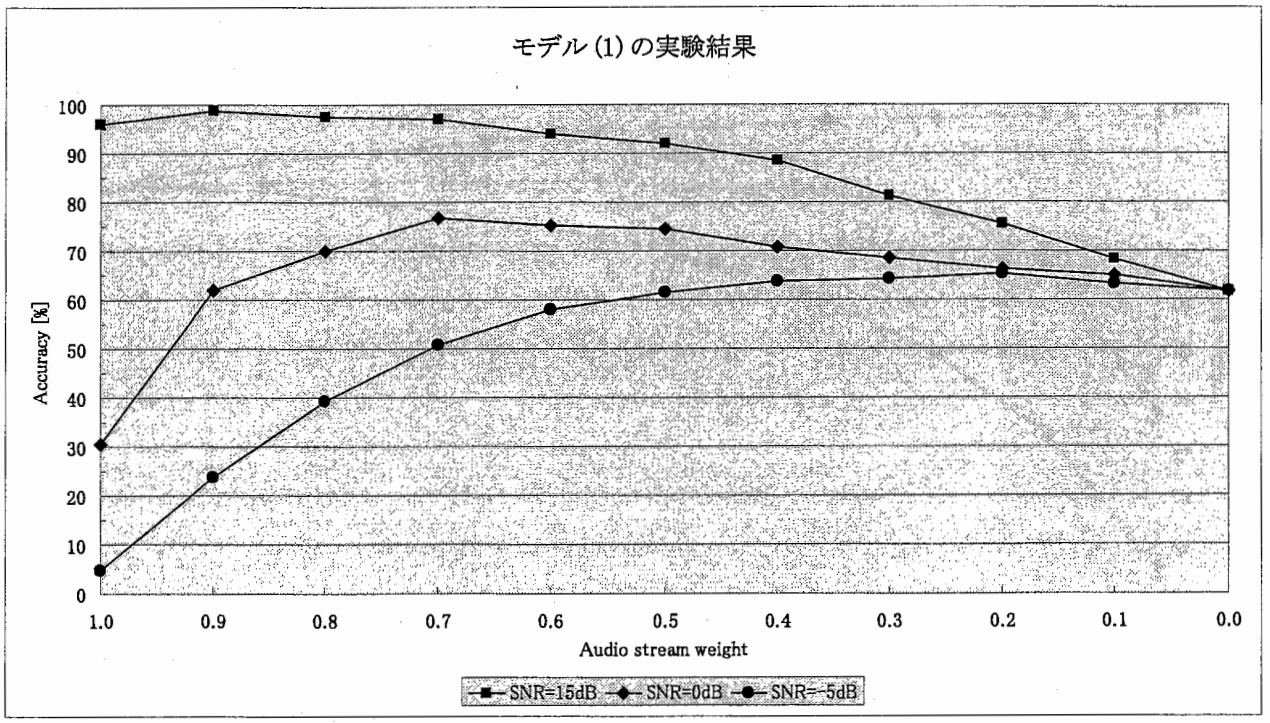
モデル (1) と従来モデル（学習なし）ではほぼ全ての場合においてモデル (1) の方が高い性能を示しており、同じストリーム重みで比較した場合、最大で 10% を超える認識率の改善がみられていることから、Pseudo-biphone 変換によるモデルの改善が効果があることが改めて示された。その一方で学習あり [A][B] の認識性能までは達していないことから、合成後もしくは変換後に再学習することの必要性・重要性が判明した。

またモデル (3) とモデル (4) においても、全体的には従来モデル（学習あり）と同程度かそれ以上の認識率となっており、最適なストリーム重みのところで見てみると、SNR=0 dB ではおよそ 3% の向上、同様に SNR=−5 dB では約 4.5 ～ 5.8% の向上が認められた。ところで $\lambda_a = 0.0$ （ $\lambda_v = 1.0$ ）のところでみると、モデル (3) では従来の学習ありモデルと比べて 7% 近く改善しているのがわかる（モデル (4) のモデルでも約 6% の改善）。これはモデルを改良したことによって、連結学習で音声と画像の強制同期が弱くなったために、よりアライメントが適正化されたためと考えられる。結果として、音声情報を学習時にのみ加えるだけで画像のみによる認識結果が向上したことになるが、逆に SNR=15 dB で音声ストリーム重みが 1.0 のところを見るとモデル (3)、モデル (4) のみ 2 ～ 3 % 他よりも認識率が高くなっており、こちらは画像情報が音声のみの認識性能を上げたものと思われる。

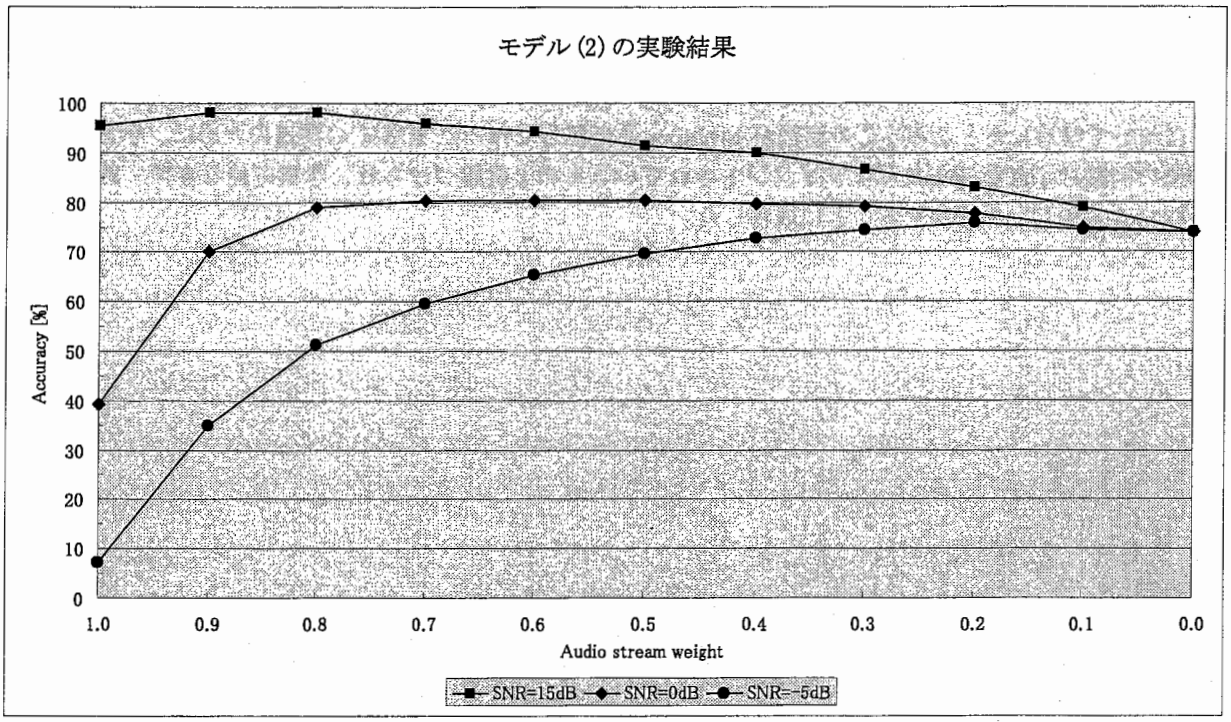
4.4 まとめ

本章では、前章で問題となっていた再学習について biphone モデルの考え方を参考した新しいモデルを作り、実験の結果、このモデルが従来の HMM 合成法で作られるモデルに比べて十分に高い認識性能を示すことを確認した。

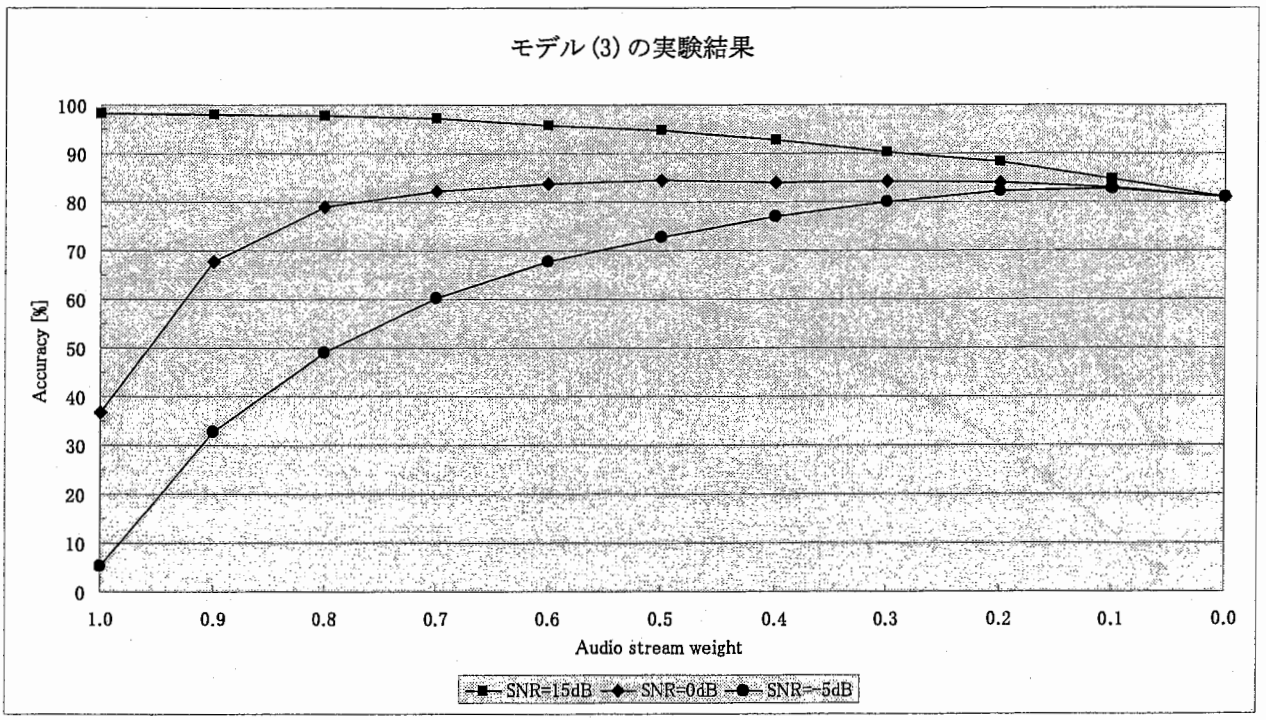
これについて検討したところ、この要因が biphone を参照して作った HMM の構造のために、学習によって音声と画像の同期のずれを適正にリアライメントされているものと結論づけられ、学習可能な音声・画像の非同期性を吸収できるモデルを構築することに成功した。



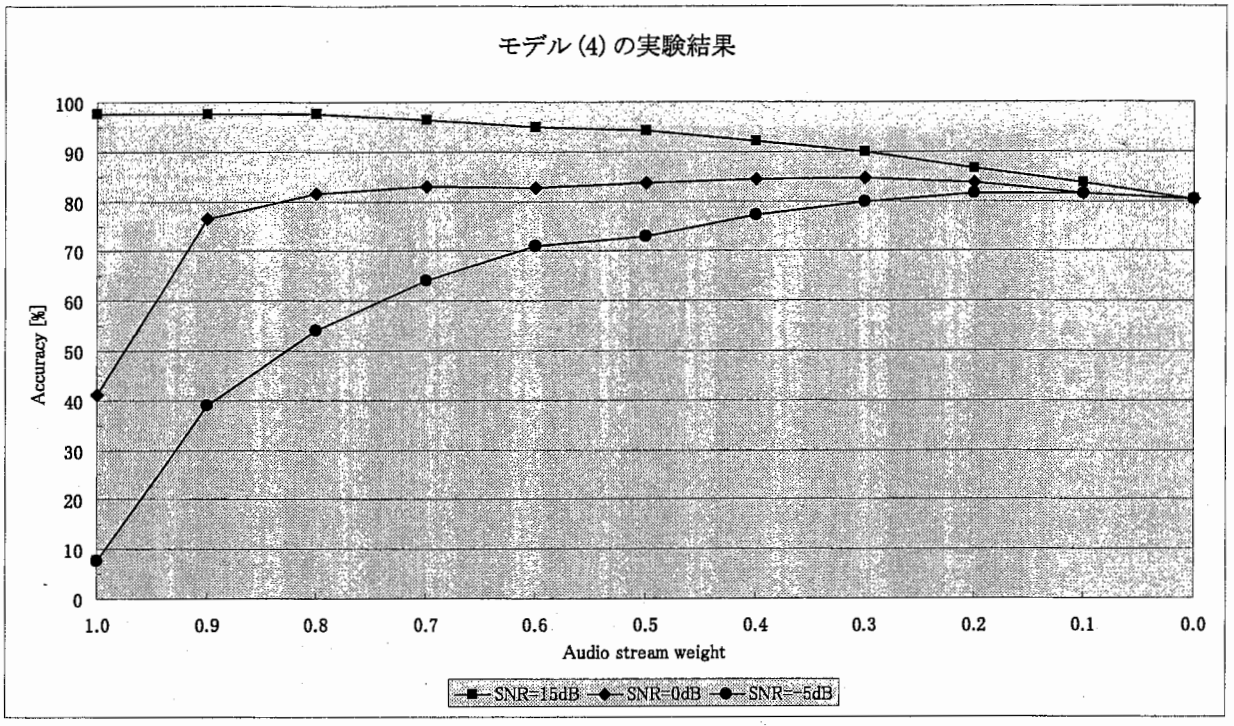
	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
SNR=15dB	96.00	98.67	97.50	97.00	94.00	92.00	88.50	81.17	75.50	68.17	61.67
SNR= 0dB	30.50	62.00	70.00	76.67	75.17	74.50	70.83	68.50	66.17	65.00	61.67
SNR=-5dB	4.83	23.83	39.33	50.67	58.00	61.50	63.67	64.17	65.17	63.33	61.67



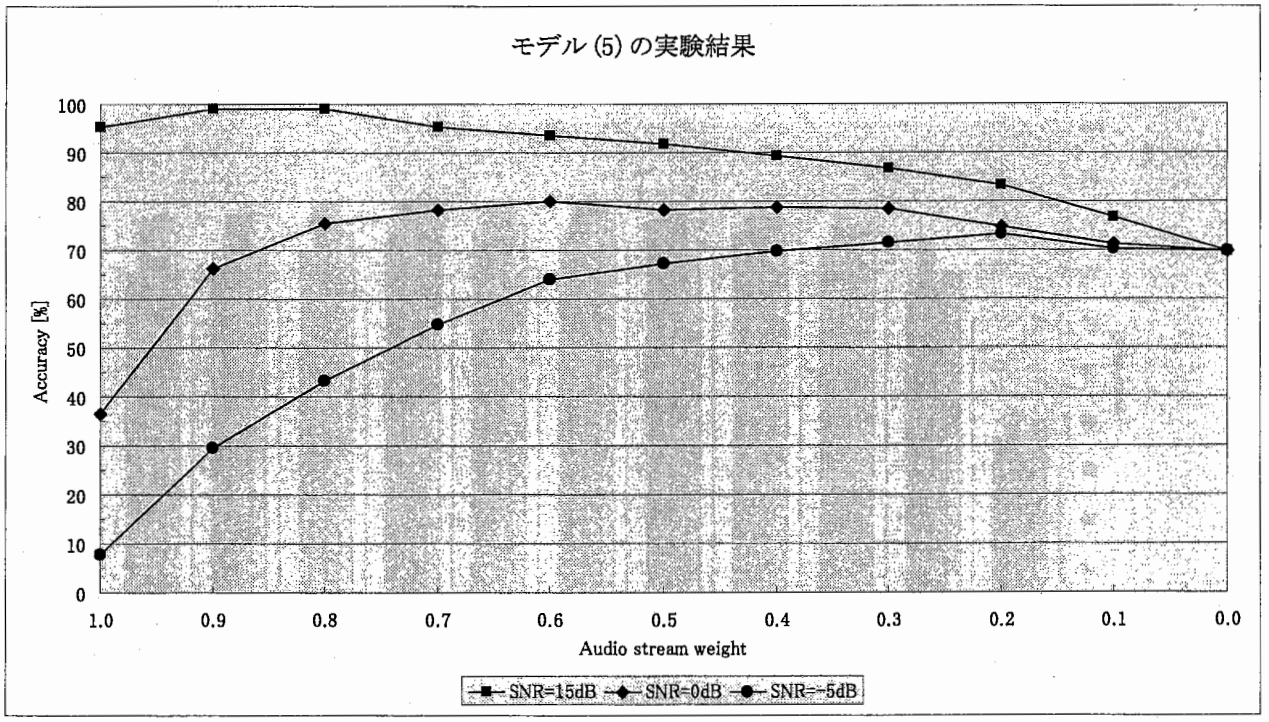
	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
SNR=15dB	95.50	98.00	98.00	96.00	94.33	91.50	90.00	86.50	83.00	78.83	73.83
SNR= 0dB	39.17	70.00	79.00	80.33	80.50	80.50	79.67	79.17	77.83	75.00	73.83
SNR=-5dB	7.17	35.00	51.17	59.50	65.33	69.67	72.83	74.50	75.83	74.33	73.83



	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
SNR=15dB	98.17	98.00	97.83	97.33	95.67	94.67	92.67	90.17	88.17	84.67	81.00
SNR= 0dB	36.83	67.83	79.00	82.17	83.83	84.50	84.00	84.17	84.00	83.00	81.00
SNR=-5dB	5.33	32.83	49.00	60.33	67.67	72.67	77.00	80.00	82.17	82.67	81.00

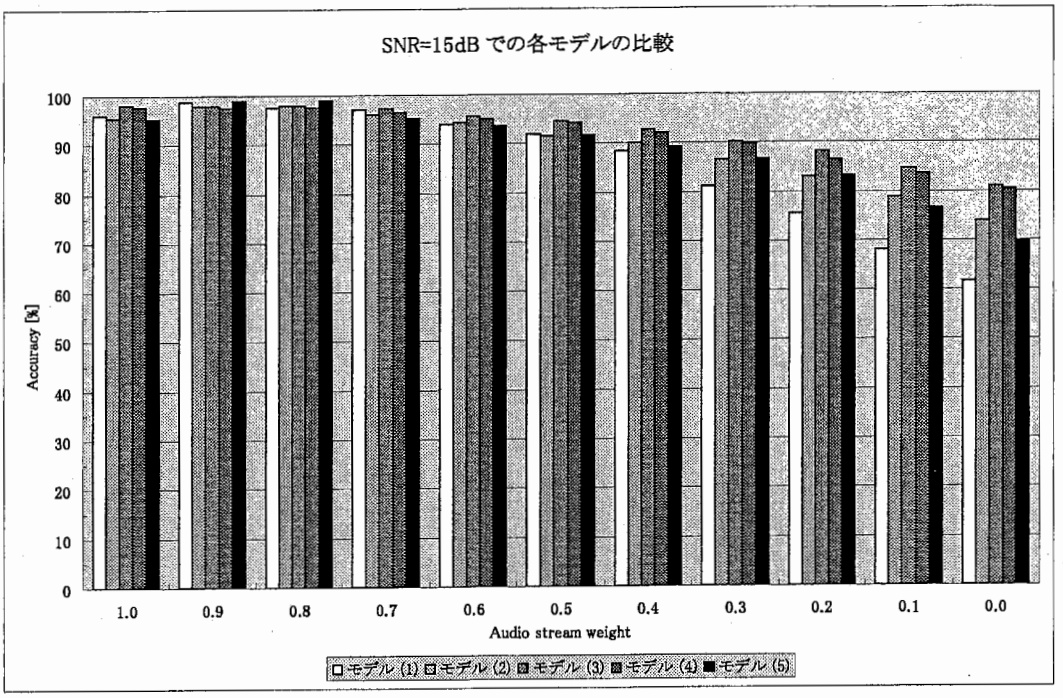


	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
SNR=15dB	97.67	97.50	97.50	96.50	95.00	94.17	92.17	89.83	86.50	83.67	80.33
SNR= 0dB	41.17	76.50	81.67	83.00	82.83	83.83	84.33	84.67	83.67	81.67	80.33
SNR=-5dB	7.67	39.00	54.00	64.00	71.00	73.00	77.33	80.00	81.50	81.67	80.33

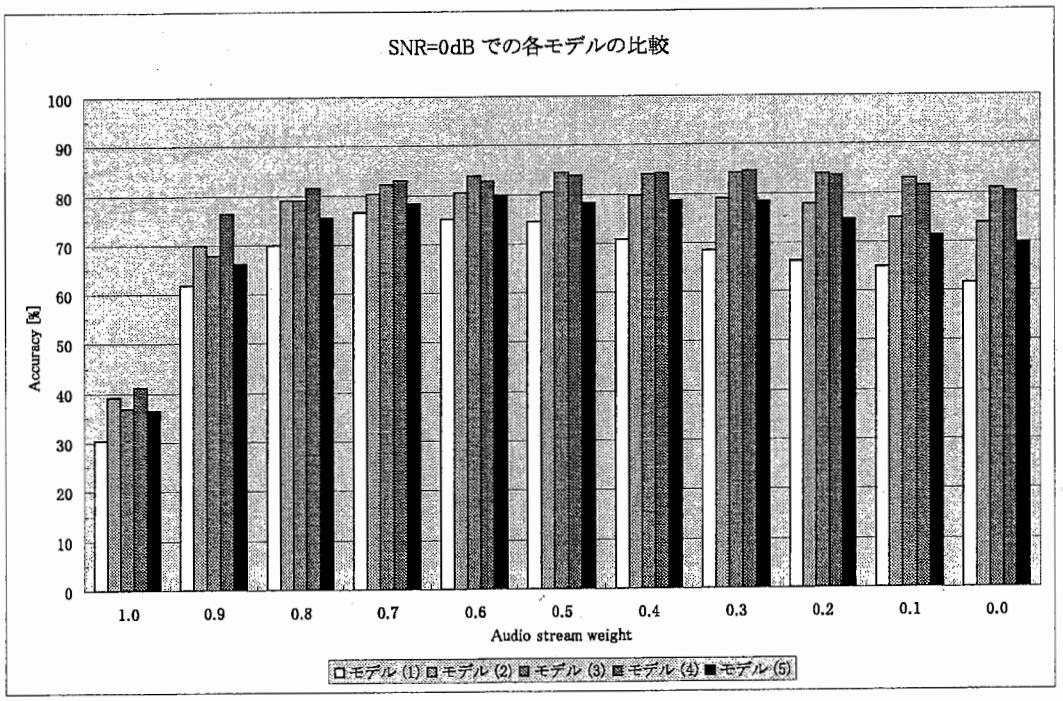


	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
SNR=15dB	95.33	99.00	99.00	95.33	93.50	91.67	89.33	86.83	83.33	76.67	69.83
SNR= 0dB	36.50	66.33	75.50	78.33	80.00	78.17	78.67	78.50	74.83	71.33	69.83
SNR=-5dB	7.67	29.50	43.33	54.67	64.00	67.33	69.67	71.50	73.33	70.17	69.83

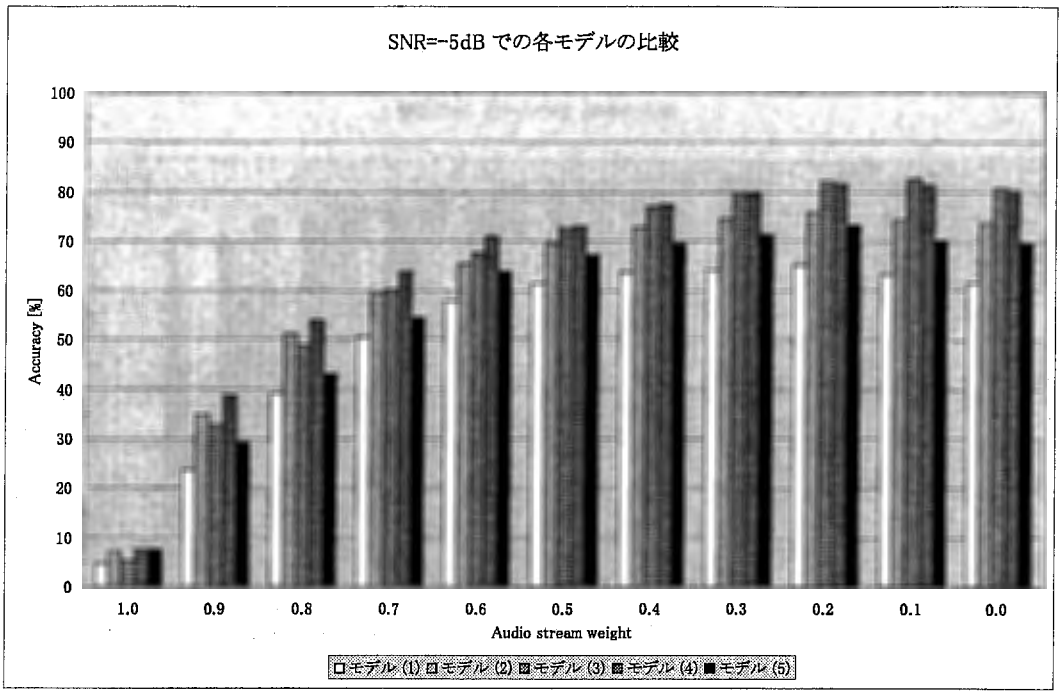
図 9: 各モデルにおける SNR・ストリーム重み別の認識結果



	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
モデル (1)	96.00	98.67	97.50	97.00	94.00	92.00	88.50	81.17	75.50	68.17	61.67
モデル (2)	95.50	98.00	98.00	96.00	94.33	91.50	90.00	86.50	83.00	78.83	73.83
モデル (3)	98.17	98.00	97.83	97.33	95.67	94.67	92.67	90.17	88.17	84.67	81.00
モデル (4)	97.67	97.50	97.50	96.50	95.00	94.17	92.17	89.83	86.50	83.67	80.33
モデル (5)	95.33	99.00	99.00	95.33	93.50	91.67	89.33	86.83	83.33	76.67	69.83

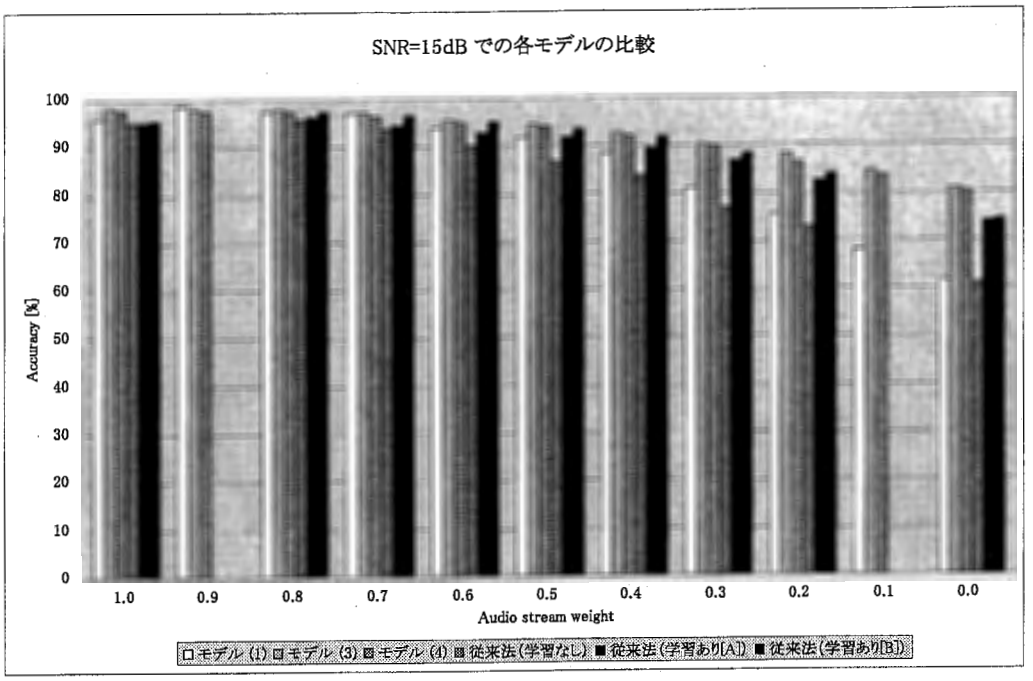


	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
モデル (1)	30.50	62.00	70.00	76.67	75.17	74.50	70.83	68.50	66.17	65.00	61.67
モデル (2)	39.17	70.00	79.00	80.33	80.50	80.50	79.67	79.17	77.83	75.00	73.83
モデル (3)	36.83	67.83	79.00	82.17	83.83	84.50	84.00	84.17	84.00	83.00	81.00
モデル (4)	41.17	76.50	81.67	83.00	82.83	83.83	84.33	84.67	83.67	81.67	80.33
モデル (5)	36.50	66.33	75.50	78.33	80.00	78.17	78.67	78.50	74.83	71.33	69.83

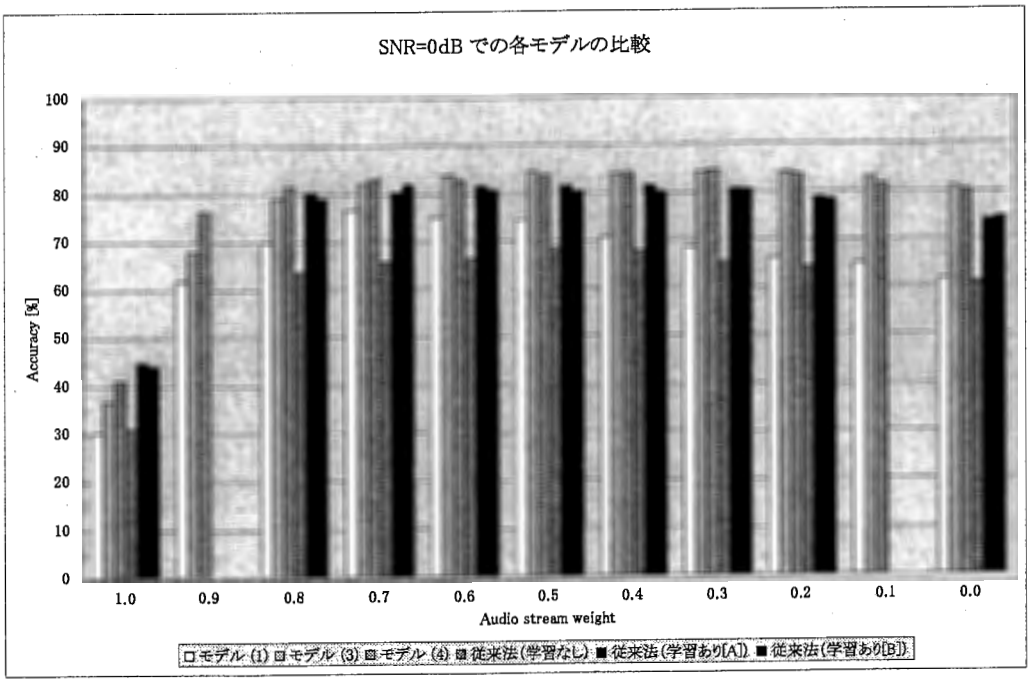


	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
モデル (1)	4.83	23.83	39.33	50.67	58.00	61.50	63.67	64.17	65.17	63.33	61.67
モデル (2)	7.17	35.00	51.17	59.50	65.33	69.67	72.83	74.50	75.83	74.33	73.83
モデル (3)	5.33	32.83	49.00	60.33	67.67	72.67	77.00	80.00	82.17	82.67	81.00
モデル (4)	7.67	39.00	54.00	64.00	71.00	73.00	77.33	80.00	81.50	81.67	80.33
モデル (5)	7.67	29.50	43.33	54.67	64.00	67.33	69.67	71.50	73.33	70.17	69.83

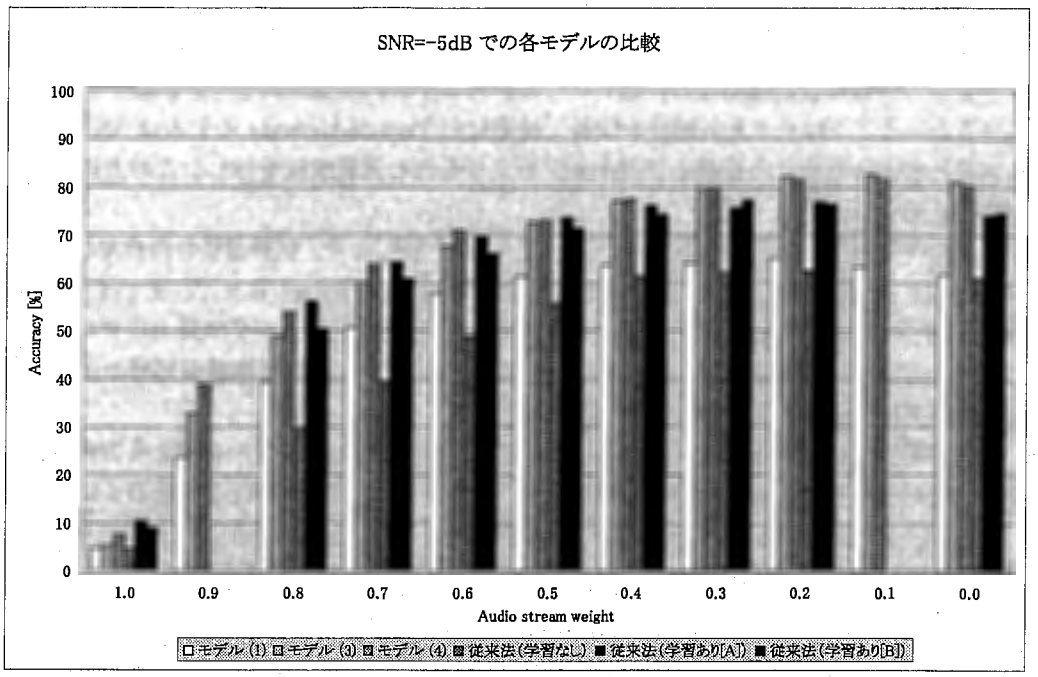
図 10: SNR 毎の各モデルのストリーム重み別の認識結果



	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
モデル (1)	96.00	98.67	97.50	97.00	94.00	92.00	88.50	81.17	75.50	68.17	61.67
モデル (3)	98.17	98.00	97.83	97.33	95.67	94.67	92.67	90.17	88.17	84.67	81.00
モデル (4)	97.67	97.50	97.50	96.50	95.00	94.17	92.17	89.83	86.50	83.67	80.33
学習なし	95.33		95.83	93.83	90.17	87.00	83.83	77.00	73.00		61.00
学習あり[A]	95.17		96.50	94.67	93.00	91.83	89.67	87.00	82.67		74.17
学習あり[B]	95.67		97.50	96.67	95.33	93.50	92.00	88.33	84.00		74.50



	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
モデル (1)	30.50	62.00	70.00	76.67	75.17	74.50	70.83	68.50	66.17	65.00	61.67
モデル (3)	36.83	67.83	79.00	82.17	83.83	84.50	84.00	84.17	84.00	83.00	81.00
モデル (4)	41.17	76.50	81.67	83.00	82.83	83.83	84.33	84.67	83.67	81.67	80.33
学習なし	31.33		63.67	65.83	66.50	68.67	68.00	65.33	64.33		61.00
学習あり[A]	45.00		80.33	80.33	81.67	81.50	81.67	80.67	78.83		74.17
学習あり[B]	44.33		79.17	82.00	80.67	80.33	80.17	80.50	78.50		74.50



	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
モデル (1)	4.83	23.83	39.33	50.67	58.00	61.50	63.67	64.17	65.17	63.33	61.67
モデル (3)	5.33	32.83	49.00	60.33	67.67	72.67	77.00	80.00	82.17	82.67	81.00
モデル (4)	7.67	39.00	54.00	64.00	71.00	73.00	77.33	80.00	81.50	81.67	80.33
学習なし	4.83		30.00	39.50	49.00	55.83	61.50	62.33	62.50		61.00
学習あり[A]	10.50		56.33	64.33	69.83	73.67	76.17	75.83	76.83		74.17
学習あり[B]	9.17		50.50	61.17	66.17	71.50	74.33	77.17	76.50		74.50

図 11: SNR 毎の従来／提案モデルのストリーム重み別の認識結果

5 おわりに — 今後の課題

本研究では、音声と画像の音素境界におけるアライメントのずれの問題に対して、HMM 合成により得られるモデルを改良するというアプローチから解決を試み、提案した Pseudo-biphone モデルは SNR が小さなところ、画像ストリーム重みが大きなところを中心に従来モデルよりも高い認識性能を示し、この方法および提案したモデルの有効性が確かめられた。

にもかかわらず、今回提案したモデルにおいてもなお、同期を要したり、状態遷移の自由度が音素境界のために小さくなっていたりする部分があり、完全に音声と画像の同期の問題を解決するところまでは至っていない。しかし現状の HMM のトポロジーの範囲内、すなわち初期状態と最終状態をひとつずつ持つモデルという制約の下においては、これ以上のモデルの改良は難しく、また仮にできるとしても非常に複雑な処理や構造を要するものと見込まれ、もはや限界にきていると言わざるを得ない。

以上から、もし本研究以上にこの問題に取り組むとすれば HMM のトポロジーを変える方向で考えていくべきであると思う。例えば初期状態と最終状態を複数許すようなモデルを構築すれば、学習および認識において各音素の HMM を連結して評価する際に、より柔軟な組み合わせを調べることができるようになり、本研究で述べた Word モデルや Pseudo-biphone モデルのような構造を音素 HMM だけで表現可能となる。このように HMM の組み合わせの自由度が増すことで、音声と画像のミスマッチングを解決し、認識率を向上させることが期待できる。ただ HMM のトポロジーを変えるためには、特徴量の生起確率の計算式や学習時の HMM 更新の方法など HMM に関する数学的理論を再度検討する必要があること、その結果として既存のアプリケーションが使えなくなればそれらを自作しなければならないこと、などの課題があり、これらは容易にできるものではない。それゆえ予め予備実験を行うなどして、認識精度が向上する新たなトポロジーを慎重に見極めるとともに、それを実現するための理論およびプログラムを着実に構築していくことが重要となる。

謝辞

本研究にあたって、機会を与えてくださった ATR 音声言語通信研究所・山本誠一社長に深く感謝いたします。また研究を進める上でお世話になりました第一研究室の皆様、および TSG の方々に心から感謝いたします。

参考文献

[1] 熊谷健一、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" 音講論集、2-Q-11、2000 年 9 月