

手話認識のためのHMM構造の自動生成

松尾 直志^{*1}, 白井 良明^{*1}, 島田 伸敬^{*1}

Automatic Generation of HMM Topology for Sign Language Recognition

Tadashi Matsuo^{*}, Yoshiaki Shirai^{and} Nobutaka Shimada^{*1}

Abstract – Sign language is used for communicating to people with hearing difficulties. Recognition of a sign language image sequence is challengeable because of the variety of hand shapes and hand motions.

We propose a method to automatically construct a transitional structure(topology) of a Hidden Markov Model(HMM) for recognizing sign language words. Unlike conventional HMM, 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. The topology is constructed from an initial topology by modifying it. With experiments, we show the effectiveness of the proposed method.

Keywords : 手話認識, 動画処理, HMM, 動き抽出, 状態遷移構造推定

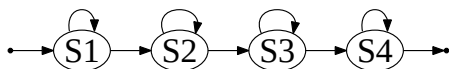


図1 単純な線形遷移構造
Fig.1 A linear topology

1. はじめに

画像列からの手話認識処理は特徴抽出処理と抽出された特徴の認識処理から成る。抽出される特徴は一般に手の位置や速度、形状等の情報であり、認識処理においては各手話単語にひとつの隠れマルコフモデル (Hidden Markov Model, HMM) を対応させるという手法が広く用いられてきた^{[1],[2]}。各々の単語に対応するHMMはいくつかの状態とそれらの間の遷移構造で表される。手話認識においては各々の状態が「手を上げる」、「両手の間隔を左右に広げる」等の単一の意味のある動きに対応しており、状態遷移構造によってこれらの部分動きの連続関係が表現される。

HMMによる手話認識における重要な問題として、同じ意味の手話単語動作であっても、そこから得られる特徴量系列が発話によって大きく異なるという点がある。

Starnerらの手法^[2]では、HMMの状態数は単語に関係なく固定されている。単純な動作であっても複雑な動作であっても同じ状態数のモデルを用いるため、単語ごとに異なる動作の複雑さを反映した学習をさせるのが難しい。

川東らの手法^[3]では、各モデルの状態数は対応する単語動作から自動的に推定されるため、単純な動作に対しては状態数の少ないモデルを、複雑な動作に対しては状態数の多いモデルを用いて学習を行える。しかし、状態数の推定に用いられる閾値は話者毎に手動で調整する必要がある。また、状態遷移構造は図1のような単純な構造に限定されており、同じ動きが同じようにつながった動作にしか対応できない。

Gaolinらの手法^{[4],[5]}ではHMMでの認識処理に加えて新たな認識処理を追加することで、複数種類ある動きのうち、どれであっても許されるような状況に対応している。しかし単語に依らず状態遷移構造は固定されているため、手話単語動作内に含まれる動きの数に応じたモデルにはなっていない。

「各単語動作の複雑さに応じた状態数」、「同じ意味となる複数種類の動きを許容する構造」の問題を解決する方法として、ひとつの単語に複数のHMMを対応させる方法もあるが、その場合各単語動作の学習にはより多くのサンプルが必要となってしまう。

そこで提案法では、ひとつの手話単語に対してその動作の複雑さに応じた枝分かれを含む状態遷移構造を持つ単一のHMMを用いる。単語動作に応じた状態遷移構造の自動生成は「動作の区間分割」と「複数発話からの区間列の統合」によって行う。まず手話単語動作を撮影した画像列を、単一の意味のある動きの区間列に分割し、各区間を状態の候補として扱う。複数の発話に渡って状態候補を収集し、それらから同じ動きと見なせるものを見つけて縮約することで状態遷移構造を生成する。

*1: 立命館大学 情報理工学部

*1: Department of Human and Computer Intelligence, Ritsumeikan University

2. 学習用サンプルの区間分割

本発表で用いる特徴は^[3]の方法で抽出する。各区間が意味のある部分動きに対応するように区間分割するための特徴として、手の動きの向きと速度を用いる。これは、手の位置や形状等の特徴量は話者によって、また同じ話者であってもその発話毎に大きく異なるためである。区間分割の第一段階として、手領域の重心位置の移動速度によって各フレームを静止しているフレーム、あるいは移動しているフレームに類別する。第二段階として、連続する静止フレームは静止区間としてまとめ、連続する移動フレームは、向きが揃っているものがある程度連続していれば直線動き区間としてまとめる。移動フレームではあるものの短時間の内に向きが頻繁に変わるものは振動区間として扱う。

3. 認識に用いる特徴量

一般に手話においては、手の細かい動きに重要な意味のある場合、手は顔に近い場所で動かされる傾向がある。そこで、手領域の重心位置を表す特徴量として、顔領域の重心からの相対座標 x をとり、それを以下の (1) 式で対数的に変換したものをを用いる。

$$L_{r_0} \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) = \frac{\log \left(1 + \frac{\sqrt{x^2 + y^2}}{r_0} \right)}{\sqrt{x^2 + y^2}} \begin{bmatrix} x \\ y \end{bmatrix} \quad (1)$$

r_0 は定数であり、ここでは初期フレームでの顔・手間の距離としている。手の位置を表す特徴量 y は、顔領域の重心位置からの相対座標 x を用いて (2) 式のようにして得る。

$$y = L_{r_0}(x). \quad (2)$$

手を動かす速度は話者によって大きく異なるので、速度の値そのもので単語を識別するのは難しいが、静止しているかどうかという情報は認識に有用である。そこで手の移動速度を表す特徴量としては通常の画素単位の座標系で表した速度 v を (3) 式で対数変換した $\tilde{v}$ を用いる。

$$\tilde{v} = \log(\max\{v, 1\}) \quad (3)$$

ほとんど静止している状況で移動速度が非常に遅い場合には移動方向が安定に定まらないので、手の移動方向を表す特徴量として単純に速度ベクトルを長さ 1 に正規化したものを採用するのは望ましくない。また、手の移動方向も顔から離れた領域では意味が弱いと考えられるので、顔を中心とした相対座標をさらに顔からの距離によって対数的に変換した座標系で速度を求める。手の移動方向を表す特徴量としては、通常の座標系での速度ベクトルの長さを顔を中心とする対数座

標系で求めた長さに変換して得られるベクトル $\hat{v}$ を用いる。

左右の手の相対位置関係は、両者が接近している場合には細かい位置が重要となるが、両者が離れている場合には大まかな位置が重要で細かい情報はそれほど重要でない。そこで、両手の相対位置関係を表す特徴量 y_{rel} は、手の重心位置を表す特徴量と同様、(1) 式を用いて対数的に変換した座標を用いる。

$$y_{\text{rel}} = L_{r_1}(x_{\text{right}} - x_{\text{left}}) \quad (4)$$

(4) 式の r_1 は定数であり、ここでは初期フレームでの両手間の距離としている。 x_{left} 及び x_{right} はそれぞれ左手、右手の重心位置ベクトルである。

4. 状態遷移構造の生成

状態遷移構造の生成は、単一の発話から得られる特徴量列を初期モデルを生成し、他の発話を区間分割して得られる区間列の情報を初期モデルに統合していくことで行う。

4.1 初期モデル

提案法では、初期モデルと他の発話から得られる区間列を比較し、まだ含まれていない動きを追加していくので初期モデルは多くの発話に共通して現れる動きを含んでいるのが望ましい。ここではある単語について得られた発話の内、区間分割の結果の区間数が最も少ないものを初期モデルの基として採用する。この区間列に含まれる区間のそれぞれを状態と見なして初期モデルとする。単一の発話に含まれる動きを並べたものなので初期モデルの構造は図 1 のような、枝分かれのない構造となる。

4.2 区間列の統合

ある発話から得られる区間列を入力とし、これをモデルに統合する処理は以下の 2 つの手順から成る。

1. 入力区間列の各区間での動きと、モデル中の状態が表す動きとを比較し、同じ動きを表す区間と状態の対応関係を求める。
2. モデル中に対応する状態を持たないような区間があれば、新たな動きを許容できるようにモデルに新たな状態と遷移構造を追加してモデルを拡張する。

手順 1 において、対応関係は状態 S_i と区間 seg_j 間の類似度 $C(S_i, \text{seg}_j)$ を用いて定める。ここでは (5) 式で表される総合的な評価値 S を最大にする対応関係 $\{S_{i(k)}, \text{seg}_{j(k)}\}$ を採用する。但し S を最大化する際には、区間のスキップは許すものの区間列の順序は保つような対応関係のみを考える。状態についても、どの k についても $S_{i(k)}$ から $S_{i(k+1)}$ までモデルの遷移構造を辿ることが可能な組み合わせ $\{S_{i(k)}, \text{seg}_{j(k)}\}$ に制限

する.

$$S = \sum_k C(S_{i(k)}, \text{seg}_{j(k)}), \quad (5)$$

式中の S_i はモデル中の i 番目の状態, seg_j は入力区間列の j 番目の区間である. 条件を満たす最良の対応関係は DP マッチングによって求められる. 本発表では類似度 $C(S_i, \text{seg}_j)$ として (6) 式のものをを用いる.

$$C(S_i, \text{seg}_j) = 1 - \frac{1}{T_j} \sum_{t \text{ in seg}_j} \left\{ \frac{d_i(\mathbf{f}_t)}{\sigma} \right\}^2, \quad (6)$$

$$d_i(\mathbf{f}_t) = \sqrt{(\mathbf{f}_t - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{f}_t - \boldsymbol{\mu}_i)}$$

(6) 式中の $\mathbf{f}_t$ はフレーム t での手の移動速度を対数的に変換した (3) 式の $\tilde{v}$ と, 対数的に変換した長さを持つ手の速度ベクトルを両手についてまとめて得られる 6 次元の特徴量ベクトルである. $\boldsymbol{\mu}_i$ と Σ_i はそれぞれ状態 S_i での特徴量ベクトルの平均と共分散行列であり, d_i は状態 S_i の特徴量の平均値 $\boldsymbol{\mu}_i$ と特徴量 $\mathbf{f}_t$ とのマハラノビス距離を表している. σ はマハラノビス距離の閾値を表しており, ここでは 3.0 としている. つまりマハラノビス距離が 3.0 のとき, 類似度 $C(S_i, \text{seg}_j) = 0$ となる. $C(S_i, \text{seg}_j)$ が負となるような組み合わせは評価値 S を減少させるので, 最良の対応関係にはそのような組は現れない. したがって σ は区間が状態と「同じ動き」と見なされるためのマハラノビス距離の閾値といえる.

続く手順 2 では, 条件を満たしつつ S を最大化する対応関係 $\{S_{i(k)}, \text{seg}_{j(k)}\}$ に含まれなかった区間, つまりモデル中のしかるべき場所に「同じ動き」の状態の存在しなかった区間に対応する新しい状態を, モデルに追加する. また, 対応関係 $\{S_{i(k)}, \text{seg}_{j(k)}\}$ に現れる状態遷移 $S_{i(k)} \rightarrow S_{i(k+1)}$ の中でモデルには含まれないものがあれば, それをモデルにも追加する. この統合処理によって, 入力の発話の動きが含まれるよう, モデルが拡張される.

4.3 統合処理の例

提案法による状態遷移構造生成の実験例を示す. 図 2 は「合う」を意味する手話単語動作を 2 名の話者に発話したときの画像列の一部である. ここでは図 2(a) の発話 A から得られる区間列を初期モデルとし, 図 2(b) の発話 B から得られる区間列を統合する場合について考える. これらの発話における右手領域の重心位置の軌跡は図 3 のようになっており, 右手を左手に合わせるまでに描く弧の大きさ, 合わせた後の手を戻す動作の軌跡が異なっている. 発話 A から得られる区間列を基に生成した初期モデルが図 4, 発話 B から得られる区間列が図 5 である. この初期モデル, 区間列について, 状態, 区間の類似度 $C(S_i, \text{seg}_j)$ を求めると

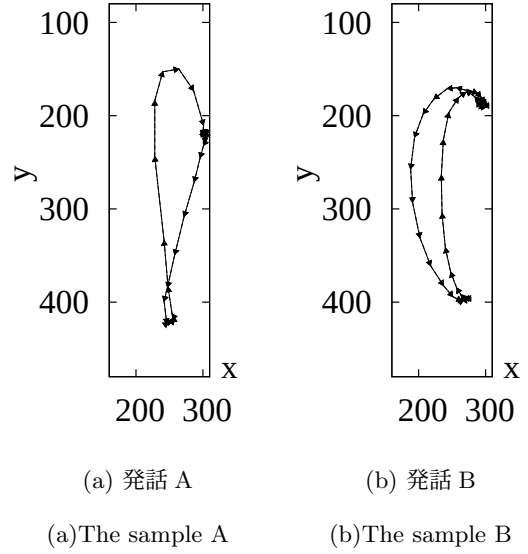


図 3 右手重心位置の軌跡
Fig. 3 The trajectory of right hand motions

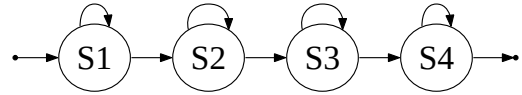


図 4 発話 A から得られる初期モデル
Fig. 4 The initial model from the sample A

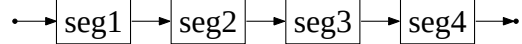


図 5 発話 B から得られる区間列
Fig. 5 The segment sequence from the sample B

表 1 のようになった. 四角で囲われた数字は, 評価値 S を最大の対応関係で採用された組み合わせである. この組み合わせを図示したものが図 6 である. ここで, 図 4 の初期モデルにおいて, 状態 S_1 は手を上げる動作, S_2 は弧を描くように手を動かして両手を接触させるまでの動作, S_3 は接触して静止している動作, S_4 は手を直線的に動かして初期位置に戻す動作にそれぞれ対応している. 一方, 発話 B でも手を上げた (seg_1) 後, 弧を描く動作があるが, その軌跡は非常に小さな弧となっている (図 3(b)) ため意味のある動き区間としては採用されず, 手を上げる区間 seg_1 の後すぐに両手を接触させて静止させる区間 seg_2 が続いている. 発話 B では手を接触させた後, 大きく弧を描くように動かした (seg_3) 後, 直線的に手を戻している (seg_4). 図 6 から, モデルと区間列の動きの内, 「同じ動き」に対応する状態と区間が正しく対応付けられているのが分かる. この対応関係から, 初期モデルには状態 S_1 から S_3 への状態遷移及び区間 seg_3 に対応する状態が不足していることが分かる. これを補

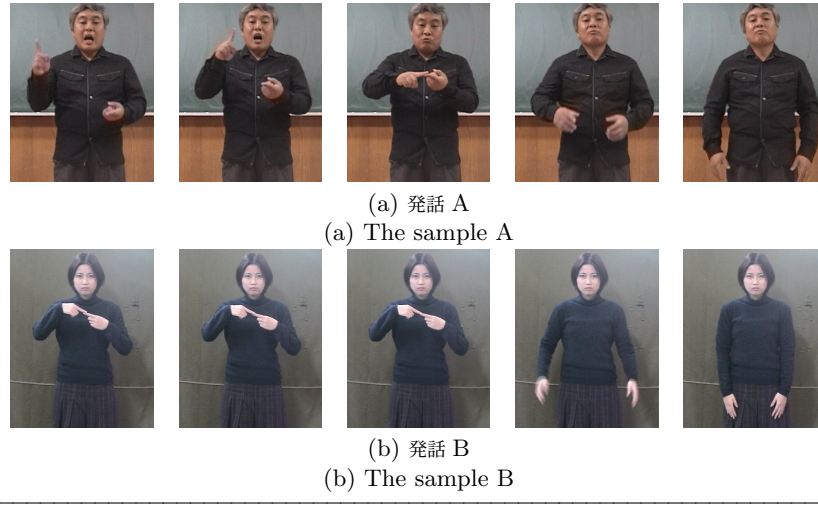


図 2 「合う」の手話単語動作
Fig. 2 The motion of the word “match”

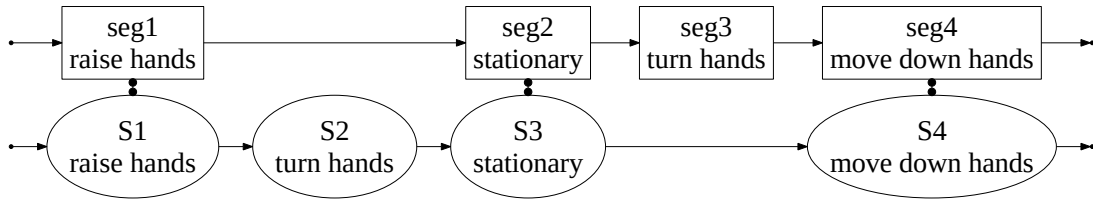


図 6 評価値 S を最大にする対応関係
Fig. 6 The matching result

表 1 状態と区間の類似度
Table 1 The similarity between a state and a segment

	seg ₁	seg ₂	seg ₃	seg ₄
S_1	0.94	-11.7	-12.3	0.38
S_2	0.25	0.53	-10.26	0.55
S_3	-15.8	0.61	-20.1	-19.3
S_4	0.42	-11.4	0.01	0.53

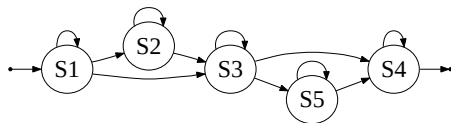


図 7 生成された状態遷移構造
Fig. 7 The generated topology

うと図 7 のような構造が得られる。このモデルは初期モデルに採用された発話 A (図 2(a)) の動きと発話 B (図 2(b)) の動きの両方を含むモデルとなっている。

2 つ目の実験例として、各発話での動作位置が違っていても正しく動きを認識できる例を示す。図 8 は「暖かい」を意味する手話単語動作での手重心位置の軌跡を示したものである。「暖かい」の動作は体の前で手を前後に回転させるような動作となっており、別々

の話者から得られた発話 C と発話 D (図 8(a),(b)) では回転の中心位置も回転の大きさも異なっている。しかしこれは区間分割を行えば「両手を同時に上げる」動きと「両手を同時に下げる」動きの繰り返しとなっており、「同じ動き」と見なすことができる。これらの発話から、提案法により生成した状態遷移構造が図 9 である。状態 $S_1, S_2, \dots, S_5$ は手を上げる動きと手を下げる動きに交互に対応しており、発話 C と発話 D とが同じ動きであると思なされている。状態 S_6, S_8 は手の回転動作が終わった後の速度の遅い状態である。どちらの発話でも速度の遅い状態ではあるが、速度の特徴量を求める際の対数変換の効果が顔からの距離によって異なるため顔に近い位置いるか、遠い位置で静止しているかによって対数変換の効果が異なるため、速度の値が違うと思なされ別の状態となっている。

3 つ目の実験例として、より複雑な動きを持つ単語「冬物」についての結果を示す。この単語動作は手を顔の近くまで上げ、手を握った状態で震わせた後、円を描くように両手を動かして手を上げ、そして戻すという動作である。この単語についての構造推定結果は図 11 のようになった。前半の状態 S_1, S_8 は手を上げる動きに対応しており、二人の話者で手を上げる経路が異なるため別々の状態となっている。 S_{10}, S_{12} は手

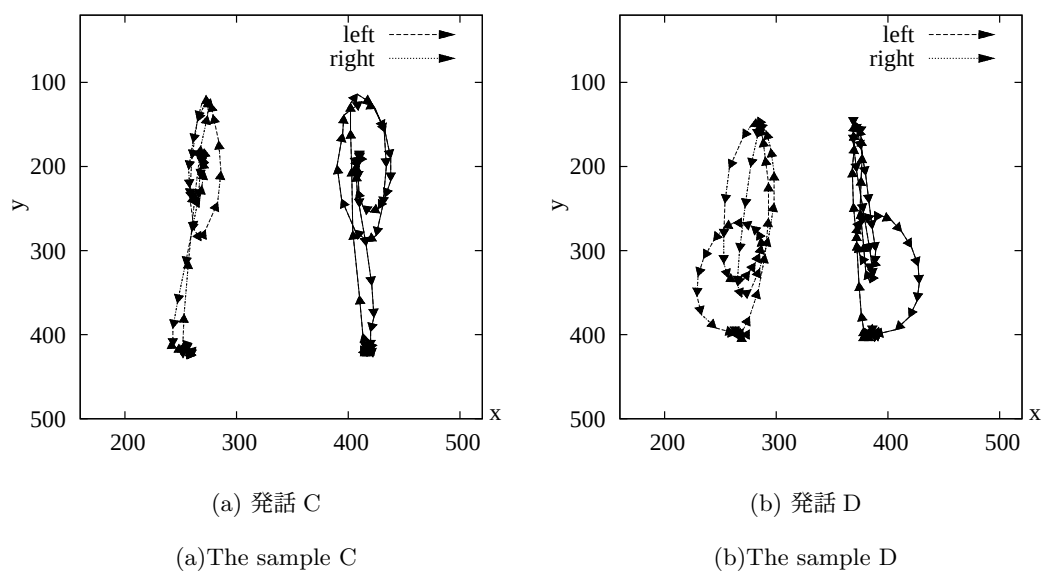


図8 「暖かい」を意味する手話単語動作での手重心位置の軌跡
Fig.8 The motions for the word “warm”

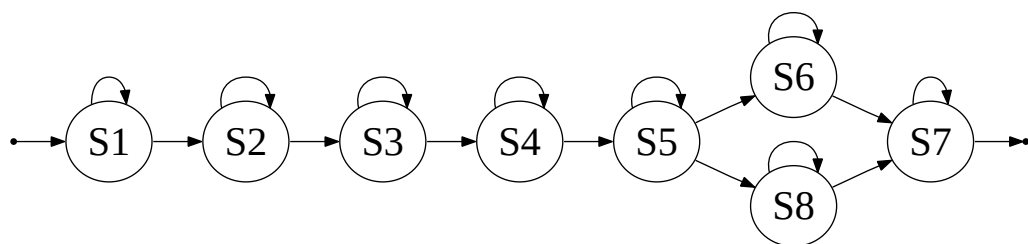


図9 「暖かい」の発話からの遷移構造の推定結果
Fig.9 The estimated topology for the word “warm”

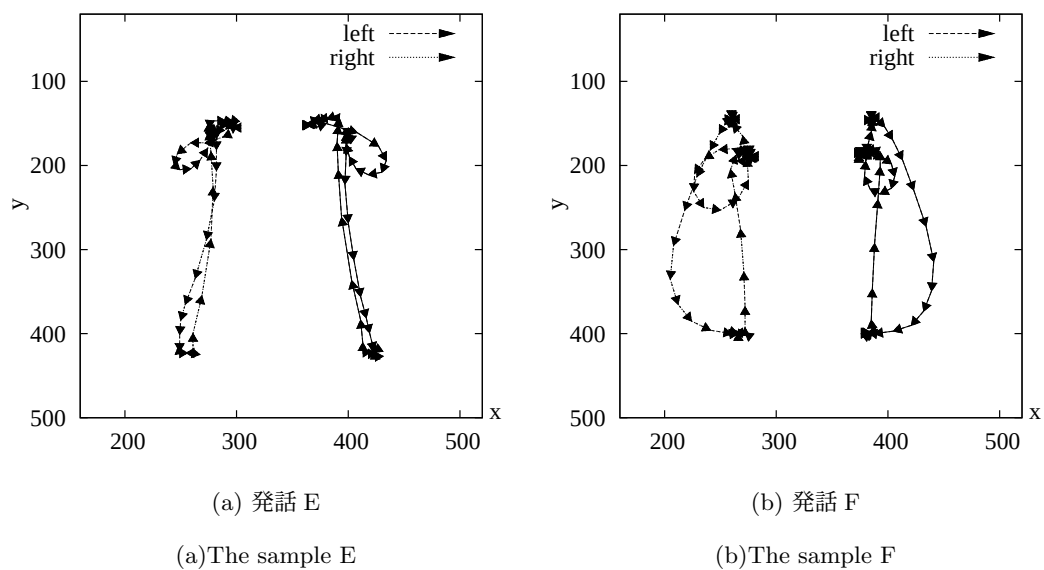


図10 「冬物」を意味する手話単語動作での手重心位置の軌跡
Fig.10 The motions for the word “winter clothing”

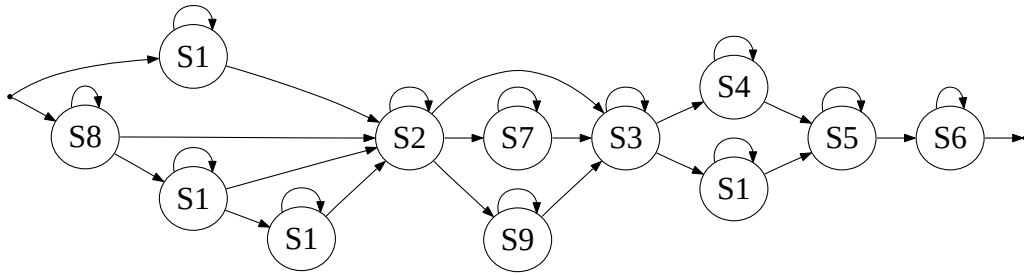


図 11 単語「冬物」についての状態遷移構造推定結果
Fig. 11 The estimation result for the word “winter clothing”

を震わせる動きに対応しているが、震わせる動きは発話によって速度が速い場合も遅い場合もあり、また振れ幅も大きい場合も小さい場合もあるため複数の状態と状態遷移が生成されている。状態 S_2 は震わせた後の静止状態に対応するが、震わせる際の振れ幅が非常に小さい発話の場合、実際には震わせている動きが画像上では静止していると見なされて、この状態に含まれることもある。状態 S_2 以降の枝分かれは手を円を描くように動かす部分であるが、ある話者は発話 E のように、手を震わせていた位置から水平に移動した後、円軌道を描き、またある話者は震わせていた位置から円軌道を描き始める等の違いによるものである。円軌道を描いた後、手を静止させる状態 S_5 、手を初期位置に戻す状態 S_6 は共通の動きと見なされている。手を振動させる動きは発話毎の違いが大きいためまとめた状態を作るのは難しいが、後半の部分に関しては単語「冬物」として許容される動きの多様性を反映した構造になっている。

発話によって手の動きの大きさや動かす位置は異なるものの、手を上げる、下げる、という動きを単位として区間分割を行っているため、発話毎の重要な差異に依存せずモデルが生成できているといえる。

5. 認識実験

手話単語動作認識では、節 3. で定義した以下の特徴量を用いる。

- y 顔を中心とする対数的座標系で表現した手の重心位置
 - $\tilde{v}$ 手の移動速度を対数変換した値
 - $\hat{v}$ 顔を中心とする対数的座標系で表現した手領域重心の速度ベクトルを変換したもの
 - y_{rel} 左手領域の重心位置に対する右手領域重心の相対座標を対数変換したもの
- $y, \tilde{v}, \hat{v}$ についてはそれぞれ左右の手について求めるので、特徴量の次元は $(2 + 1 + 2) \times 2 + 2 = 12$ 次元である。

この実験では、2 人の話者のそれぞれに各単語につき 3 回ずつ動作を行ってもらい、23 単語についての

表 2 提案法、従来法による手話単語動作の認識結果

Table 2 The results of sign language recognition for words with both hands

	Speaker 1	Speaker 2
previous method [3]	59(85.5%)	59(85.5%)
proposed method	62(89.9%)	53(76.8%)

23 words, 3 motion for each word

単語動作を集めた。各単語につき 1 つの発話を認識対象として、残りの 5 つの発話から HMM を生成して認識実験を行った。比較のため、話者について手動で微調整を行う手法である川東らの方法 [3] でも同様の実験を行った。実験の結果は表 2 の通りである。提案法の認識率は 70% 以上で、手動での調整を必要とする川東らの方法に近い認識性能が得られている。提案法は手動調整によって得られた HMM に近い認識性能を持つ HMM を話者毎の調整等なしで生成できるといえる。

6. おわりに

手話単語動作を認識するための HMM を、状態遷移構造も含めて自動生成する方法を提案した。提案法は手話単語動作を意味のある単一の動きに分割する区間分割処理と、複数発話から共通する動きを見つけ縮約する統合処理から成っている。提案法を用いれば、各々の単語の持つ動きの多様性を反映したモデルを自動的に生成することができる。

参考文献

- [1] Grobel, K. and Assan, M.: Isolated sign language recognition using hidden Markov models, *Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation', 1997 IEEE International Conference on*, Vol. 1, pp. 162–167 vol.1 (12–15 Oct 1997).
- [2] Starner, T., Weaver, J. and Pentland, A.: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 12, pp. 1371–1375 (1998).
- [3] 川東香奈, 白井良明, 島田伸敬, 三浦純: 手話の HMM

作成のための状態分割, 信学技報, **Vol. 105**, No.67, pp. 55–60 (2005).

- [4] Gao, W., Ma, J., Wu, J. and Wang, C.: Sign language recognition based on HMM/ANN/DP, *International Journal of Pattern Recognition and Artificial Intelligence*, **Vol. 14**, No. 5, pp. 587–602 (2000).
- [5] Fang, G., Gao, W. and Zhao, D.: Large vocabulary sign language recognition based on fuzzy decision trees, *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, **Vol. 34**, No. 3, pp. 305–314 (May 2004).

