手話認識のための状態遷移構造推定

松尾　直志† 　白井　良明† 　島田　伸敬†

† 立命館大学情報理工学部 〒 525–8577 滋賀県草津市野路東 1–1–1
E-mail: †matsuo@i.ci.ritsumei.ac.jp, ††{shirai,shimada}@ci.ritsumei.ac.jp

あらまし　本研究では，発話ごとに異なる手話単語動作の多様性に応じて適切な状態遷移構造を持つ HMM を構成する手法について述べる．HMM を用いた手話認識手法は既にいくつか提案されているが，従来法ではモデルの状態遷移構造は単語に関係なく固定されていた．しかし同じ手話単語でも動作は一部が省略されていたり変形している場合がある．各単語につき複数の HMM を考えて動作の省略・変形に対応することもできるがその場合共通する動きも複数の HMM に分散してしまうので学習の効率が悪い．そこで本研究では手話動作の変化に応じた枝分かれを持つ単一の HMM を自動的に構成する方法を提案する．提案法は各発話での動作をまとまった動きの系列に分解し，複数の発話について動き系列を統合することでその単語の状態遷移構造を推定する．実験の結果，手話動作のパターンに応じた状態遷移構造が得られることが確認できた．

キーワード　手話認識，動画像処理，HMM，動き抽出，状態遷移推定

Estimation of Transition Topology for Sign Language Recognition

Tadashi MATSUO†, Yoshiaki SHIRAI†, and Nobutaka SHIMADA†

† College of Informatio Science and Engineering, Ritsumeikan University　Noji-higashi 1–1–1, Kusatsu-shi, Shiga, 525–8577 Japan
E-mail: †matsuo@i.ci.ritsumei.ac.jp, ††{shirai,shimada}@ci.ritsumei.ac.jp

Abstract　We propose a method to automatically construct a transitional structure(topology) of a Hidden Markov Model for recognizing a word in sign language from a sequence of images. The constructed topology has branches and junctions in order to represent a flexible structure. The proposed method consists of segmentation of a motion and construction of the topology from segments. A motion is divided into consistent segments, which correspond to states of the model. The topology is constructed from an initial topology by modifying it according to a learning sequence of the segments. With experiments, we show the effectiveness of the proposed method.

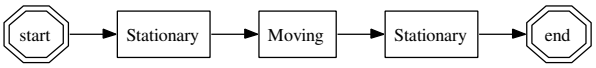
Key words　sign recognition, video processing, HMM, motion extraction, estimation of transitional topology

1. はじめに

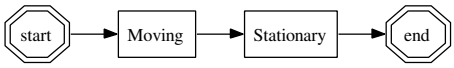
手話動作は，同じ単語であっても発話のそれぞれで速度や動きが異なるため，時間的な伸縮に対応できる隠れマルコフモデル (Hidden Markov Model, HMM) [1] を用いた手話認識が研究されている．HMM は複数の「状態」の時間的な遷移関係と各状態での特徴量の分布の組で表され，手話認識においては各々の状態が手話動作に含まれる「手を上げる」,「手を広げる」などのまとまった動きのそれぞれに対応する．学習の際には各手話単語のそれぞれに対応した HMM を生成し，認識時は入力特徴量を出力する尤度が最も高い HMM に対応する単語を認識結果とする．

手話単語動作では全く同じ動きが必ず現れるわけではなく，発話の度に動作の一部が省略されたり動きが多少変形されたりすることがある.「セーター」を意味する手話では手を胸の前から下げる動作があり，下げる前に静止状態が入る発話も入らない発話も起こり得る (図 1(a)，(b))．このような動作の変化が尤度計算のためのモデルに反映されるのが望ましい．

従来，HMM の状態数や状態遷移構造をあらかじめ単語に関係なく固定し，各状態の特徴量分布のみを学習するという手法 [2] が提案されているが，これでは動きの省略や変形に対応するのが難しい上，図 2 のような構造を全ての単語について用いるので単語毎の動作の複雑さの違いも反映されない．また，ひとつの単語につき複数の HMM を割り当てて動きの違いを表現する方法も考えられる [3]．しかし発話のそれぞれについて HMM を推定すると動きが部分的に省略されただけでも別の HMM となるため，図 3 の後半部分のように共通して現れる動きも別の状態となり，各状態用の学習データが少なくなってし



(a) 動作前の静止状態を含む場合の動き列



(b) 動作前に静止状態がない場合の動き列

図 1　同一手話単語の発話毎の動作の違い（「セーター」）
Fig. 1　Variations of sign with the same meaning (“sweater”).

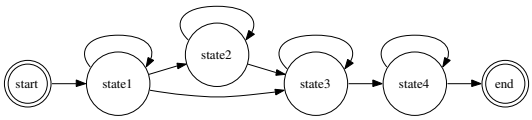


図 2　Starner の用いた遷移構造
Fig. 2　A topology used by Starner.

まう．

そこで本研究では手話単語動作から，枝分かれを持つ単一の HMM を自動的に推定する手法を提案する．動きの省略や追加は図 4 のように HMM 中の状態遷移構造の枝分かれで表現されるため，共通する動きの学習には，その動きを含む全ての学習データを用いることができる．提案法では各単語について

- 単一の発話での動作からまとまった動きの系列を抽出し，
- 複数発話の動き系列から共通する部分を見つける，

という方法で状態遷移構造を構成する．動き属性を分析してから状態を構成するので，各状態に属しているフレームがあらかじめ求まり，学習時のフレームの割り振りの問題も解決でき，動き属性に応じた学習の調整も可能である．

2.　学習用データからの動き系列の抽出

学習用データから「手を上げる」「静止させる」等の意味のある動きに対応したフレーム区間を抽出する．両手を用いる手話の場合には，それぞれの手について動き系列を求めた後，同時刻と見なせるものをペアにし，その列を動き系列と考える．

片手の動きについては，動きの速度と向きに応じて以下のような区間に分割する．

（1）　静止区間.

（2）　移動区間 (手の速度が十分速い区間)．移動区間はさらに以下の 2 種類の区間に分割される．

(2a)　直線動作区間: 一定時間以上に渡って手がほぼ直線的に動いている区間.

(2b)　曲線動作区間: 短い時間に動きの向きが変わっている区間

両手手話においては左右の手の動きの組み合わせが重要となるため，それぞれの手の動きについて求めた区間の内，同時に起こっていると見なすべきペアを見つける必要がある．ここでは図 5 のように，左右の区間が十分長くフレームを共有する場

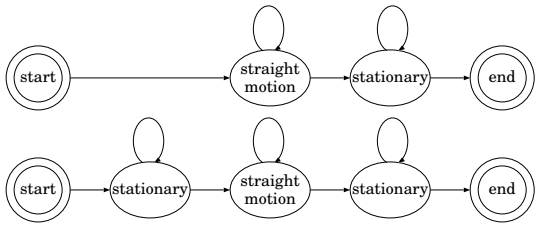


図 3　複数の HMM を用いた動き省略の表現
Fig. 3　Multiple models for representing omission.

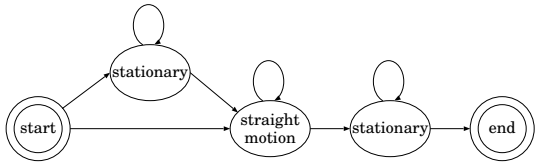


図 4　枝分かれを持つ単一の HMM による動き省略の表現
Fig. 4　Single model with junctions for representing omission.

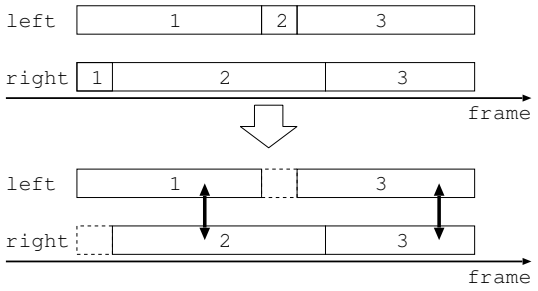


図 5　左右の手についての区間の同期
Fig. 5　Synchronization of intervals of hands.

合，あるいは左右の区間が共有するフレーム内での総移動距離が十分長い区間をペアとする．ペアのつくれない区間はフレーム数も少なく，移動距離も短いうえ，動作のタイミングも揃っていないので重要でないと見なして除外し，ペアとなる区間のみを抽出する．

図 6 は「暖かい」という手話について得られた区間列である．手を上下させる動きのそれぞれが抽出されていることが分かる．

3.　状態遷移構造の推定

2. の手法で，ひとつの学習用画像列からひとまとまりと見なせる動きの区間列を得ることができる．この区間のそれぞれを状態に対応させ，各状態に自分自身への遷移と次の区間に対応した状態への遷移を付加すれば状態遷移構造と見なせる．ここではある区間列から初期モデルを構成し，残りの区間列の情報を初期モデルに追加していくという方法を用いる．

3.1　初期モデル

本手法では必要に応じて状態を追加していくので，初期モデルとしては注目する手話の発話時に共通して現れる特徴を捉えているのが望ましい．ここでは簡単のため，検出される動き区間が最も少ない発話が，その手話に必要な最低限の動きを表していると考える．この発話で検出された動き区間のそれぞれをひとつの状態と見なして初期モデルとする．動き区間の持っていた属性は状態に引き継がれる．すなわち「静止区間」は「静止状態」に,「直線動作区間」は「直線動作状態」に,「曲線動作

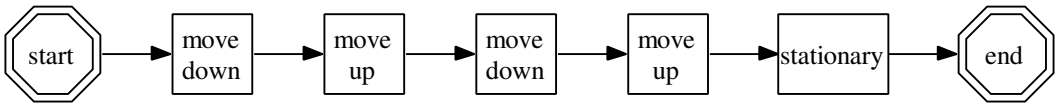


図 6 「暖かい」という発話から抽出した区間列
Fig. 6 Extracted motion segments of the sign “warm”.

区間」は「曲線動作状態」とする．

3.2 状態と状態遷移の追加アルゴリズム

ある状態遷移構造が与えられたときに，新たな区間列を統合する方法は以下の 2 つの段階からなる．

（1） 現在の遷移構造で許される遷移系列の内，新たな区間列に最も良く一致するものを見つける．このとき，区間列のスキップを許す．

（2） 上記の対応関係を元に，スキップされた区間がそれぞれ 1 つ対応する状態を持つように，遷移構造に状態と遷移経路を追加する．

区間と状態の間に一致度を定義し，その総和が最大になるような遷移系列を 1 の「最も良く一致する遷移系列」とする．但し遷移系列としては現在の状態遷移構造で許されるものを考え，対応する区間列もスキップは許すものの元々の順序は保つように選ぶものとする．最適な対応関係は DP マッチング法で求められる．本研究では一致度 C を以下のように定義した．

$$C(\text{状態}, \text{区間}) = \begin{cases} w_p, & \text{(静止状態と静止区間)} \\ -w_n, & \text{(静止状態と移動区間，あるいは移動状態と静止区間の組み合わせ)} \\ \cos(\theta_{\text{segment}} - \theta_{\text{state}}), & \text{(移動状態と移動区間の組み合わせ)} \end{cases} \quad (1)$$

ここで，θ_{segment} はその区間中の総移動方向を，θ_{state} はその状態に割り当てられた全ての区間に渡る平均移動方向をそれぞれラジアンで表したものである．w_p と w_n はそれぞれ静止属性の一致と不一致に関する重みを表しており今回は $w_p = w_n = 1$ とした．

状態と区間の対応関係が得た後，対応する状態を持たない区間毎に新たな状態を遷移構造に追加する．追加される新状態は自分自身への遷移と，区間列中で次に現れる区間に対応した状態への遷移とを持つ．

図 7 は初期遷移構造 (a) に新しい区間列 (b) の情報を統合して，新たな遷移構造 (c) を構成した例である．初期遷移構造は直線動作の前に静止状態を持たないが，静止区間を含む新たな区間列を統合することで新たな遷移経路が追加されている．ここで追加された枝分かれにより，直線動作の前に静止区間を含む動きも含まない動きも許容できるようになる．

3.3 両手の動きのずれを吸収するための短期的状態

3.2 の方法で色々な発話から得られる動きに対応した状態遷移構造が得られるが，実際の発話の場合は両手の動きが完全に同期しているわけではない (図 5)．このため主要な動きを表現する状態の他に，その動きの前後で起こる左右の手の動きのずれを吸収するような状態が必要となる．ここでは各状態遷移に，自分自身への遷移確率が低く，特徴量の分散が大きい「短期的状態」を追加する (図 8)．新しく追加される状態は特徴量の分散が大きいので観測された特徴量の値に関わらず尤度が高いが自分自身への遷移確率が低いのでこの状態には短い時間しか滞在することができず，動きのずれを吸収するのに望ましい性質を持つ．

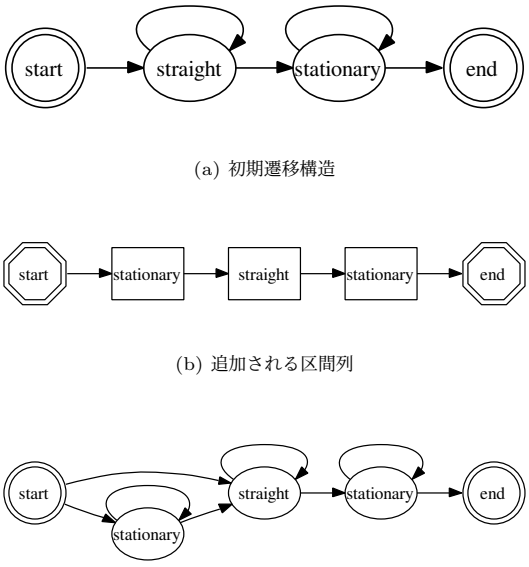


図 7 遷移構造への区間列の統合
Fig. 7 The integration of a series of segments into the initial topology

4. 状態遷移構造の推定実験

実際の手話画像から手領域を抽出して，その重心位置の軌跡から 3. の方法で状態遷移構造の推定を行った結果を示す．以下では簡単のため 3.3 の短期的状態は省略して表示している．

まず，図 7 は「大きい」を意味する手話動作について状態遷移構造の推定を行った結果である．「大きい」の動作は手を胸の前で左右に広がるように動かす動作であるが，実験に用いた画像列では広げる動作の前に静止状態がある場合とない場合とがある．同図 (a) は静止状態のない場合の発話のみから推定した結果であり，静止状態が含まれない発話については許容しにく

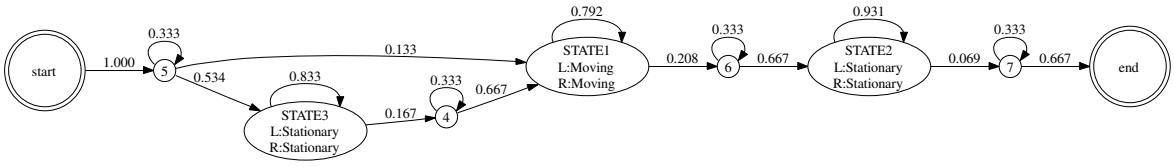


図 8　短期的状態の挿入
Fig. 8　Insertion of temporary states.

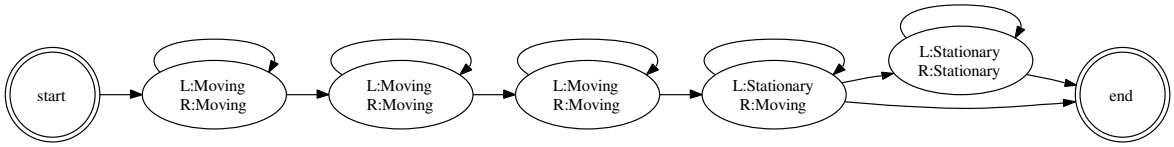


図 9 「暖かい」の遷移構造推定結果
Fig. 9　The estimation result of the sign “warm”.

い構造となっている．一方，同図 (b) は静止状態のない発話も含めて推定した結果であり複数の経路があるため，静止状態が入る発話も入らない発話も許容可能であることが分かる．

図 9 は「暖かい」という意味の手話単語について推定した結果である．この手話は，両手を胸の前まで上げた後，両手を前後に回転させるように上下させる動作である．実験に用いた画像列では，手を上下させる動作の後にすぐ終了し静止状態のないものと，動作の後に静止状態が入るものとがある．共通の 4 状態の後，すぐに終了する経路と静止状態を経て終了する経路の 2 つがこの違いに対応している．

図 10(a)，(b) は「背が低い」を意味する，片手で行う手話での手重心位置の軌跡である．この単語の手話動作は片手を高い位置から下げるという動作となっているが，その動作の前には手を高い位置に移動する準備動作があり，また手話動作終了後には手を下に戻す終了動作がある．図 10(a) は準備動作や終了動作が適切に除外され，手話動作のみを抽出できたときの重心軌跡である．実際に意味のある「片手を高い位置から下げる」という動作のみが含まれている．一方，図 10(b) は同じ手話単語の別の発話における軌跡であるが，準備動作，終了動作の途中で動きの向きが変わっているために手話動作の開始と終了が正しく検出されなかった場合の重心軌跡である．このような場合の状態遷移構造の推定結果が図 10(c) である．同図中央の「手を下げる移動状態」と「静止状態」はこの単語の発話に共通しているが，その前後については複数の経路があり，図 10(a), (b) のような違いを吸収できる構造となっている．

より複雑な手話単語「冬物」に提案法を適用した結果が図 11 である.「冬物」の動作は手を顔の近くで震わせるような動作ののち，両手を上に動かす，というものである．震わせる動作では重心位置の移動が非常に細かいため，静止状態と見なされる場合も移動状態と見なされる場合もあるため，前半部分に複数の経路が生成されている．さらに，後半部分についても手を上に動かした後の，手を戻す動作の有無を吸収するための経路が生成されている．本来，手を震わせる動作に対しては静止状態と移動状態の様々な組み合わせではなく,「震わせる動き」そのものに対応した状態を定義すべきであるが，提案法を単純に用いるだけでも震わせる動作にある程度対応したモデルを構成できる．

以上のように，本論文で提案した状態遷移構造の推定法を用いればその手話単語動作に現れ得る動きを柔軟に許容する状態遷移構造が自動的に構成できる．複数の学習用データから起こり得る動きの情報を総合してモデルを生成するので，固定した構造を画一的に用いるよりも手話動作に適したモデルが得られた．状態遷移構造の推定の際に各状態に属するフレームも求まるので，属するフレームでの特徴量から，各状態のパラメータ (平均，分散) を学習することも容易である．

5. 認 識 実 験

提案手法を実際の手話画像列に適用し，認識実験を行った．今回，特徴量としては以下のようなものを用いた．

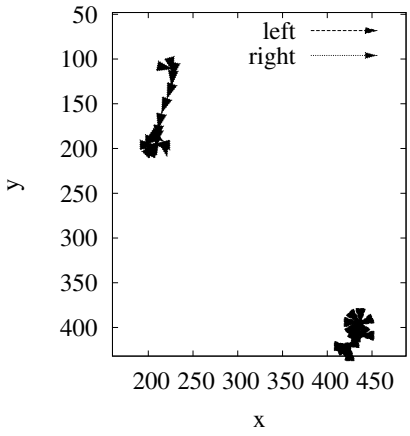
- 対数的な位置ベクトル $\vec{x}_{\mathrm{loghand}} = \mathrm{logvec}\,(\vec{x}_{\mathrm{hand}} - \vec{x}_{\mathrm{face}})$
- 速度ベクトルの向き $\vec{v}_{\mathrm{direction}} = \dfrac{1}{\|\vec{v}_{\mathrm{hand}}\|}\vec{v}_{\mathrm{hand}}$
- 速度の対数 $\log\|\vec{v}_{\mathrm{hand}}\|$

ここで，$\vec{x}_{\mathrm{face}}$，$\vec{x}_{\mathrm{hand}}$ はそれぞれ顔重心と手重心の位置である．また，$\vec{v}_{\mathrm{hand}}$ は手重心の速度ベクトル，$\|\vec{v}_{\mathrm{hand}}\|$ は速度を表している．$\mathrm{logvec}(\vec{x})$ はベクトルを長さに関して対数的に変換する関数であり，

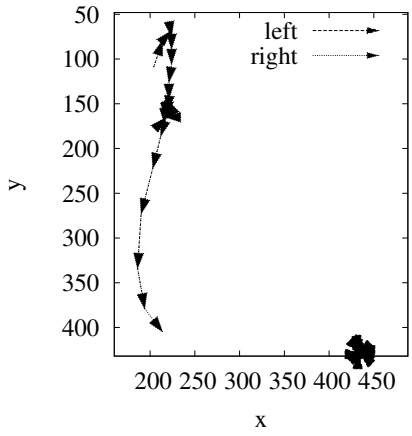
$$\mathrm{logvec}\,(\vec{x}) = \left\{\log\left(1 + \frac{\|\vec{x}\|}{R}\right)\right\}\vec{x} \qquad (2)$$

のように定義される．$\|\vec{x}\|$ が R より十分大きい場合には長さが対数的に変換されるが，$\|\vec{x}\|$ が R に比して小さい場合には線形変換に近くなる．したがってベクトル $\vec{x}_{\mathrm{loghand}}$ は手重心位置が顔重心位置から遠い場合には実際の距離について対数的であり大まかな位置の情報と見なせる．一方，手重心位置が顔重心位置に近い場合には実際の距離について線形的になるので細かい位置についての情報を反映させることができる．手話動作においては顔に近い領域では手の細かい動作が重要になるものの顔から離れた場所では細かい位置はあまり重要でないと思われるので，このような特徴量を導入した．以上のものを両手について考え，$(2+2+1)\times 2 = 10$ 次元の特徴量ベクトルとした．

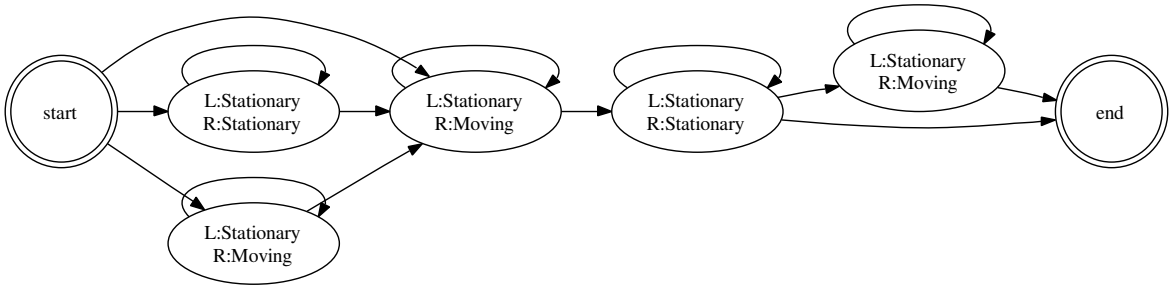
状態遷移構造を推定した後，特徴量ベクトルは各状態毎にガウス分布に従うものと仮定して平均と分散から各状態での特徴量分布を学習した．手話動作としては発話者 2 名で各発話者に



(a) 準備動作が含まれない場合の手重心の軌跡



(b) 準備動作も含まれる場合の手重心の軌跡



(c) 状態遷移構造の推定結果

図 10 「背が低い」の遷移構造推定

Fig. 10 The estimation result of the sign “short”.

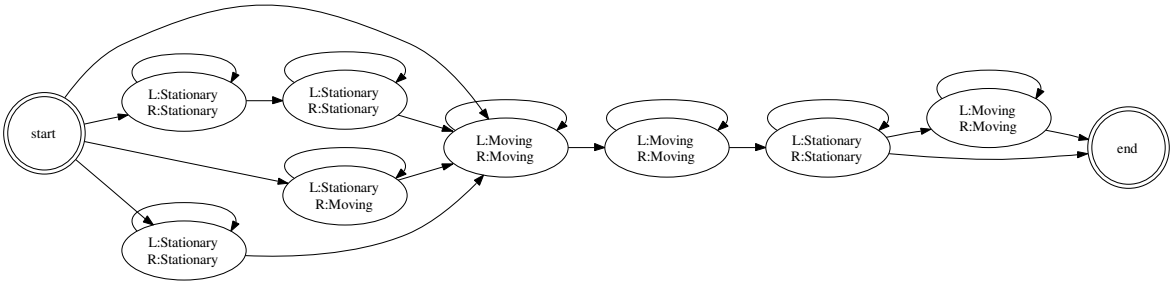


図 11 「冬物」の状態遷移推定結果

Fig. 11 The estimation result of the sign “winter clothing”.

3 回ずつ同じ単語の動作を行ってもらいデータを取得した．各単語につき $3 \times 2 = 6$ 通りの動作データがあるので，このうちの 1 つを認識対象とし，残りの 5 つから推定したモデルで認識を行った．また，比較のため状態遷移構造の推定を行わず，発話毎の動きの違いを複数の HMM を用いた場合の認識実験も行った．認識実験の結果は表 1 ， 2 の通りである．枝分かれのない HMM を複数用いる場合に比べて，概ね認識率が改善していることが分かる．特に両手手話において改善が大きい．両手手話においては両手の動きがどちらも揃っていなければ同じ動きとは見なされないため，発話によって異なる動きが現れやすい．枝分かれのない HMM を複数用いると，それら異なる動きを個別に学習するため，学習データが非常に少なくなってしまう．提案法では異なる動きは別々の状態として扱うが，共通する動きはまとめて学習できるので，認識率が改善したと考えられる．今回の実験では発話者に応じたパラメータ調整等は全く行っておらず，状態遷移構造の推定も自動的に行ったが比較的良好な認識結果が得られた．

6. おわりに

本論文では HMM を用いた手話認識における従来法の問題について述べ，それを解決するため状態遷移構造を自動的に求める方法を提案した．提案法により同じ意味をもつ手話単語の動

表 1　片手手話についての認識実験結果

Table 1　The result of recognition for words with one hand

		発話者 A			発話者 B		
		動作 1	動作 2	動作 3	動作 1	動作 2	動作 3
認識	複数 HMM	8	5	6	6	3	6
正解数	提案法	6	6	7	7	5	6
認識	複数 HMM	100%	63%	75%	75%	38%	75%
成功率	提案法	75%	75%	88%	88%	63%	75%

the total number of words:8

表 2　両手手話についての認識実験結果

Table 2　The result of recognition for words with both hands

		発話者 A			発話者 B		
		動作 1	動作 2	動作 3	動作 1	動作 2	動作 3
認識	複数 HMM	2	2	2	10	7	10
正解数	提案法	14	13	12	13	11	12
認識	複数 HMM	13%	13%	13%	67%	47%	67%
成功率	提案法	93%	87%	80%	87%	73%	80%

the total number of words:15

作の多様性に応じた状態遷移構造を推定することができる．実験の結果，動きのはっきりした手話単語については概ね適切な遷移構造が求められているようである．また，提案法では状態遷移構造の推定時に学習用データとモデル内の各状態の対応関係が得られるので，移動状態であるか静止状態であるかといった動きの属性を用いて，不適切なパラメータ（移動状態における各フレームでの重心位置など）が尤度に悪影響を及ぼさないように調整することも可能である．

今回の実験では手の形状情報を用いていないが，実際の手話動作では特に静止状態において形状情報が重要である．一方，移動状態においては形状情報を精度良く抽出するのが難しい上，移動中の形状自体はそれほど重要でないと思われる．提案法では各状態に対応する学習用データが求まるので，移動状態では形状情報の精度が低いものと仮定して学習するなど，色々な調整が可能である．また，手を震わせる動作や扇ぐ動作など，手重心位置の移動速度が閾値に近い動作が続く場合に，非常に短い静止状態と移動状態が繰り返し検出され状態数が増えすぎる可能性もある．この場合には微小な動きの繰り返しが意味をもつので，一度分割した区間列を再度統合する必要がある．この点についても今後研究を行う予定である．

文　献

[1] 中川：“確率モデルによる音声認識”, 電子情報通信学会，コロナ社 (1988).

[2] T. Starner, J. Weaver and A. Pentland: “Real-time american sign language recognition using desk and wearable computer based video”, IEEE Transactions on Pattern Analysis and Machine Intelligence, **20**, 12, pp. 1371–1375 (1998).

[3] 川東, 白井, 島田, 三浦：“手話の HMM 作成のための状態分割”, 信学技報, 第 105 巻 of WIT2005-21, pp. 55–60 (2005).