修 士 論 文

隠れマルコフモデルを用いた
手話単語認識システム

指導教官　堀口 進 教授

北陸先端科学技術大学院大学
情報科学研究科 情報システム学専攻
マルチメディア統合システム講座

810012 伊藤 徳広

2000 年 2 月 15 日

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 研究の背景と目的

近年、聴覚障害者の社会進出にともなって、聴覚障害者と健聴者がコミュニケーションをとる機会も増えている。聴覚障害者と健聴者がコミュニケーションをとる手段として、主に手話や筆談、および手話通訳士による手話の通訳などが考えられる。しかし、筆談は聴覚障害者、健聴者双方に負担がかかる上、コミュニケーション速度も遅いという難点がある。一方、手話通訳士を介する場合、双方のコミュニケーションは最も容易な域になるが、手話通訳サービスは予約が必要であり、手話通訳士の数も限られていることから、利便性に問題がある。そのため、聴覚障害者の大きなにコミュニケーション手段である手話を自動認識するシステムへの要求が高まっており、さまざまな手法での手話認識が試みられている。

手話認識システムでは、入力データとして画像を使う場合と、3 次元空間中での手指の位置座標を使う場合がある。後者は近年のデバイス技術の発達により、位置及び手形状を数値化する手形状入力装置を用いて細かな手形状やその向きを読みとる事が可能なので、認識可能な単語数の増加が容易となってきている。この手形状入力装置を用いた手話認識システムとして佐川ら [1] が行なった DP マッチング法を用いたものが挙げられる。このシステムでは 620 語もの単語を 98.7 %という高い認識率で認識可能であった。しかしな

がら、この認識結果を出すためには個人データで構成されている辞書が必要であるという問題がある。DP マッチング法はその認識率を上げる為に拘束条件を厳しくすると、個人の癖などの影響を受けやすくなる傾向があるため、使用者を限定する事になってしまう。より多くの話者が使用できる手話認識システムを構築するためには、個人の癖を取り除くような認識手法を検討する必要がある。

近年、音声分野で使われていた隠れマルコフモデル（HMM）を手話認識手法に使用する試みが行なわれている。HMM は学習という過程を通して統計的にパターンを処理出来るので、データの揺らぎに強いという特徴を持っており、不特定話者の単語認識に適している。そこで本論文では、入力デバイスとしては装着型の手形状入力装置を、認識モデルとしてはHMM を用い、より多くの話者に対応した手話認識システムの構築を試み、その結果について検証した。

1.2 本論文の構成

本論文の構成は以下の通りである。第二章で過去の研究における問題点をまとめ、本論文で使用するHMM の概要を述べる。第三章では本研究で提案する手話単語認識システムの構成を示す。さらに、手話単語認識で必要となる手話単語切り出し手法、およびHMM の学習収束判定に関して検討する。第四章では手話単語認識システムの性能評価を行ない、認識率や未知話者での認識結果に関して議論する。第五章において本研究で得られた結果をまとめ、結論とする。さらに今後の課題を示す。

第 2章

隠れマルコフモデルでの手話単語認識手法

2.1 初めに

本章では、従来の手話単語認識システムについて述べ、その特徴や問題点を提示する。次に本論文で使用する隠れマルコフモデルの概要を述べ、このモデルを使用した手話単語システムの認識過程を示す。

2.2 従来の手話単語認識手法

聴覚障害者は他者とコミュニケーションを取る手段として、主に手話や筆談などを用ている。特に手話はコミュニケーション速度という点において筆談より優れているため、聾唖者の日常的なコミュニケーション手段としてよく用いられている。しかし現状では手話を理解出来る人間の数は少なく、聴覚障害者が社会生活の上で他者との意志疎通にストレスを感じるケースも多い。また、他者との意志疎通の方法として手話通訳士を挟んで行なう場合もあるが、この場合でも手話通訳士の利用制限やプライバシー問題などの数多くの課題がある。この様な状況から、コンピュータによる手話認識システムの要望が発生しており、現在までにさまざまな研究がなされている。

2.2.1 DP マッチング法

手話動作の認識において最も有名な手法の一つとして佐川ら [1] の DP マッチングによる認識方法が上げられる。この手法は識別対象を特定の個人に限定した場合、90 %以上という非常に高い認識率を得る事が出来る。

DP マッチング法では、まず動的計画法（Dynamic programing）に基づいて各パターン間の距離を定義し、辞書中のパターンと入力パターン間の距離を計算して、最も距離の小さいパターンを認識結果として採用する手法である。

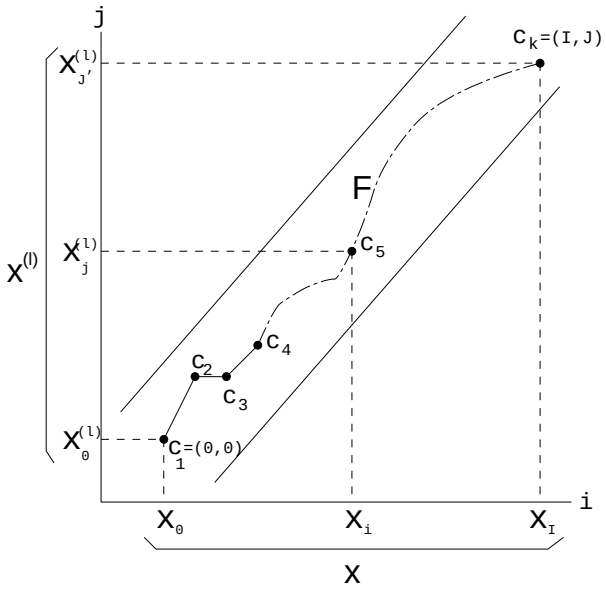


図 2.1: ＤＰマッチング

例えば長さが I、J' の時系列パターン X、$X^{(l)}$ があるとする。この時、図 2.1 に示した様に $X, X^{(l)}$ を対応づける関数 F を考えると、関数 $X, X^{(l)}, F$ は次の様に定義する事が出来る。

$$X = x_0, x_1, x_2 \cdots, x_i, \cdots, x_I \tag{2.1}$$

$$X^{(l)} = X_0^{(l)}, x_1^{(l)}, x_2^{(l)} \cdots, x_i^{\prime(l)}, \cdots, x_I^{\prime(l)} \tag{2.2}$$

$$F = c(1), c(2), c(3) \cdots, c(k) \tag{2.3}$$

ここで、F はパターン $x, x^{(l)}$ で構成される i, j 平面上の点を結んだ関数である。これに重み w_k を定義すると、時系列パターン $x, x^{(l)}$ の距離 d は、

$$d(x, x^{(l)}) = \min_{F} \left\{ \frac{\sum_{k=1}^{K} w_k d_k(x_{i'}, x_i^{(l)})}{\sum_{k=1}^{K} w_k} \right\} \tag{2.4}$$

と表せる。なお、重み w_k とは関数 F に関連した正の係数である。ここで $w_k = (i_k - i_{k-1}) + (j_k - j_{k-1}), (i_0 = j_0 = 0)$ とすると $\sum_{k=1}^{k} w_k = I + J$ となる。

これにより式 (2.4) の分母は F の内容に依存せず、つねに一定値 $I + I'$ となる。よって時系列パターン $x, x^{(l)}$ の距離 $d(x, x^{(l)})$ は次のようになる。

$$d^{(l)}(x) = \frac{1}{I + I'} \min_{F} \left\{ \sum_{j=1}^{J} w_j d_j(x'_i, x_i^{(l)}) \right\} \tag{2.5}$$

この様に各パターンとの距離を算出し、最も距離が近いものを選択パターンとして決定する手法が DP マッチングである。しかし、この重みづけと整合窓の大きさのパラメータを一般化するのが困難である、という問題点が上げられる。特に、複数話者を対象とした場合、パラメータ設定が個人差により大きな影響を受ける為、より一層設定が困難になってしまう傾向がある。

2.2.2 FFT による認識手法

DP マッチングではその性質上、同一カテゴリに属するデータの分散はある程度少なくなければならないが、不特定話者を対象とした場合、どうしてもデータに大きな分散が生じるので認識が困難となる。そこで、時系列データである手話をフーリエ変換して、周波数成分で識別する手法を鈴木 [2] が提案している。

入力されるデータを N 個の離散データ x_n とした場合、有限区間の離散的フーリエ変換法である DFT を用いると、離散データ x_n とそのフーリエ係数 X_k は式 2.6,2.7 で表せる。

$$X(k) = \begin{cases} \sum_{n=0}^{N-1} x(n) W_n^{kn}, & 0 \leq k \leq N-1 \\ 0, & other \end{cases} \tag{2.6}$$

$$x(n) = \begin{cases} \frac{1}{N} \sum_{k=0}^{N-1} X(k) W_N^{-kn}, 0 \leq k \leq N-1 \\ 0, & other \end{cases} \tag{2.7}$$

時系列データを変換し、周波数成分で比較するには通常、時間領域で同じ長さの区間を区切りFFTを行なうが、手話単語データは時間方向に非線形な伸縮を伴う。そこで、線形補間を用いて固定の長さに時間軸を正規化した後、FFTを行なう事により時間軸上の伸縮問題を解決している。

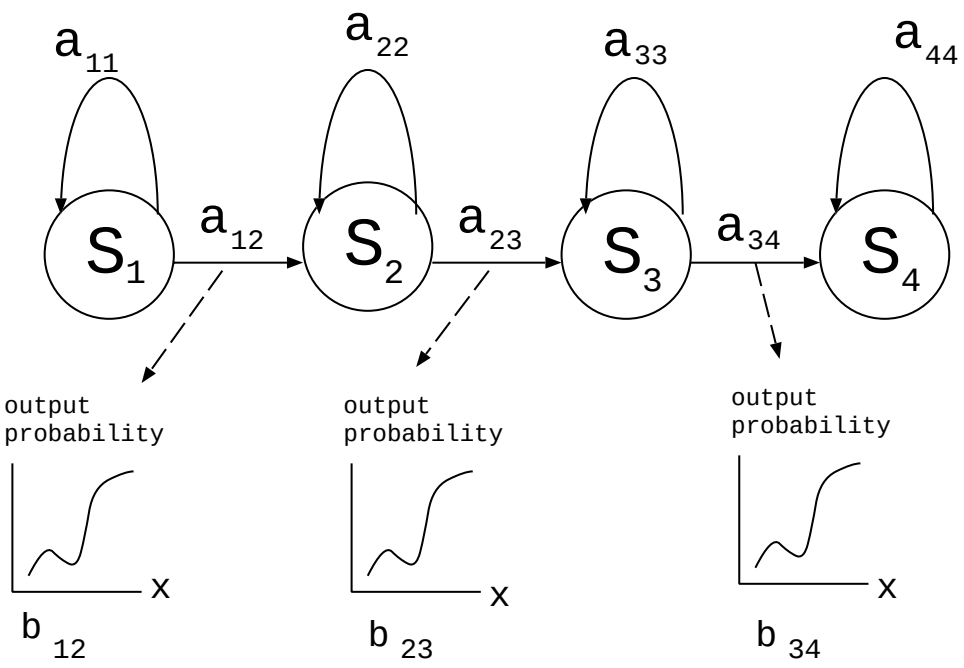
これにより、特徴ベクトルの要素数がN、時系列データを長さMに正規化したデータをFFTで処理すると、周波数領域では複素数値をとるため$N \times 2M$の行列データとなる。これを改めて$N \times M$の行列にし、このデータでそれぞれの要素の平均値と分散を求める事でパターンマッチングが可能となる。

この手法での認識結果3名の複数話者を対象とした場合は83.4 %と高い認識率を示した。しかし、人数が増えていくにしたがって認識率の低下が著しくなるために、不特定話者の認識には十分ではない。

2.2.3 HMMによる認識手法

音声分野で使われていた隠れマルコフモデル（HMM）をジェスチャー認識に使用するという試みが近年行なわれており、その一環としてHMMによる手話認識の研究がなされている。HMMの理論的なモデル説明については、2.4章で詳しく述べることとし、本章では過去の研究成果などついて概要を簡単に述べる。

HMMとは図2.2の様に幾つかの遷移状態S_i（$i = 1, 2, 3, \cdots$）をもつオートマトンの事である。このモデルは各状態に移る遷移確率（a_{ij}）と、その遷移時に、入力事象が発生する出力確率（b_{ij}）をもっている。このモデルに、ある入力事象$X(x_1, x_2, x_3, \cdots, x_I)$が入力されると、入力事象$x_i$ごとに状態$S_i$がさまざまに変化する。その中で、最終の事象$x_I$の時に最終状態（図2.2では$S_4$）となる遷移過程で得られる確率の合計を受理確率と呼ぶ。例えば、図2.2に入力事象$X(x_1, x_2, x_3, x_4)$が入力されると状態S_1から$_4$へ変化す



なお出力確率 b_{ij} は各 a_{ij} にそれぞれ存在する

図 2.2: 隠れマルコフモデル

る経路は４つとなるので、この時の HMM の受理確率 $P(X \mid M)$ は次の様になる。

$$
\begin{aligned}
S_1 \to S_1 \to S_2 \to S_3 \to S_4 : \quad & P_1 = a_{11}b_{11}(x_1) \cdot a_{12}b_{12}(x_2) \cdot a_{23}b_{23}(x_3) \cdot a_{34}b_{34}(x_4) \\
S_1 \to S_2 \to S_2 \to S_3 \to S_4 : \quad & P_2 = a_{12}b_{12}(x_1) \cdot a_{22}b_{22}(x_2) \cdot a_{23}b_{23}(x_3) \cdot a_{34}b_{34}(x_4) \\
S_1 \to S_2 \to S_3 \to S_3 \to S_4 : \quad & P_1 = a_{12}b_{12}(x_1) \cdot a_{23}b_{23}(x_2) \cdot a_{33}b_{33}(x_3) \cdot a_{34}b_{34}(x_4) \\
S_1 \to S_1 \to S_2 \to S_3 \to S_4 : \quad & P_1 = a_{12}b_{12}(x_1) \cdot a_{23}b_{23}(x_2) \cdot a_{34}b_{34}(x_3) \cdot a_{44}b_{44}(x_4) \\
P(X \mid M) = P_1 + P_2 + P_3 + P_4 &
\end{aligned}
\tag{2.8}
$$

HMM とは、こうしてある入力事象 X が、あるマルコフモデル M で起きうる確率を受理確率 $P(X \mid M)$ として表現できるモデルの事である。

これを手話などの動作認識に応用するには、認識動作ごとにマルコフモデルを作成し、各マルコフモデルの遷移確率と出力確率を認識すべき事象が入力された時に、最も受理確率が高くなる様に設定する。これにより、ある入力事象が与えられた時、各モデルの受理確率の違いから、入力事象の識別が可能となる。HMM を使用した手話認識方法 [3] では、画像から得られた手の位置と方向を用い 40 種類のアメリカ手話を 95 パーセント以上の認識率で認識する事に成功している。

HMM は学習によって得られたデータを統計的に処理できるモデルなので、入力データの揺らぎに強いという特徴を持つ。その為、不特定話者での認識などに向いていると考えられる。しかし、マルコフモデルは使用前に予め遷移確率や出力確率を「学習」という形で決定する必要があり、その為に多量のサンプルデータが必要となる。その結果、手話単語の増加に伴って用意するべきサンプルデータ量も膨大となるために、現在約 3000 語といわれる手話の全てを認識するのは困難であることが問題である。

2.3 認識モデル

DP マッチング法などと違って、HMM では学習したデータでパラメータを設定しているのでデータの揺らぎに強く、不特定話者による手話の認識の手法として期待出来る。しかし、従来の HMM を用いた研究では認識率こそ良いものの、各手話単語ごとに一つの HMM を使用していた為、単語数増加に伴って必要となる HMM の数が大幅に増加し、そ

れに比例して学習サンプルも必要になってくる。そこで、本研究では手話単語一つを幾つかの部分単語に分割し、部分単語ごとに認識する方法をとる。これによって、複数の部分単語から構成されている手話単語は、この組合せで表現が可能となる。これによって従来法で問題となる単語学習サンプル増加の問題を解決する。本節では単語の分割に際し、分割の参考となった手話の音韻表現と、それを元に分類した基本動作の概念、両者を統合した単語認識方法について述べる。

2.3.1 手話単語の構造

本研究での認識対象である日本手話は、現在でもその動作を記号などにより表記する完全な方法は確立されてはいない。しかしながら、比較的よく使われる表現方法というものは存在する。そこで、本研究ではこの中で神田ら [4] が考案した手話表記法を元に動作などの認識分類を考えることにした。これによると、手話は大まかには「手の形」「動き」「位置」の 3 要素で構成されており、「手の形」は幾つかを基本的な手形とその変形過程などを、「動き」は空間に描く軌跡やその大きさなどを、「位置」はその描かれた軌跡がどの位置か（例：頭・顔・肩など）などを表現している。表 2.1 に神田らの手話表記法の概要を示す。

手話の音韻表現とは、これら手形や動きなどに記号を割り当て、記号の並びによって動的な情報である手話を、記号という静的なもので表現する事である。

2.3.2 手話単語の基本動作認識

手話使用者の腕の動きに注目すると、その動きは一つ、または複数からなる数種類の線形状から構成されているケースが多い。事実、手話の音韻表現においても直線や曲線・四角などでその動作を表現している。この様な種類の動きは、手話を構成している基本的な動きの要素であると考えられる。以後、この様に手話の動作を構成している基本的な運動を基本動作と呼ぶ。

表 2.1: 日本手話の音韻表記方法

分野	表記方法
手の形	基本的な指の形として４６種類のパターンを採用 （図 2.3 参照） これに指の変化で 6 パターン
動き	方向・軌跡・様態・位置などで構成 ここでの軌跡とは線形状で、直線と曲線の 2 種類 様態は動作の大きさや繰り返しなど
位置	位置は手話が行なわれている場所 例えば顔・頭・肩など

2.4 隠れマルコフモデル

本研究で用いる隠れマルコフモデルとはオートマトンの一種で、元々は音声分野において音声認識モデルとして使われていたモデルである。しかし、認識などで一番困難であるパラメータの設定が、学習という作業で決定できるために、近年さまざまな認識に使われる様になった。オートマトンは初期状態から最終状態までの道筋が、入力によって一意に決まる決定性オートマトンと、どの様な道筋を通るかは不明な、非決定性オートマトンの 2 種類がある。HMM はこのうち非決定性オートマトンに分類される。

HMM は先に図 2.2 に示した様に、幾つかの状態 S_i とそれを結ぶ遷移確率 a_{ij}、各遷移確率に付随する出力確率 $b_{ij}(k)$ から構成されている。HMM にあるパターン系列 $y = y_1, y_2, \ldots y_r$ を入力すると初期状態から最終状態へと状態が変化し、そのパターンが HMM で生起する確率 $P(y \mid M)$（M は HMM によって表現されモデル）を知る事が出来る。この確率 $P(y \mid M)$ は $q = q_{i0}, q_{i1}, \ldots, q_{iT}$ を状態遷移系列とすれば式（2.9）のように書ける。

$$P(y \mid M) = \sum_{i_0, i_1, \ldots, i_r} P(y \mid q, M) \cdot P(q \mid M) \tag{2.9}$$

一般に、$P(y \mid M)$ の値は次の様に求められる。まず、入力として N 個の観測データ $Y = (y_1, y_2, \ldots, y_N)$ が得られたとする。この時、時刻 t に観測データ $y_1, y_2, \ldots, y_t$ を生成して状態 s_i に滞在する前向き確率（フォワード変数）を式（2.10）の様に定義する。

$$\alpha(i,t) = \sum_j \alpha(j, t-1) a_{ji} b_{ji}(y_t) \tag{2.10}$$

なお、a_{ji} は状態 s_j から状態 s_i への遷移確率を、$b_{ji}(y_t)$ は状態 s_j から状態 s_i への遷移の際にシンボル y_t を生成する確率である。今、初期状態から最終状態への遷移可能な全ての状態間遷移ではなく、最大確率を与えるパスのみを求めるとすると、

$$\alpha(i,t) = max_j\{\alpha(j, t-1) a_{ji} b_{ji}(y_t)\} \tag{2.11}$$

式（2.11）を対数変換すると、

$$\log \alpha(i,t) = max_j\{\log \alpha(j, t-1) + \log a_{ji} + \log b_{ji}(y_t)\} \tag{2.12}$$

となり、対数尤度の和により確率を求める事が出来る。この手法をビタビ・アルゴリズム (Viterbi alogrithm) と呼んでいる。

2.4.1 隠れマルコフモデルの特徴

第 2.4 節で述べた様に HMM はあるパターン系列が入力された時、そのパターンが初期状態から最終状態まで遷移しうる確率を求められる。つまり、識別したい時系列パターンを入力した時に、高い確率で最終状態まで行き着けるように遷移状態確率と出現確率を設定しておけば、それに近いパターンでは高い確率で、遠いパターンは低い確率で最終状態に到達する様にする事が出来る。HMM ではこのパラメータ設定過程を学習と呼び、多く

の学習サンプルを使ってパラメータの推定を行なう。学習過程において、各パラメータには学習したパターンの統計的な情報が保存されることとなるので、データの揺らぎに対して強くなるという特性がある。

2.4.2 マルコフモデルの学習法

HMM のパラメータは、入力されたデータに対して起きる状態遷移が観測できないため、直接最尤推定する事ができない。そこで、バウム-ウェルチのパラメータ推定法により、観測シンボル系列 Y が与えられた時 $P(Y \mid M)$ (M は初期確率:π_i, 遷移確率:a_{ij}, 出現確率: $b_{ij}(k)$ で構成されている HMM）が最大となるパラメータを推定することにする。

まず、2.4 節で述べた前向き確率に加えて、時刻 t に状態 s_i に滞在し、観測データ $Y = y_t, y_{t+1}, \ldots, y_T$ を生成する後向き確率 (backward porbability)β を 2.14、2.14 に定義する。

$$p(y_t, y_{t+1}, \ldots, y_T) = \sum_i \beta(i,t) \tag{2.13}$$

$$\beta(i,t) = \sum_j a_{ij} bij(y_t) \beta(j, t+1) \tag{2.14}$$

さらにモデル M が Y を出力する場合において、時刻 t に状態 s_i から s_j へ移行し、シンボル y_t を出力する確率として γ を 2.15 で定義する。

$$\gamma(i,j,t) = \frac{\alpha(i,t-1) a_{ij} b_{ij}(y_t) \beta(j,t)}{P(Y \mid M)} \tag{2.15}$$

すると式（2.10）（2.14）（2.15）を用いて、HMM の各パラメータは次の再推定の繰り返しによって求める事が出来る。

$$\hat{\pi}_i = \frac{\sum_j \gamma(i,j,1)}{\sum_i \sum_j \gamma(i,j,1)} \tag{2.16}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \alpha(i,t-1) \cdot a_{ij} \cdot b_{ij}(y_t) \cdot \beta(i,j)}{\sum_t \alpha(i,t) \cdot \beta(i,t)} = \frac{\sum_t \gamma(i,j,t)}{\sum_t \sum_j \gamma(i,j,t)} \tag{2.17}$$

$$\hat{b}_{ij}(k) = \frac{\sum_{t,y_t=k} \gamma(i,j,t)}{\sum_t \gamma(i,j,t)} \tag{2.18}$$

複数の学習サンプルがある場合は、全ての学習用サンプルに関してこの計算を行なってからパラメータを一回更新し、その値が収束するまで繰り返す。

2.4.3 基本動作を用いた隠れマルコフモデルでの手話単語認識

本研究ではこの HMM を使って手話の基本動作を認識し、手形の情報と併せて最終的に手話単語認識システムを構築する。基本動作の区分は手話の音韻表現を参考にする。しかし、音韻表現においての記述区分では、表記の汎用性を持たせるために動作区分を細かく分けているうえ、その定義も曖昧である。その為、この区分をそのまま認識区分として使用するには問題がある。そこで、本研究ではある程度まとまった動作を一つの動作として扱う事により、音韻表記の場合に起こる定義の曖昧性を抑える。

2.5 ベイズ法による手形認識

手話の音韻表現では指の動きが無いかぎりは、一つの記号で手形を表現している。しかし、実際の手話動作では手話の最中に手形は微妙に変化してしまう。特に手首を動かす運動では、手首の動きにつられて手形も変化しやすい。しかし、通常の会話で使われる手話は、腕の動きが停止している事は殆んど無いと言ってよい。停止した場合でも、すぐに他の手話へ移行する事が多いため、運動停止時のみの手形認識は誤認識しやすいと予想される。その為、手話動作の手形認識は、運動中・停止中どちらとも一つの手形として特定するのは困難である。

そこで本研究では得られたデータの各フレームごとに手形を特定し、一つの手話動作中に含まれる各手形の分布割合を手形認識の方法として用いる事とした。これにより、一つの手形にする場合に問題になる指の曲げ伸し運動にも対応する事ができるうえ、手首などの運動による手形の変化もある程度反映させる事が可能となる。

各フレームにおける手形の認識にはベイズ識別法を用いた。ベイズ識別法を用いた手形の認識では、後藤 [9] の実験結果より、ほぼ 100 %と非常に高い結果認識率が得られている。そこで、今回の研究においてはこの手法を手形認識に利用することとした。

2.6 手話単語認識法

本システムでの手話単語認識システムは、手形・動作をそれぞれ別々の手法で認識し、それを統合する形で手話単語を特定する。これらの認識する部分をまとめると表 2.2 の様になる。なお、ここで運動を行なう位置を考慮に入れないのは、今回認識対象とする手話

表 2.2: 本システムでの認識分類

認識対象	認識分類	使用識別方法
手形	基本的な指の形４６パターンから、 指の形が違う２５パターン （四角で囲った形：図 2.3 参照 ）	ベイズ識別法
動き	軌跡形状や様態を組み合わせた、 手話単語を構成する基本的な動作 軌跡の運動面は KL 法で求める	HMM KL 法
位置	考慮に入れない	

単語が位置情報に依存しないためである。

この手形・動きを手話単語として識別する為には、単語と動作一式を対応づけた辞書を用意する必要がある。そこで、本実験では図 2.4 に示す様な構成の辞書を作成する事にした。辞書内部は単語ごとに手形・基本動作・運動平面の３要素の情報が記されており、各要素はそれぞれ、手話単語中に出現する手形の割合、基本動作ごとの受理確率、運動平面ベクトルが記載されている。

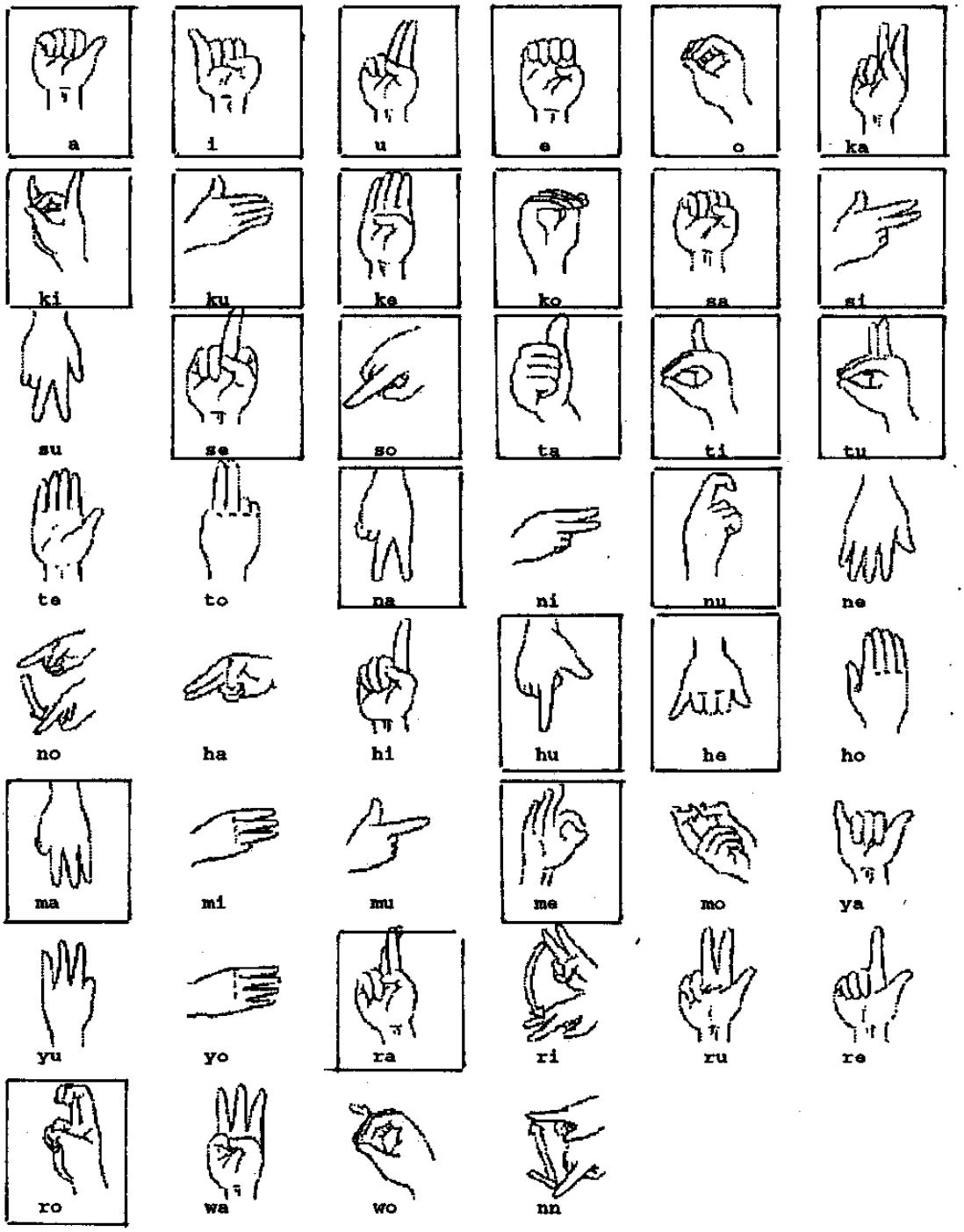

a
i
u
e
o
ka
ki
ku
ke
ko
sa
si
su
se
so
ta
ti
tu
te
to
na
ni
nu
ne
no
ha
hi
hu
he
ho
ma
mi
mu
me
mo
ya
yu
yo
ra
ri
ru
re
ro
wa
wo
nn

図 2.3: 指形状一覧図

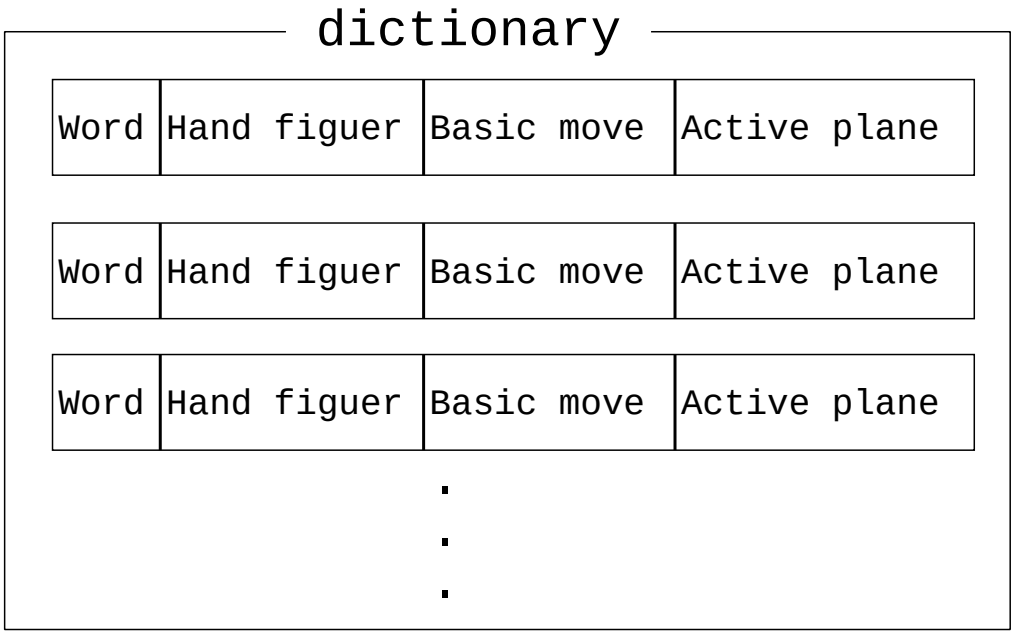


図 2.4: 辞書データ構成

2.7 まとめ

本章ではまず、従来法について述べその問題点を示した。次に本システムで使用する HMM やベイズ識別法について説明した。そして、最後に単語認識に使用する辞書の構成について述べた。

第 3章

手話単語認識システム構成

3.1 はじめに

本章では手話単語認識システムの入力装置について述べる。次に手話単語を入力データから自動的に切り出す為の切り出し法を示した後、HMM の学習方法について説明する。最後に、作成した辞書仕様の詳細などについて記す。

3.2 システム構成

本研究で構築する手話単語認識システムを図3.1、3.2 に示す。システムは手の位置や形を取り込み、データとして格納する入力部分と、入力されたデータから手形・基本動作を認識し、最終的な手話単語を選定する認識部分の 2 部構成になっている。まず、図3.1 の「Sensor Unit」から手形の情報は「CyberGlove」が、姿勢と位置は「FASTRACK」から取り込み「Host computer」内部に「Sign language data」として保存する。この部分までがデータ入力部分となる。

次に、先ほど取り込んだデータを元に認識システムを通過し単語を識別する。流れとしては、「Sign language data」から手形データである「finger data」を取り出し手話開始から終了までの全フレームを「bayes recognition system」内部でベイズ識別法をつかって手

形を識別する。得られたデータは手形ごとの出現頻度として正規化される。

一方、動作「move data」は開始から終了を 1 セットとして、KL 法により運動平面を求める。「HMM recognition system」はこの運動平面上に描かれた 2 次元の運動軌跡を認識し、基本動作ごとの受理確率を求め、結果を正規化する。

そして最後に、ここまでで得られた、手形データと動作データ、それに運動平面データを 1 セットとして「word pick up algorithum」に送り、辞書との比較で最終的な単語を決定する。

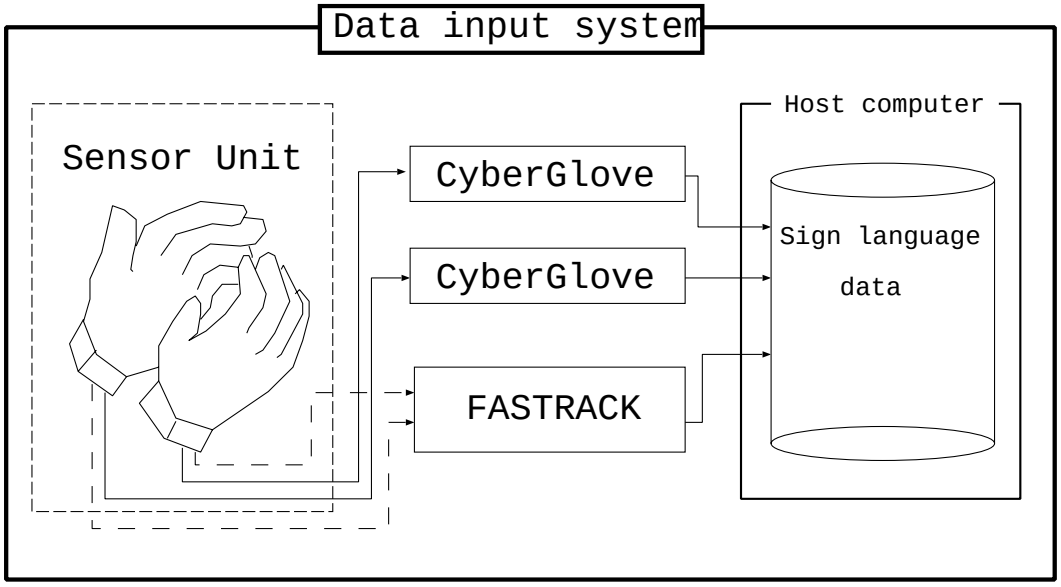


図 3.1: 手話単語データ入力システム

3.3 入力装置

本システムでは手話単語の入力装置として Virtual Technologies 社製の CyberGlove[7] と POLHEMUS 社製の FASTRACK[8] を用いている。各装置の概観を図 3.3、図 3.4 に示す。

CyberGlove はグローブの各関節部分に光ファイバを縫い込んだもので、指を曲げるとこの光ファイバも一緒に曲がる。光ファイバの曲げによって、光ファイバの屈折率は変化

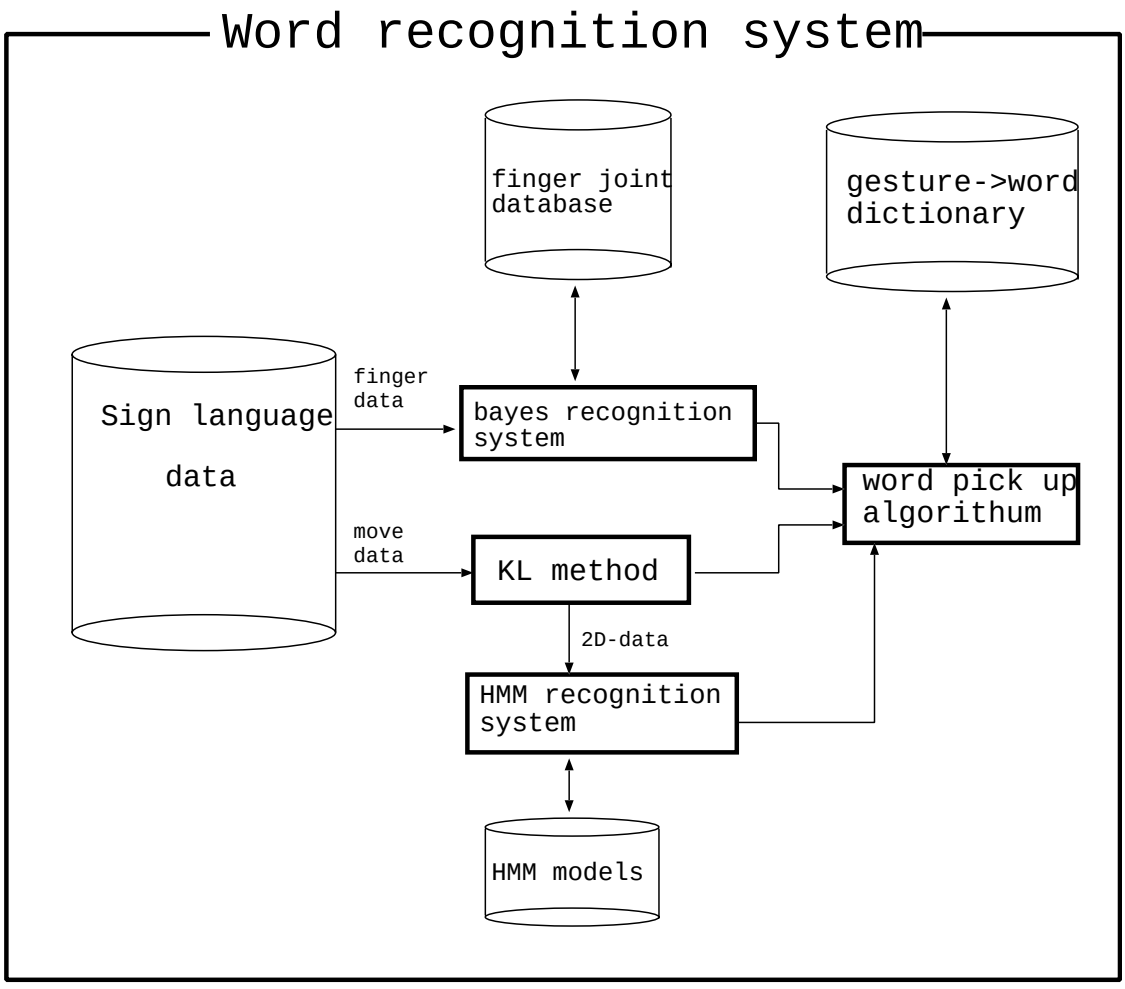


図 3.2: 認識システムモデル

図 3.3: Cyberglorve

図 3.4: FASTRACK

するので、その値を測定する事により曲げ角を検出する事が出来る。センサー自体は図3.5の位置に付けられており、それぞれ指の曲がり具合や開き具合などを検出出来るようになっている。

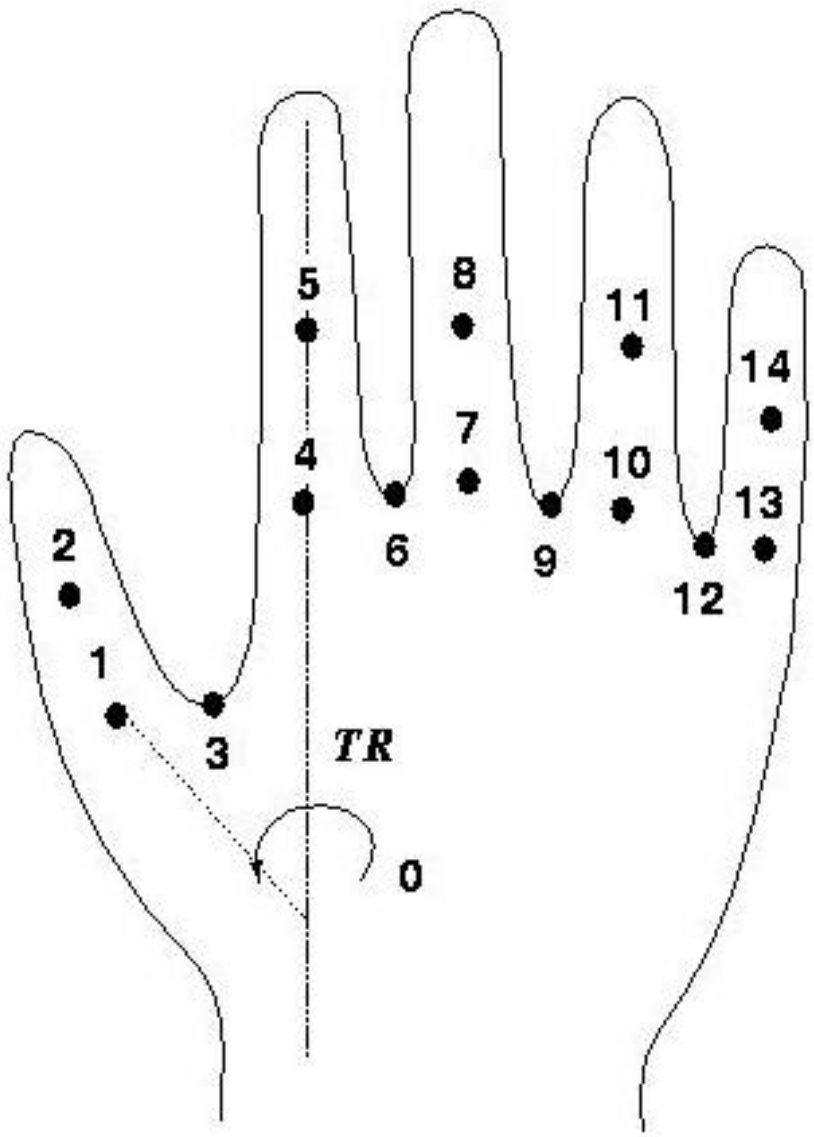


図 3.5: 手形状入力装置の関節角測定点

FASTRACKは、トランスミッタと観測基準点となるレシーバの組合せからなり、磁気によってトランスミッタの3次元位置及び姿勢を計測する事が可能である。本システム

ではトランスミッタを手首の位置に装着し、手話動作中の運動軌跡データを取ることとした。

3.4 手話単語の切り出し法

手話単語認識において、連続している手話動作の中から必要な手話単語部分を抽出する事は非常に重要であり、同時に非常に困難でもある。単語抽出を完全にすることが出来れば、手話単語認識の完全自動化に大きく近付いたと言っても良い。しかしながら実際の手話において、手話単語と手話単語間を結ぶ遷移動作の境界は非常に曖昧で容易に分離出来ないのが現状である。その為、手話単語認識システムの多くは、一つの単語を行なうごとに動作を数秒停止させたり、特定の場所に手を戻したり、手動で切り分けたりし、認識に必要な部分の抽出が容易になるよう入力データを工夫をしている。

本研究においても、認識単位として基本動作を採用している為、入力データは何らかの方法で基本動作単位に分割する必要がある。基本動作は手話単語中に概ね１つである為、手話単語の動作のみを切り出すことは、基本動作を切り出すことと等価である。しかし、単語によっては基本動作が2つ以上含まれている場合も存在する。そこで、複数の基本動作が含まれる手話単語から基本動作を自動抽出可能かどうか検討した。

3.4.1 速度による切り出し

基本動作を切り出す手法として最も有効と思われるのは速度による切り出しである。但し手話動作において、その動作が完全に停止する事は稀なので、ある閾値を設け、その閾値以下の速度が一定時間連続した場合を基本動作の切れ目とした。この方法は、一つ一つ別々の意味を持っている手話単語を２つ並べて一つの単語としてみなすタイプの手話に有効であると考えられる。

3.4.2　角度変移による切り出し

もう一つの手段として、急激な移動方向の変化を用いて基本動作を分離できないか検討した。つまり、現在の速度ベクトルと、前フレームでの速度ベクトルの内角差を角度変移として抽出し、それを用いて分離を行なう。この方法だと、往復運動を複数の直線動作として分離でき、より手話の音韻表記に近い基本動作の区分けが可能となる。しかし、角度変移のみだと停止中の微妙な前後の動きの時も間違って反応してしまう可能性がある。そこで、停止中は速度が小さい事を利用して、角度変移と速度の絶対値を掛ける事により、手話を行なっている時のみの切り出しを検討した。

3.5　切り出し性能の評価

実際にFASTRACKから得られた3次元位置データを用い、その速度変化や角度変化についてプロットした結果の例を図3.6〜3.17に示す。なお、図上の縦線は、目視による切り出し位置を重ねたものである。データをプロットした単語は、1つの単語に基本動作が複数含まれる単語であり、表3.1に示す２１単語を対象に行なった。その結果、速度による切り出しが不可能な単語が一部あるものの、目視による切り出し位置は速度が大きくなる前後に存在することから、閾値とフレームの調整によって、ある程度有効であると考えられる。一方、角度変移については図3.11,3.13の様に、基本動作の区切りと思われるフレーム前後での傾向や、さらには同一単語中での傾向なども変化の仕方が安定していなかった。そのため、一般的な切り出しルールを作成する事は困難であり、使用出来ない事が分かった。

次に、速度による切り出しで、実際に手話単語の切り出しを行ない、手動による切り出し位置と自動切り出しで切り出された位置との差を比較する。切り出した際には、図3.18の様に動作と動作を繋ぐ谷の幅が一定以上で、かつ閾値を下まわっている時に谷の中間を切断位置とした。

このルールにより切断した分割数と、目視による分割数とを比較し、その一致率を表

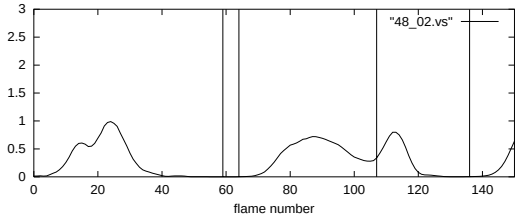


図 3.6: 医者１（速度変位）

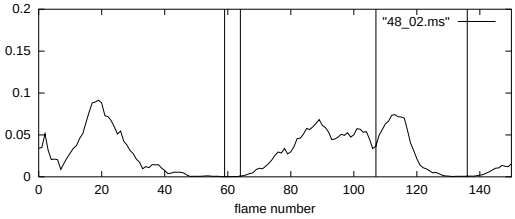


図 3.7: 医者１（角度×速度変位）

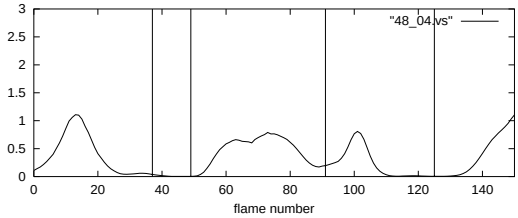


図 3.8: 医者２（速度変位）

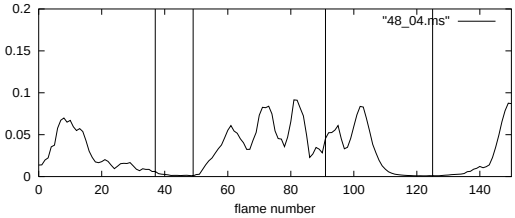


図 3.9: 医者２（角度×速度変位）

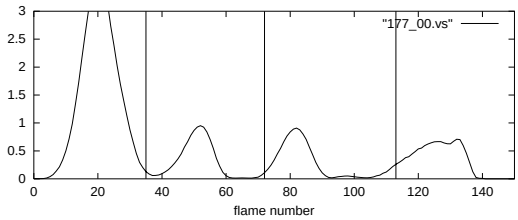


図 3.10: 北１（速度変位）

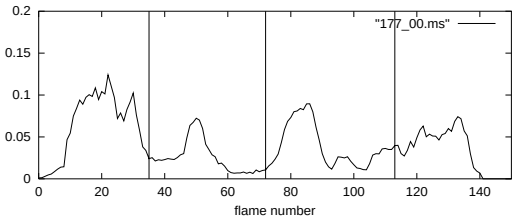


図 3.11: 北１（角度×速度変位）

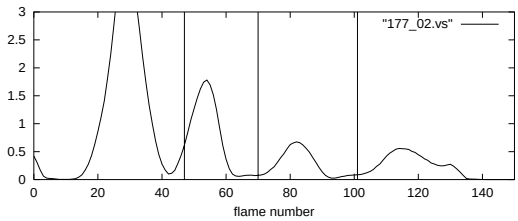


図 3.12: 北２（速度変位）

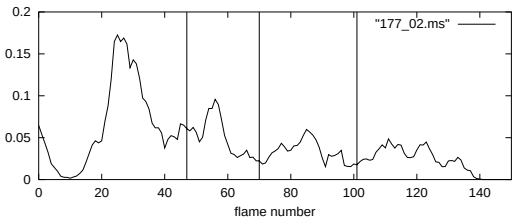


図 3.13: 北２（角度×速度変位）

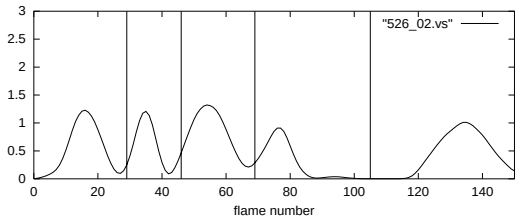


図 3.14: 部屋１（速度変位）

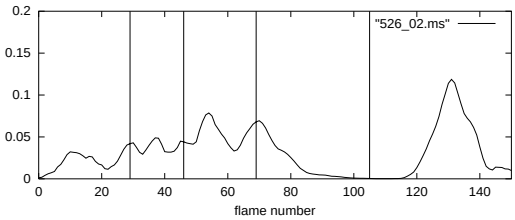


図 3.15: 部屋１（角度×速度変位）

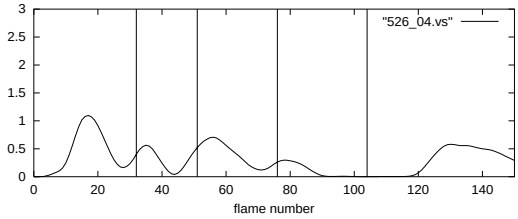


図 3.16: 部屋 2（速度変位）

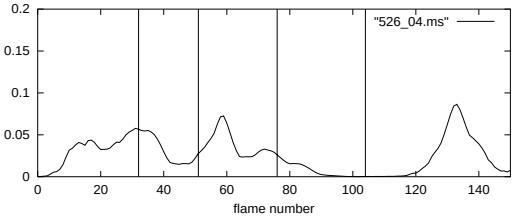


図 3.17: 部屋 2（角度×速度変位）

表 3.1: 自動切り出し対象単語

石川, 医者, 北, 決心, 健康, 恋人, 小学,
住所, すみません, 大学, 父, 中学, どこ, 何故,
日曜日, 入学, 恥ずかしい, 母, 部屋, 盲人, 両親,

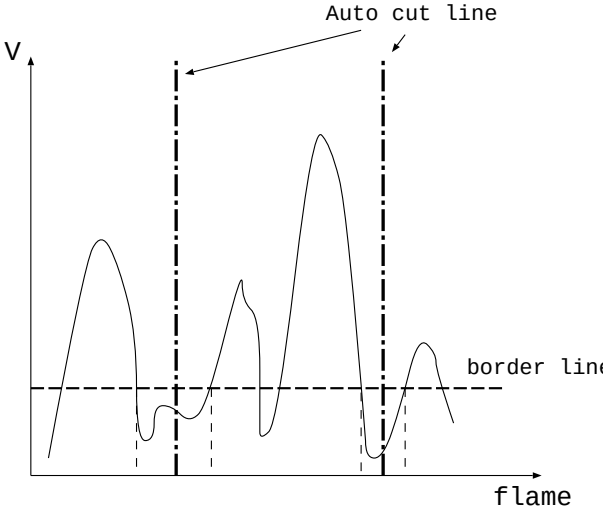


図 3.18: 自動切断例

3.2 にまとめる。

表 3.2: 分割数一致率（%）: 話者 A

速度閾値	継続フレーム数		
	f = 3	f = 4	f = 5
0.2	18.4	13.6	12.9
0.25	23.5	23.2	16.2
0.3	22.4	22.4	17.3
0.35	16.9	16.5	12.9
0.4	18.4	9.2	8.5

表 3.2 より、話者 A の場合は、速度閾値が 0.3〜0.25, 継続フレーム数が 3,4 付近が最も手動での切り分けた数と自動切り分けで切り分けられた基本動作の数が一致した。この時、切り出したフレーム位置と目視による切り出し位置との相対距離を図 3.19 に示す。

目視と自動切り分けとの相対距離５割は目視位置から１０フレーム以内と比較的近くで切り分けに成功している。一方、目視での切り分けフレームから 30 フレーム以上離れてしまったケースも４割ほど見受けられた。本システムでは毎秒 60 フレームのデータを使用しているので、時間としては 0.5 秒以上の遅れとなる。手話動作を行なっている時間は、2〜3 秒という事を考慮にいれると、これだけ離れた場合は違う位置で切り分けていると考えて良い。

以上の結果より、目視による分離と分離数が一致したのは最高でも 23.5 %であり、その内４割は正しくない位置で切り分けていると思われる。この事から、正しい位置でのきり分けを速度だけで手話単語を分離するのは困難である事が分かった。特に手の向きを変えるなどの動作をしている場合、センサーには殆んど速度ベクトルとして現れないため分離が出来ず、目視の分離数とずれるケースがかなりあった。また、継続フレーム数を小さくし過ぎると、本来一動作と考えていた円運動などを途中で分離してしまうケースなども

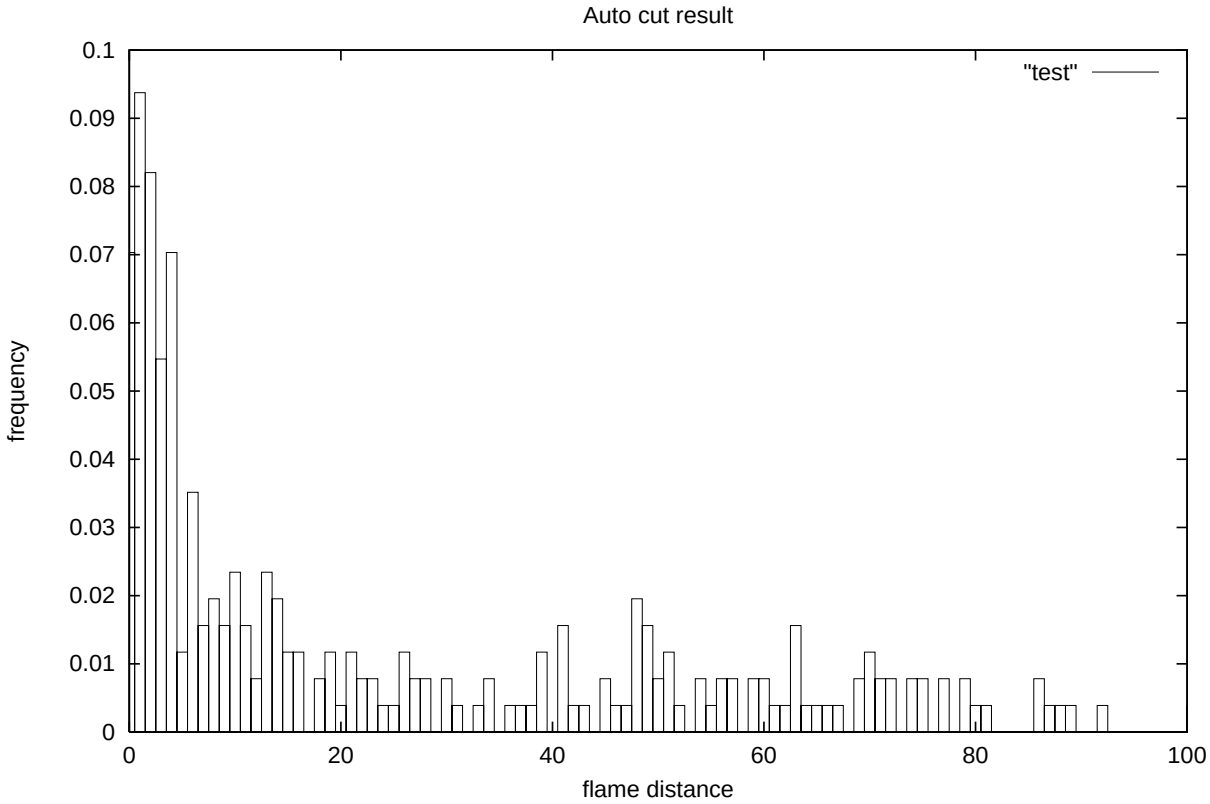


図 3.19: 自動切り出し精度

見受けられた。

3.6 基本動作の認識過程

HMM が認識するべき基本動作は、その方向や描かれている面（例えば水平面や垂直面など）には関係なく線形状のみである必要がある。したがって、従来法の様に、3 次元空間上で行なわれる手話動作のデータをそのまま認識用データとして使用することは出来ない。そこで、本研究では、3 次元空間上のデータを 2 次元平面に落す処理を行なうこととにし、その手法として KL 法を採用した。KL 法は取り扱うパターン集合に依存して特徴抽出を行ため、今回の様に一般的なルールがないパターン集合に適している。これにより HMM の認識する基本動作とは、KL 法で得られた平面に投影された軌跡形状となる。また、手話動作とは認識された基本動作と KL 法で求めた平面（以後、運動平面と呼ぶ）の組合せとなる。

3.6.1 KL 法

KL(karhunen-Loeve) 法とは、取り扱うデータに依存した形で特徴抽出が出来る手法であり、多変量解析の応用の一つである。対象パターンの集合（本研究では各フレームごとの 3 次元位置）を M 次元ベクトル $\boldsymbol{x}_n(n = 1, 2, \cdots, N)$ とし、また、求めたい正規直交ベクトル $\boldsymbol{\varphi}_k, (k = 1, 2, \cdots, K(k \leq M))$ とする。ここで、$\{\boldsymbol{x}_n\}$ に依存して $\boldsymbol{\varphi}$ を決定するのに、平均 2 乗誤差を考えると

$$J(\{\boldsymbol{\varphi}\}) = \frac{1}{N}\sum_{n=1}^{N} \parallel \boldsymbol{x}_n - \sum_{k=1}^{K} y_k^{(n)}\boldsymbol{\varphi}_k \parallel^2 \tag{3.1}$$

$$\text{ここで、} y_k^{(n)} = (\boldsymbol{x}_n, \boldsymbol{\varphi}_k) \tag{3.2}$$

式（3.2）が最小になるように $\{\boldsymbol{\varphi}_k\}$ を決めれば、その時、$\{\boldsymbol{\varphi}_k\}$ は、最もデータの分散を表現出来るベクトルとなる。式（3.2）の解法は数学上明らかにされており、$\{\boldsymbol{x}_n\}$ の共分散行列 K

$$K = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T \qquad \text{（但し、}\boldsymbol{\mu} : \{\boldsymbol{x}_n\}\text{ の平均ベクトル）} \qquad (3.3)$$

を作成して、式（eqn:k）の固有値と固有ベクトルを計算する。そして、得られた固有値を大きい順に並べ、固有値に対応したベクトルがそれぞれ主成分ベクトルとなる。

例えば図 3.20 の様に点の集合として円弧が描かれていた場合、そのデータの集合から KL 法で求められる主成分ベクトルは、第一・第二主成分ベクトルが円弧の向き（どの方向に軌跡を描くか）を表し、第三主成分ベクトルが円弧の描かれている面を表す事になる。本論文では、以後このように得られた平面を運動平面と呼ぶ。また、この運動平面に描かれている線形状が HMM により認識する基本動作となる。

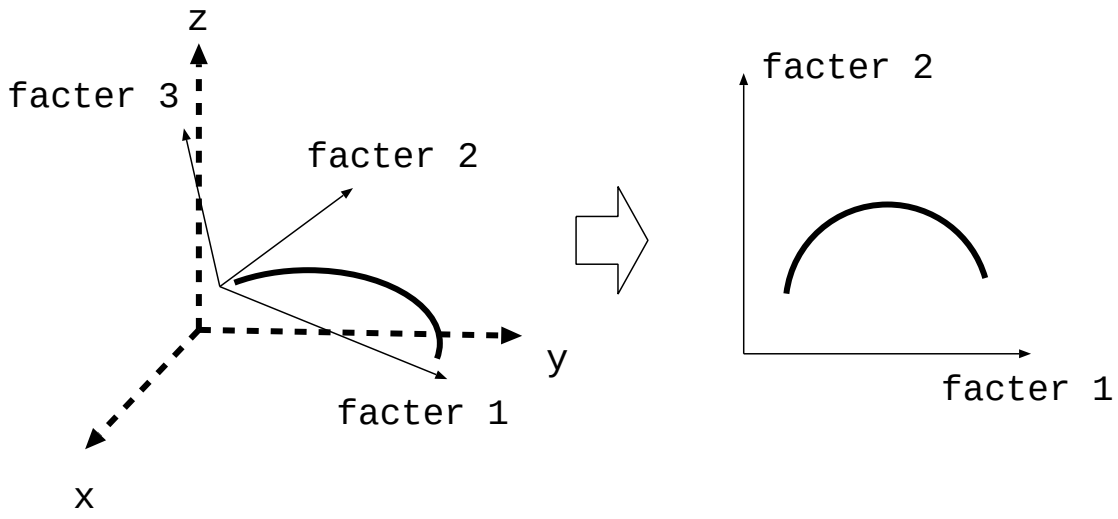


図 3.20: 平面への運動軌跡の投影

3.6.2 基本動作の選定

基本動作は図 3.20 の例でも分かる通り、線形状が一平面で完結する必要がある。また、第 2.4.2 節でも述べた様に、あまり基本動作区分を細かく分ける事は望ましくない。そこで本研究では、手話の音韻表現などを参考に基本動作としては線形状態として 7 種類を

考えた。この図形をKL法で平面に落し、さらに円などの右周り左周りなども同一の円形状として扱う為に、平面上のデータを第一象現に集めた。その結果、合計９パターンとなり、これを基本動作として仮定した。図3.21にその線形状を示す。「line」は直線動作、「half」は半円動作であり、「circle1,2」, 「dcircle1,2」はそれぞれ１重円・２重円となる。また、往復運動は「right-left move」、停止は「stop」となり、以上のカテゴリに分類されない形状は「other」とした。なお、1重円・２重円のパターンが2種類あるのは、平面に投影された図形を第一象現に集めると、2種類考えられる為である。

3.6.3 学習性能評価

HMMは学習によって統計的にデータを処理するモデルであるので、認識率は学習サンプルに依存する形となる。その為、HMMを認識に使用する前に十分に学習を行なう必要がある。そこで、本システムで認識する基本動作に必要な学習サンプル数がどの程度必要になるか実験を行なった。

データは目視で切り出した手話動作を3.6.2節で述べた基本動作9パターンに分類し、特定話者・複数話者それぞれの条件で用意した。各用意したパターンの詳細を表3.3に示す。

表 3.3: 学習サンプルデータ内訳

データ使用者	学習サンプル数（個）	テストサンプル数（個）
一名	70	10
四名	150	10

マルコフモデルの状態数は5つと6つの2種類で、その形状はベイキス型と呼ばれるものを使用した。このモデルはHMMでは最も一般的に使われているモデルで、前に戻る遷移状態がないタイプである。HMMで手話認識を行なったStarner[3]もこのモデルを採用している事から、本手法でもこのモデルを使用することにする。図3.22にその形状を示す。

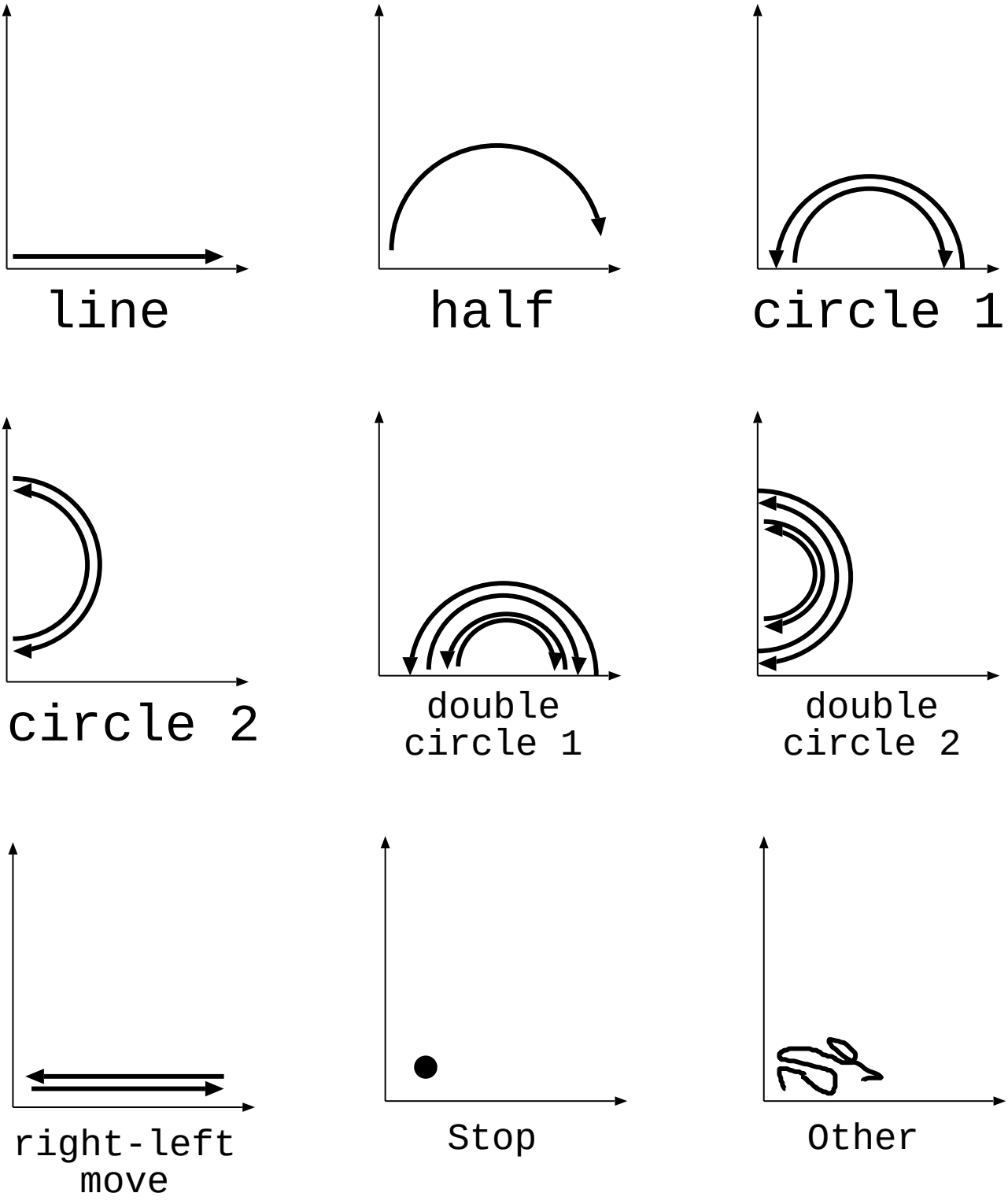

line
half
circle 1
circle 2
double
circle 1
double
circle 2
right-left
move
Stop
Other

図 3.21: 基本動作

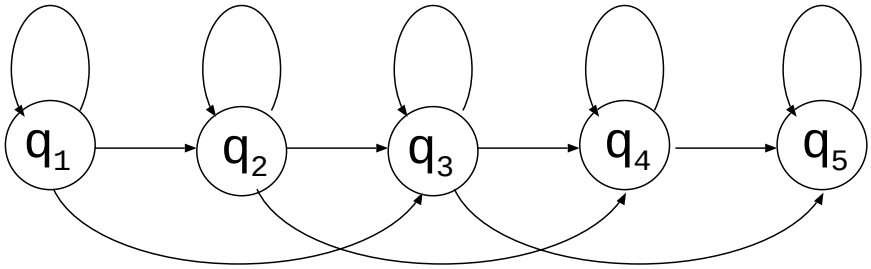


図 3.22: ベイキス型 HMM の例

認識テストは 4 回行ない、その平均値をまとめた図??〜??を本論文の章末に合わせて載せておく。実験の結果、図 3.23 の様に学習話者が一人の場合、「直線・半円・停止」はサンプル学習数が 45 前後で 70 %以上の認識率を保ち安定した。、一方、図 3.24 の様に「2 重円１」が 40 %以下と最も悪い認識率となった。また、「その他」は学習サンプル数が 50 まで進んでも比較的不安定な傾向を示している。これは基本動作として学習するパターンの幅が最も大きいものが「その他」であるために、学習収束により多くのサンプルが必要となっているのが原因と考えられる。状態数についてはあまりその差については現れなかった。ただし、認識率の悪い基本動作については若干、認識率が向上する傾向にある。認識率が悪い基本動作は複雑な線形状である事を考えると、より多くの状態数がある方が、複雑なパターンに対応出来るという HMM の特性が現れている事によると思われる。

一方、学習サンプルに複数（四人）の学習サンプルを用いた場合の学習収束には図 3.25、3.26 の例からも分かるように、一人の場合に比べかなりの学習サンプルが必要な事が分かる。また、認識率に関しても、一人に比べると悪い結果となっている。これはやはり対象を多人数にした事によるパターンのばらつきが最大の要因だと思われる。

認識率の低かった「円・2 重円」については、誤認識の原因として、円・2 重円は基本動作として 2 つのパターンを考えたが、実際のデータではこの 2 つの中間的なデータがかなり見受けられた事が最大の要因であると考えられる。事実、円・2 重円の誤認識パターンは同形状の違うパターンが多い。しかし、実際の手話では円を描いた場合、重要なのは一回の円か２回以上の円運動かという違いは重要であるが、その回転方向は人それぞれで

あり特に回転方向が指定される手話は無い。そこで、円・2 重円に関しては形状が同じ場合の合計認識率をまとめてみた。その結果を図 6.19、6.20 に示す。

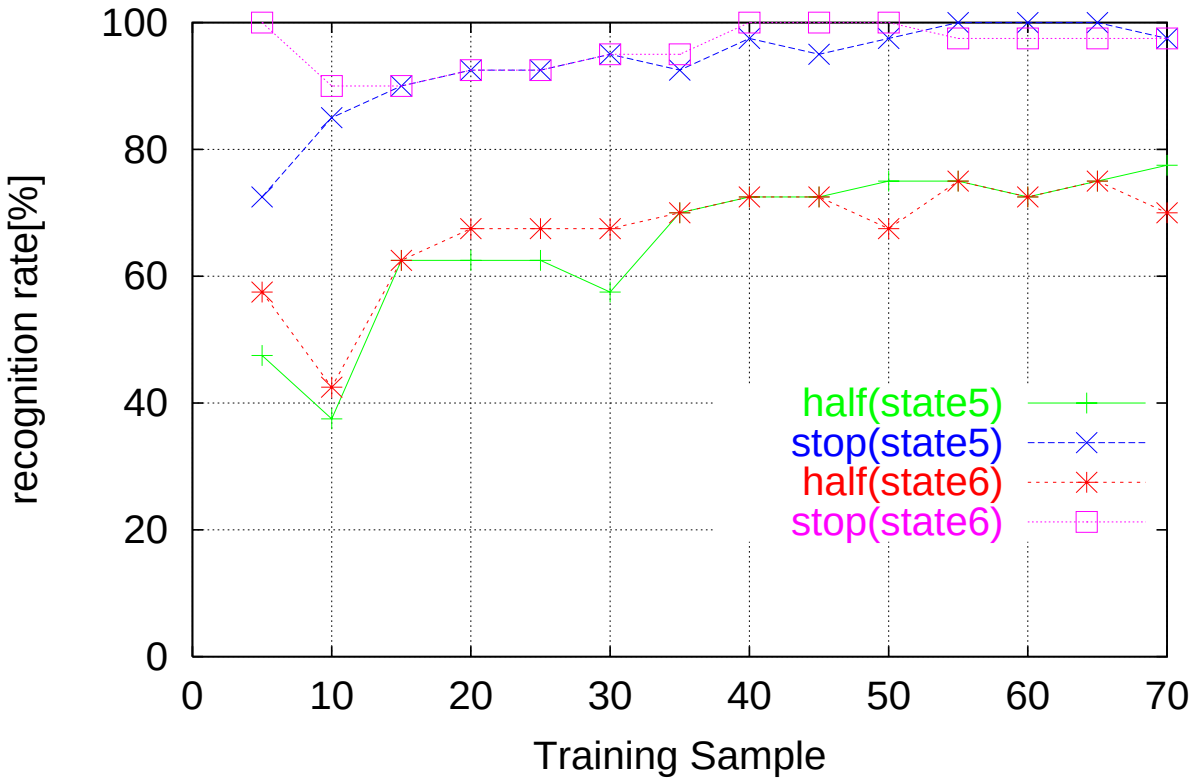


図 3.23: 学習話者一名での収束例（収束終了）

3.7 ベイズ法を用いた手形認識

ベイズ法による手形認識は、後藤 [9] の行なった実験システムをそのまま採用した。また、比較に使用する辞書も同じものを使っている。なお、この辞書は 10 人の人間から採取した 50 音のパターン組 99 個のデータを平均したものである。このシステムに手話動作中の手形を全て認識させ、その結果を手形で振り分け、正規化する事により手話動作中の手形変化を表すことにした。その流れを図 3.27 に示す。まず、切り出した手話動作中の

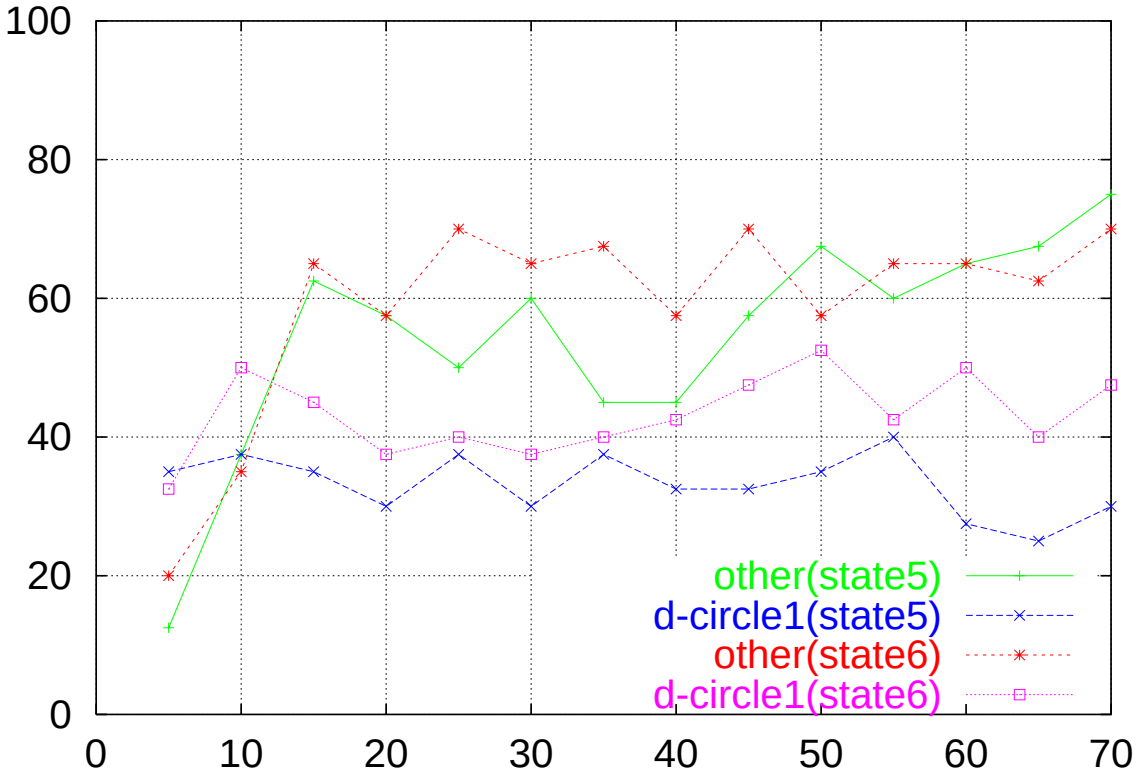


図 3.24: 学習話者一名での収束例（未収束）

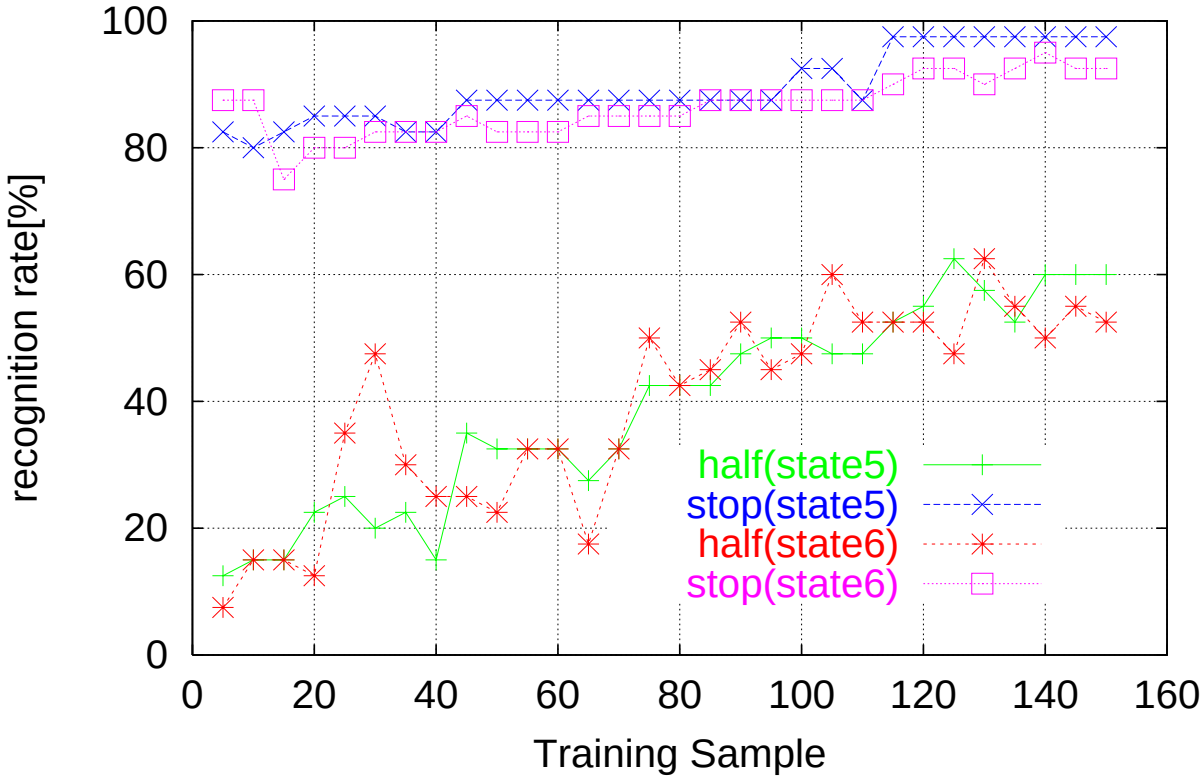


図 3.25: 学習話者四名での収束例（収束終了）

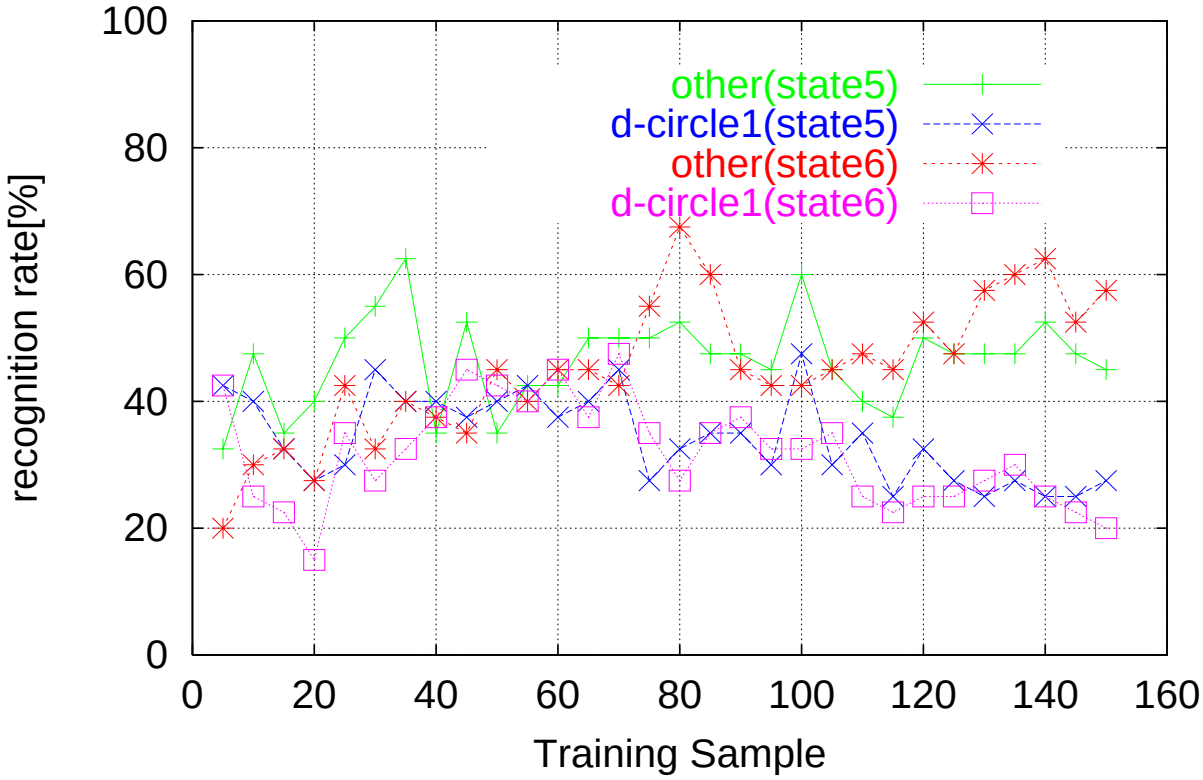


図 3.26: 学習話者四名での収束例（未収束）

データがフレーム毎に「byes recognition」に入力去れる。ここでベイズ認識法により、予め作成しておいた手型辞書と入力データが比較され、最も近い手型が選択される。選択される手型の種類は２６種類でその形状は図2.3の四角で囲まれた手形となる。選択された手型は「Hand shape」内部の各手型にカウントされていく。手形状データはこのカウントと全フレーム数で正規化したものとなる。

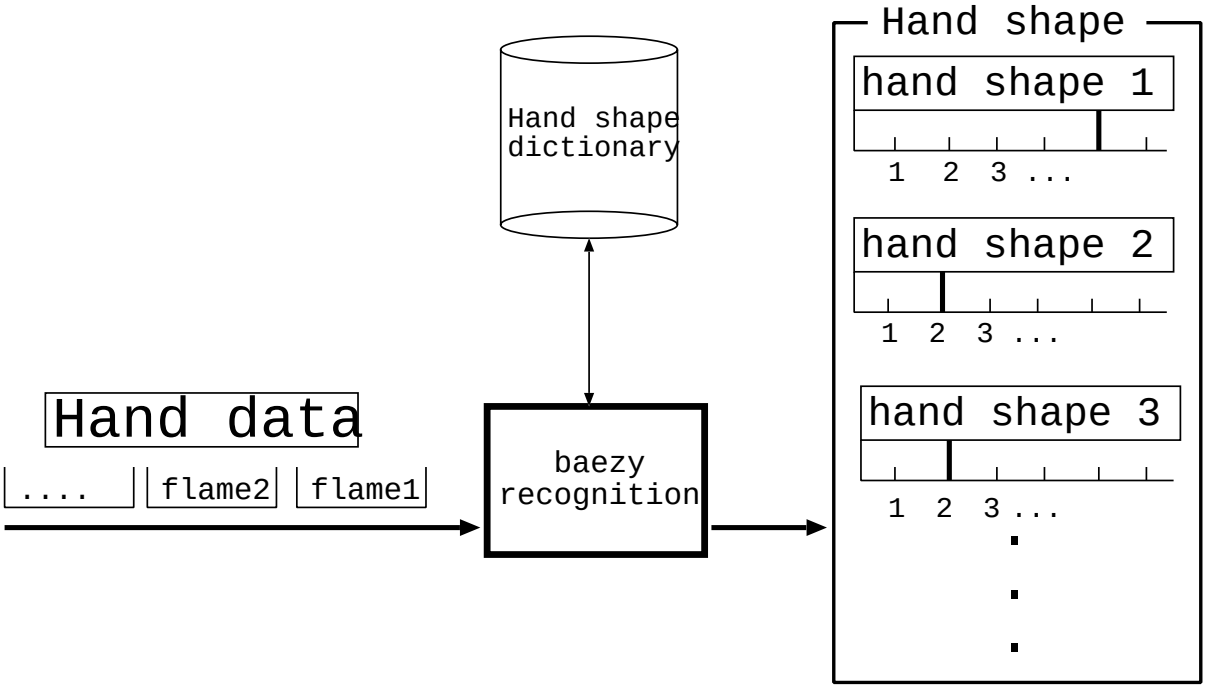


図 3.27: 手形識別処理

3.8 手話単語辞書の構成

認識できた手形・基本動作のデータより最終的に単語を決定する為には、予め単語と認識した動作との対応づけをおこなってある辞書が必要になる。この辞書と、入力された未知データの比較によって、最終的に手話単語が決定される。認識過程で得られるデータは手形・基本動作・運動平面の3種類であるので、辞書データもこれに対応する形となる。具体的なデータの構成としては図3.28の様に、単語・手形・動作・運動平面の順にデータ

が並んでいるものを作成する。手形は対象データ全フレームにおける各手形の出現割合、動作は対象データを各基本動作と仮定した時の HMM の受理確率を正規化したものである。また運動平面は KL 法で求めた主成分ベクトルとなる。辞書データ全ての認識対象単語について持っており、このデータと未知の手話単語データとを比較して、最も近い候補を認識単語とする。なお、ここでの最も近い候補とは、手型・基本動作・運動平面の差を比較し、スコアとして計算した場合にその値が最も小さいものの事を指す。このスコアの計算方法は次の 3.9 節で述べる。

Word	Hand shape data						Basic move			Active plane		
	1	2	3	....	25	26	line	half	....	facter1	facter2	facter3

図 3.28: 辞書構成詳細

3.9 単語認識アルゴリズム

本システムにおける単語認識手順を図??に示す。認識の手順としてはまず、入力された未知パターンと辞書に登録されている単語全てに対し、手形・基本動作それぞれ別々に構成データの分布差を式（3.4）（3.5）をもちいて計算する。その後、手形・基本動作それぞれのスコアに重みを掛けた結果が小さい順に幾つか候補を抜きだした上で、さらに候補の運動平面と未知データの運動平面の内角差を式（3.6）で比較し、最も差の少ない候補を最終的な認識結果とする。ここで手形・基本動作を優先して抜きだした後に運動平面の比較をするのは、運動平面は手形や基本動作と違って、特になにも処理を行なっていない為である。特になにも処理しないという事は、データに個人差やその時の状況が大きく影響してしまう可能性があるという事を意味している。その為、手形や基本動作と同列に認識すると認識結果に余計な影響を及ぼしやすくなると考えた為である。

$$scoa_H = W_h \cdot \sum_{i=1}^{26} \mid Hand_i^D - Hand^I \mid \tag{3.4}$$

$$scoa_M = W_m \cdot \sum_{j=1}^{6} \mid Move_i^D - Move^D \mid \tag{3.5}$$

$$wscoa = scoa_H + scoa_M + W_v \cdot \sum \frac{V_i^D \cdot V_i^I}{\mid V_i^D \mid \cdot \mid V_i^I \mid} \tag{3.6}$$

なお W_h, W_m, W_v は各スコアにかかる重み、$Hand_i^D, Move_i^D$ は辞書に登録されている単語の手形及び基本動作の分布データ,$Hand^I, Move^I$ は未知データの分布データである。$V_i^D, V_i^I (i = 1, 2, 3)$ が辞書・未知データの主成分ベクトルである。

3.9.1 構築辞書データの内訳

実験を行なう前に先だって、認識用辞書を作成しておいた。辞書は、予め単語が分かっているデータを認識システムに通し、その結果、出てきたデータを保存しているものである。今回実験用に作成した辞書データは、一単語につき 6 パターンの認識結果を平均したものとした。辞書の作成に用いたパターンのデータ内訳を表 3.4 に示す。

表 3.4: 辞書作成用サンプルデータ内訳

辞書名	サンプル採取者	使用サンプル数
dictionary1	話者 A	8 パターン
dictionary2	話者 A	2 パターン
	話者 B	2 パターン
	話者 C	2 パターン
	話者 D	2 パターン

第 4 章

手話単語の認識実験と評価

4.1 はじめに

本章では 3.9.1 節において作成しておいたた辞書を用い、手形や基本動作、主成分ベクトルへの重みづけの変化によってどの様に認識率が変化するか検討を行なった。次に、辞書構成時の使用話者が認識率に及ぼす影響を調べた。

4.2 認識対象単語

認識対象となる単語は電子辞書「ムサシ」[10] を参考に、任意に選ばれた 280 単語とした。その単語一覧を表 4.1 に示す。

4.3 手話単語認識実験

表 4.1 の単語データは膝に両手を置いた時点からデータが始まり、一つの手話動作を行なった後、再度膝に両手を戻すまでが一つのデータとなっている。本実験ではこのうち、手話動作を行なっている部分のみを目視により抽出して使用した。

表 4.1: 対象単語

挨拶, 会う, 赤, 明るい, 秋, 朝, 浅い, 明後日, 明日, 遊ぶ, 暖かい, 頭, 新しい, 熱い, 集まる, あなた, 兄, 姉, 危ない, ありがとう, ある, 歩く, 安心, 言う, 家, 以下, 怒る, 以外, 生きる, 行く, 幾つ, 石川, 医者, 椅子, 忙しい, 痛い, 一日, 一年, 一番, 一緒, 一般, 意味, 妹, いろいろ, 上, 嘘, 美しい, 旨い, 生まれる, 裏, 売る, 選ぶ, 多い, 大きい, 教える, 遅い, 教わる, 夫, 弟, 男, 一昨日, 大人, 同じ, 覚える, おめでとう, 重い, 思う, 面白い, 表, 女, 会社, 買う, 顔, 書く, 過去, 貸す, 家族, 固い, 悲しい, 金, 通う, 借りる, 軽い, 可愛い, 変る, 間, 考える, 関係, 簡単, 学校, 頑張る, 北, 昨日, 決める, 今日, 兄弟, 嫌い, 疑問, 臭い, 曇り, 悔しい, 暮す, 比べる, 来る, 苦しい, 車, 黒, 計算, 結婚, 決心, 健康, 現在, 恋人, 答え, 断る, 子供, 細かい, 困る, 最高, 最後, 最初, 探す, 淋しい, 寒い, さようなら, 賛成, 残念, しかし, 試験, 仕事, 自然に, 下, しっかり, 姉妹, 趣味, 手話, 障害者, 小学, 勝負, 昭和, 調べる, 白, 信じる, 自慢, 住所, 自由, 上手, 好き, 過ぎる, 少し, 捨てる, 全て, すみません, する, 座る, 生活, 相談, 卒業, 空, 大切, 高い, 立つ, 例えば, 楽しい, 食べる, 大学, 大丈夫, 騙される, 騙す, だめ, 誰, だんだん, 小さい, 近い, 違う, 父, 中学, 長, 通訳, 使う, 月, 次, 机, 作る, 都合, 続く, 妻, 強い, 適当, テレビ, 天気, 電車, 電話, 東京, 遠い, 時, 得意, 友達, 取る, どこ, どちら, 無い, 中, 長い, 泣く, なぜ, 懐かしい, 何, 名前, 苦手, 西, 日曜日, 日本, 入学, 人気, 盗む, 願う, 眠い, 眠る, 寝る, 年齢, 農業, 飲む, 入る, 始める, 恥ずかしい, 話合い, 母, 速い, 春, 晴れ, 反対, 場所, 火, 東, 引く, 飛行機, 筆談, 必要, 人, 人々, 暇, 開く, 平等, 深い, 不思議, 不満, 冬, 古い, 文化, 下手, 部屋, 返事, ほとんど, 本, 本当, 毎日, ますます, まずい, 貧しい, まだ, 町, 間違い, 待つ, まで, 短い, 水, 道, 南, 未来, 見る, 息子, 娘, 難しい, 無駄, 無理, 明治, 迷惑, 珍しい, 盲人, 目的, もし, もっと, もらう, 森, 約束, 安い, 休み, 破る, 柔らかい, 指文字, 良い, 用事, 読む, 夜, 離婚, 両親, 料理, 恋愛, 聾唖, 老人, 若い, わからない, わかる, 別れる, 分ける, 忘れる, 私, 悪い,

表 4.2: 特定話者実験条件

被験者	話者 A
使用 HMM	ベイキス型（状態数６）
HMM 学習サンプル採取者	話者 A
辞書データ作成話者	話者 A
テストデータ	話者 A

4.3.1 特定話者

まず、特定の被験者（話者 A）について実験を行なった。使用した HMM は 3.6.4 節で調べた状態数６のモデルを使い、予め 70 パターンの学習サンプルにより学習を行なった HMM を使用した。辞書は被験者（話者 A）の手話 280 単語データセット 8 パターンの結果を各単語につき平均した dictionary1 を採用し、テストパターンは辞書作成に使用していないパターンを用いた。表 4.2 に実験条件をまとめて示す。

要素別認識結果

本システムは手形と基本動作の違いをスコアとして計算し、その上位幾つかで運動平面との照合を行なっている。よって手形や基本動作の認識率が悪い場合は、運動平面での照合前に候補から外れてしまう。そこでまず始めに、作成した辞書において手形・基本動作・運動平面それぞれが、単体でどの程度、辞書データと一致しているか検討する。具体的には各要素だけでスコアを計算し、目的の単語が上位何番目に入るかを調べる。データは辞書作成に使用したデータと未使用データを使用した。その結果を図 4.1、4.2 に示す。

図 4.1、4.2 より、辞書作成に用いたデータでは上位 30 位以内に手形・基本動作が含まれる割合は 100 %となった。しかし、運動平面に関しては上位 30 位以内に含まれる割合はやや悪く 85 %となった。これは、運動平面のデータ自体にかなりばらつきがあったため、平均を取った場合にデータが鈍ってしまった結果と考えられる。一方、辞書に使用し

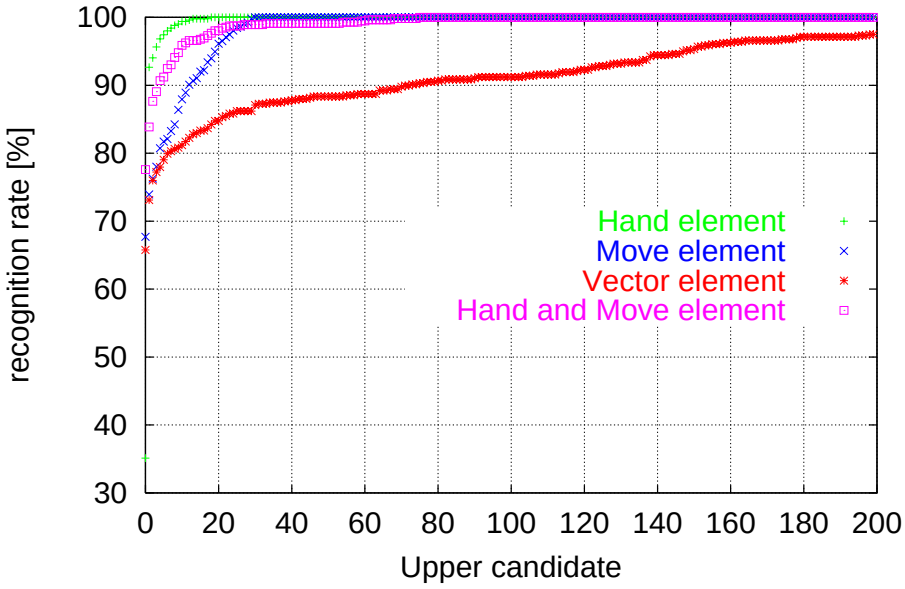


図 4.1: 要素別認識結果（辞書使用済み）

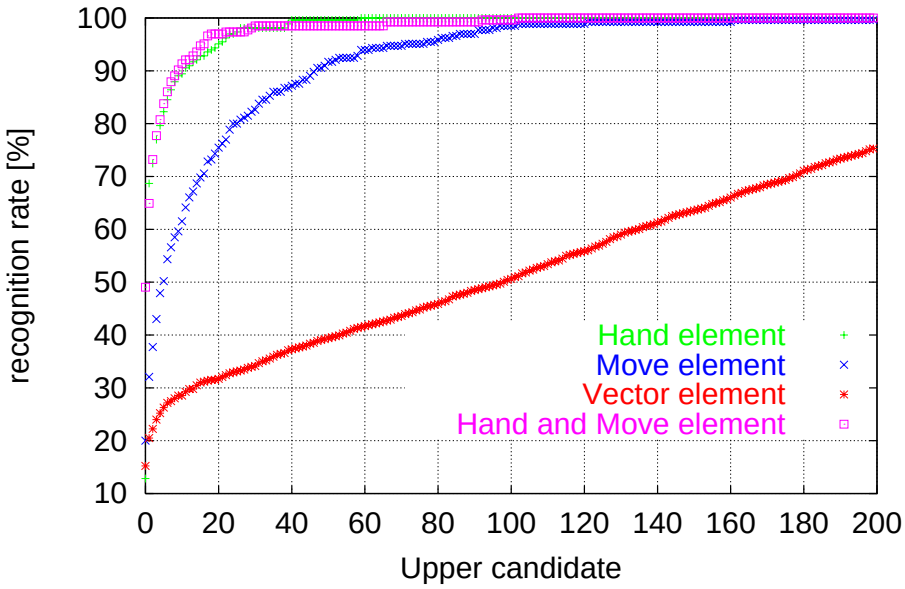


図 4.2: 要素別認識結果（辞書未使用）

ていないデータはやはり手形・基本動作共に、認識率がやや低下し、上位 30 位以内では手形が 98 %、基本動作が 82 %となった。また、手形・基本動作を組み合わせた場合は、手形や基本動作単体よりも上位では良い認識結果が得られた。この事から、手形・基本動作双方を組み合わせる事により、お互いの誤認識をある程度、補間出来る事が分かる。運動平面については、その傾きが非常になだらかになってしまった事より、運動平面と単語間にはあまり強い相関関係が存在していないと思われる。これは、単語認識における運動平面の情報の重みはあまり無い事を示しており、単語認識工程において本手法の様に運動平面を最後に持ってきたは処置は妥当であると言える。

次に、未知単語における辞書の認識率、および、各要素の重みを変化させた場合、それが認識率にどの様な影響を及ぼすかを調べた結果を表 4.3〜4.5 に示す。

表 4.3: 特定話者認識実験

使用データ	第一位認識率%	第二位認識率%	第三位認識率%
辞書作成使用データ	82.1	87.5	90.2
辞書作成未使用データ	61.8	73.2	76.4

次に、手形と主成分ベクトルの重みを固定した状態で、基本動作のみの重みを変えた時、認識率に与える影響を表 4.4 に示す。なお、運動平面との比較は手形・基本動作のスコアが良かった上位 30 単語で行ない、この単語から目的の単語が外れた場合はミスとしてカウントした。

続いて、表 4.5 の中で認識率の良かった重み比 1:1:1 のケースについて、主成分ベクトルの重みを変化させる事による認識率への影響を調べた。

表 4.5 から、手形・基本動作・主成分ベクトルの重みについては 1:1:1.5 の時が最も認識率が良く、第一位認識率で 62.8 %を得られた。表 4.4、4.5 より、やはり手形と基本動作の比率を変えると認識率に大きな影響を及ぼすが、運動平面の重み変化は認識率にあまり影響を及ぼさない事が分かる。これは要素別の結果（図 4.1,4.2 とも一致する。

表 4.4: 特定話者認識実験 (辞書作成未使用)

重み比 （手型・基本動作・運動平面）	第一位認識率%	第二位認識率%	第三位認識率%	ミス率%
1:0.75:1	46.4	61.4	69.3	1.1
1:1:1	61.8	73.2	76.4	
1:1.5:1	60.3	72.9	77.1	1.8
1:2:1	59.3	73.2	77.5	1.8

表 4.5: ベクトル重みによる認識率の変化

手形・基本動作・ベクトル比	第一位認識率%	第二位認識率%	第三位認識率%
1:1:0.75	62.1	72.9	76.4
1:1:1	61.8	73.2	76.4
1:1:1.5	62.8	72.5	77.9
1:1:2	62.5	72.1	76.4

4.3.2 不特定話者

不特定話者での認識実験に使った辞書の構成を表 4.6 に示す。この辞書を用い、話者 A,B,C,D,E について各要素別の認識率及び、単語認識率について調べたを図 4.3〜図 4.5 と表 4.7〜表 4.9 にまとめる。

表 4.6: 不特定話者認識用辞書一覧

辞書名前	HMM 学習データ採取者	辞書データ採取者	採取パターン数
辞書 1	話者 A	話者 A	8 パターン
辞書 2	話者 A,B,C,D	話者 A	8 パターン
辞書 3	話者 A,B,C,D	話者 A,B,C,D	各人 2 パターン

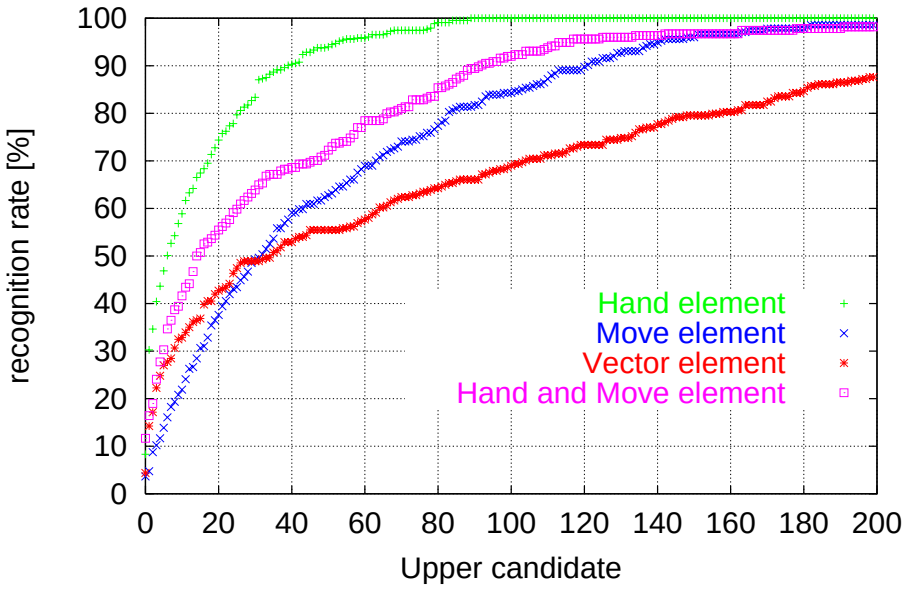


図 4.3: 要素別認識結果（辞書１：学習・登録同一人物）

表 4.7,4.8,4.9 より、学習が収束した HMM を使用した場合、認識率は辞書作成に使用した人数によって変動はするものの、HMM の学習に使用したサンプルの人数にはあまり影

表 4.7: 辞書１（学習・登録、同一話者）を用いた場合の認識率

テストデータ	第一位認識率%	第二位認識率%	第三位認識率%
辞書作成データ	70.7	80.9	85.5
話者 A	61.8	73.2	76.4
話者 B	17.8	26.3	29.4
話者 C	13.6	22.1	28.9
話者 D	12.9	19.6	22.9
話者 E	12.1	17.9	22.9

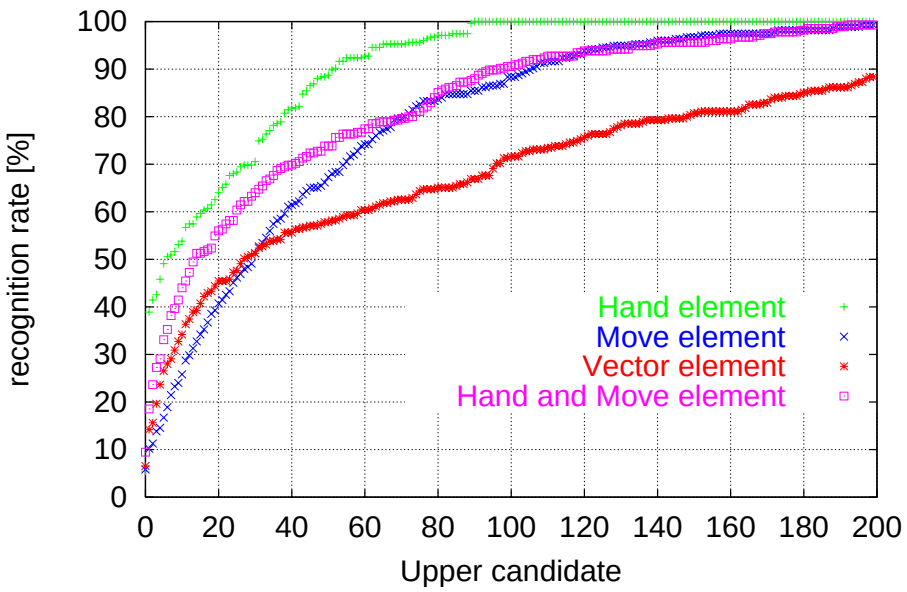


図 4.4: 要素別認識結果（辞書２：学習複数・登録一人）

表 4.8: 辞書 2（学習複数・登録同一話者）を用いた場合の認識率

テストデータ	第一位認識率%	第二位認識率%	第三位認識率%
辞書作成データ	82.3	87.0	90.0
話者 A	50.7	63.2	71.1
話者 B	13.6	24.6	31.8
話者 C	12.9	20.7	25.4
話者 D	11.4	17.1	21.1
話者 E	12.5	17.5	20.3

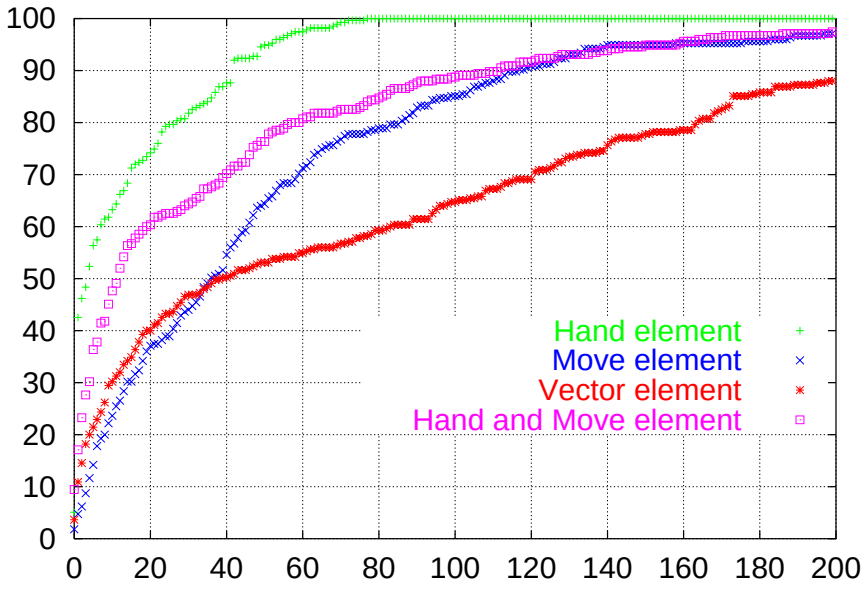


図 4.5: 要素別認識結果（辞書 3 ：学習・登録複数人）

表 4.9: 辞書１（学習・登録、複数話者）を用いた場合の認識率

テストデータ	第一位認識率%	第二位認識率%	第三位認識率%
辞書作成データ	14.5	22.3	26.3
話者 A	12.1	18.9	26.1
話者 B	18.6	29.3	33.6
話者 C	17.5	24.3	31.1
話者 D	17.5	24.3	31.1
話者 E	19.6	27.5	32.1

響を受けない事が分かる。これはHMM が学習している基本動作は、特定の単語の動きではなく、幾つかの違う単語の動きを同一動作として学習しているため、もともと学習の段階である程度分布の広いデータが用意されるために、個人・複数で特に大きなパターンのばらつきによる違いが無かった事が原因と考えられる。この結果は図 4.3,4.4 の要素別の結果において、基本動作のグラフが辞書１、辞書２双方ともあまり変わらない事からも裏付けられる。

一方、辞書作成時に使用した人数を増やした場合、図 4.5, や表 4.9 から分かる様に、話者 B,C,D,E ではやや認識率が向上したものの、話者 A は非常に認識率が低下するという結果が得られた。これは、手形・基本動作で認識された手話単語の動作データが、個人の動きに強く影響を受けている事を表している。

4.3.3 従来法との比較

特定話者での認識結果だが、最も良い認識率の場合でも 62.8 %であり、本実験より多くの単語で認識実験をおこなった DP マッチングの 98.7 %や、FFT を用いた認識実験での認識結果である 83.4 %に比べると良いとは言えない結果になった。

この結果の原因だが、一つの原因として考えられるのは学習による動作認識の鈍りであ

る。これは個性を吸収する為には必要なものでもあるが、今回の様に複数の単語を同じ基本動作でくくる場合、どうしても各基本動作にはっきり分離できないデータも数多く出てくる。この様なデータを含む学習は HMM での動作の分類を曖昧にしてしまい、結果として多数の単語分離が困難になったと考えられる。

次に、不特定話者での場合だが、全くの未知話者での認識実験ではその認識率が 12.1 %と非常に悪かった。しかし、辞書構成や学習パターンに用いる人数を変えた実験においては、最高で 19.1 %とやや向上した。これは、複数人で辞書を構成したために、辞書の個人化が緩和された事によるものだと思われる。一方、この条件においては、逆に認識率が非常に悪くなる話者もおり、辞書作成に人数を増やしても、かならずしも良い結果が得られるとは限らない事が分かった。

4.4 まとめ

本章では手形・基本動作・運動平面の 3 要素で構成されたデータを用い、本論文で提案したシステムの手話単語認識率を調べた。その結果、特定話者認識では 62.8 %、また未知話者での認識では 12.1 %という結果となった。認識率が低かった原因としては、HMM の認識率の悪さや、手形状や運動平面のばらつきなどが考えられる。また、辞書構成を変えた場合の実験においては、多くの話者では認識率が多少向上したものの、中には非常に悪くなってしまう話者もいた。この事より、辞書データに強い個人の癖が出ている事が分かる。つまり、現状の本システムでは不特定話者での認識には向いていない事になる。この様な問題を解決するためには、HMM による基本動作の分類をより細かくし、HMM の誤認識を抑えると共に、手形や運動平面に関して、個人の癖を受けにくい様な手法を考案する必要がある。

第 5 章

結論

社会福祉への関心の高まりと共に、障害を持つ人々もさまざまな場面で活躍する様になった。特に昨今、手話はテレビにおいて手話ニュースとして放映されるなど、昔に比べてその存在が広く知れ渡る様になった。これによって、ひと昔前にあった偏見などは随分減っている。また、手話を習おうという健聴者も増えてきている。しかし、やはり手話は一般の人にはあまり縁が無く、聾唖者が社会生活をする上でコミュニケーションに苦労する場面は多い。現状では、この問題に対し筆談や手話通訳士に頼らざるえないが、どちらもその利便性という点では問題が多い。

そこで本研究では、従来から課題となっていた不特定話者に対応した手話単語認識システムを目指し、HMM を用いた手話単語システムの構築と検証を行なった。

まず、手話の音韻表記を参考に手話動作を、手形・基本動作・運動平面の 3 要素に分け、基本動作をの分類を検討した。次に、HMM が認識する基本動作について、その学習サンプル数の違いにより認識率がどの様に変化するか調べた。その結果、特定話者では基本動作は 60 パターン辺りで、一部のパターンを除きほぼ安定してくる事が分かった。

次に、学習した HMM とベイズ法識別法による手形認識結果を用いて、手話単語認識システムの構築と実験を行なった。認識システムが識別する単語は 280 単語とし、手形と基本動作が似ている上位候補何個かで運動平面を比較し、最終的な単語を選択するというア

ルゴリズムで単語の認識を行なった。結果としては、特定話者認識では最高で 61.8 %という認識結果になった。また全くの未知話者での認識結果は 12.1 %という結果となった。認識結果の悪さについては、さまざまな要因が考えられるが、特に HMM の分離範囲が大まかすぎ、多くの手話動作の分類には不向きであった点が挙げられる。また、手形や運動平面は個人差や状況によってばらつきが大きく、そのまま平均データを用いると、不特定話者での認識は困難である事が分かった。

5.1 今後の課題

本研究では、手話単語を手形・基本動作・運動平面に分けて認識を行なったが、あまり良い結果は得られなかった。原因としては、HMM の識別動作分類や、単語認識アルゴリズムの不備、手形や運動平面における個人差の考慮不足などが考えられる。これを解決する為には、より細かな HMM の分類を検討や、手話動作中における手形変動の癖などを考慮する必要がある。また、作成した辞書から単語を認識するアルゴリズムに関しても、パターン同士の違いを考慮した認識方法を検討する事が重要である。

謝辞

本研究を進めるにあたり、熱心な御指導と御鞭撻を賜わりました、北陸先端科学技術大学院大学 情報科学研究科 口 進 教授, 阿部 亨 助教授にここで深く感謝の気持ちを現します。

副テーマで御指導頂きました赤木 正人教授に感謝致します。

日頃より有意義な教示と議論を頂きました, 山森一人助手, 井口 寧 助手, 林 亮子 助手に感謝致します.

研究を進める上で、お世話になりました下平 博助教授に厚くお礼申し上げます。

サブテーマと研究を進める上で、さまざまなアドバイスを頂きました水町 光徳さんに感謝いたします。

日頃より、お世話になった堀口・阿部研究室の皆様に厚くお礼申し上げます。

参考文献

[1] 佐川、酒匂、大平、崎山、阿部：“圧縮連続 DP 照合を用いた手話認識方式”, 電子情報通信学会論文誌, Vol.J77-D-II, No.4, pp.753-763, Apr.1994.

[2] 鈴木 信勝、堀口 進 : “手指運動軌跡のコード化による手話単語認識に関する研究” 北陸先端科学技術大学院大学 修士論文 1999

[3] Thad Eugene Starner “Visual Recognition of American Sign Language Using Hidden Markov Models” Massachusetts Institute of Technology, Cambridge MA, June 1991

[4] 神田和幸, 中博一：“日本手話の音韻表記法”, 日本手話学会, 手話学研究 12(1991),31-39.

[5] 武士 展照 春山 智 小林 哲則: “HMM を用いた手振り認識” 電気情報通信学会 信学技報 (1996-05)

[6] 神田和幸：“手話学講義”, 日本福村出版,1994.

[7] “*Cyber*Glove™ User's manual”, Virtual Technologies, 1993.

[8] “3SPACE USER'S MANUAL”, POLHEMUS, 1993.

[9] 後藤 岳志, 堀口 進 : “動作を伴う指文字を含む連続指文字認識”, 電気関係学会北陸支部連合大会 F-54,pp396.1996.

[10] 神田和幸：“ムサシ α 日本手話電子辞書”, アルファメディア, 1995.

[11] 古井 “音響・音声工学” 近代科学社 1992.

[12] 舟久保 登 “パターン認識” 共立出版株式会社 1991.

第 6 章

付録

6.1 特定話者学習収束実験結果

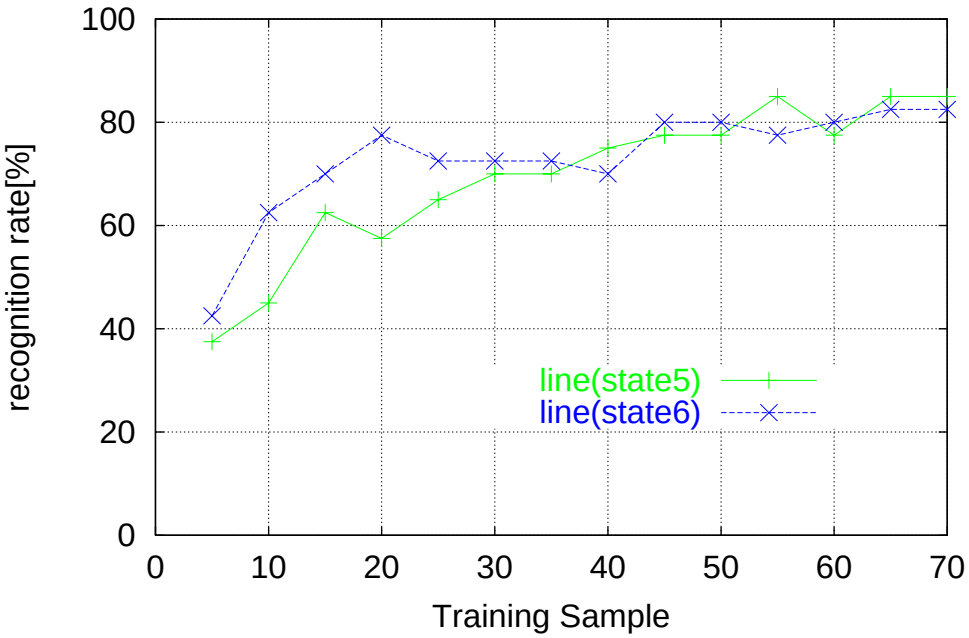


図 6.1: 直線動作の学習サンプル数と認識率の関係

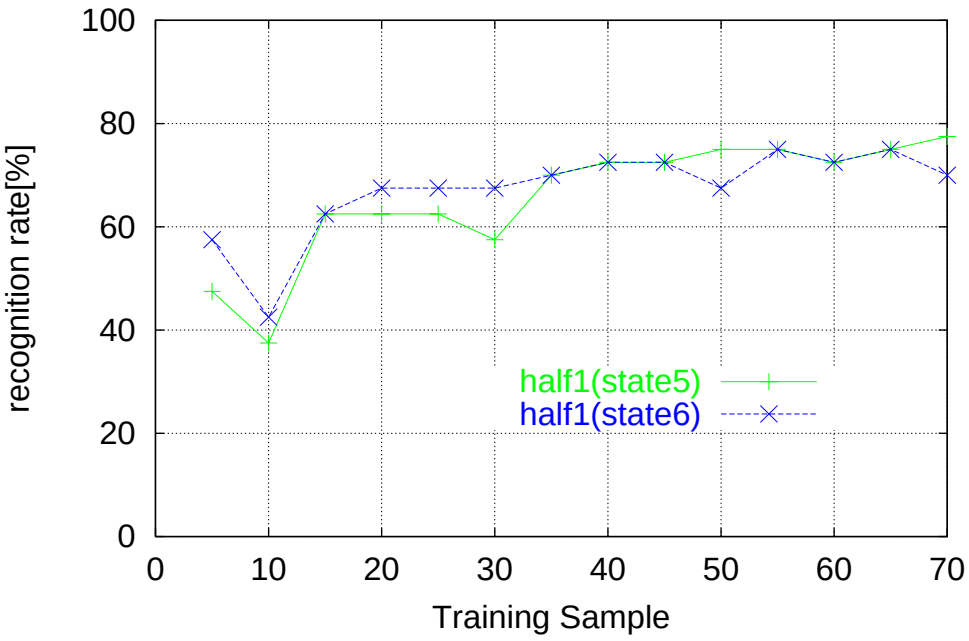


図 6.2: 半円動作の学習サンプル数と認識率の関係

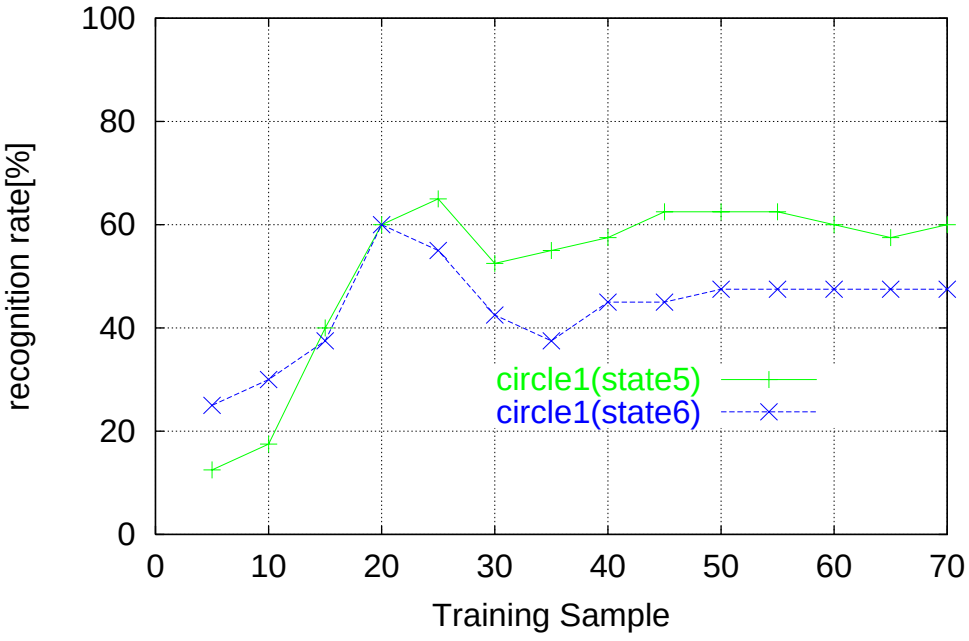


図 6.3: 円（１）の学習サンプル数と認識率の関係

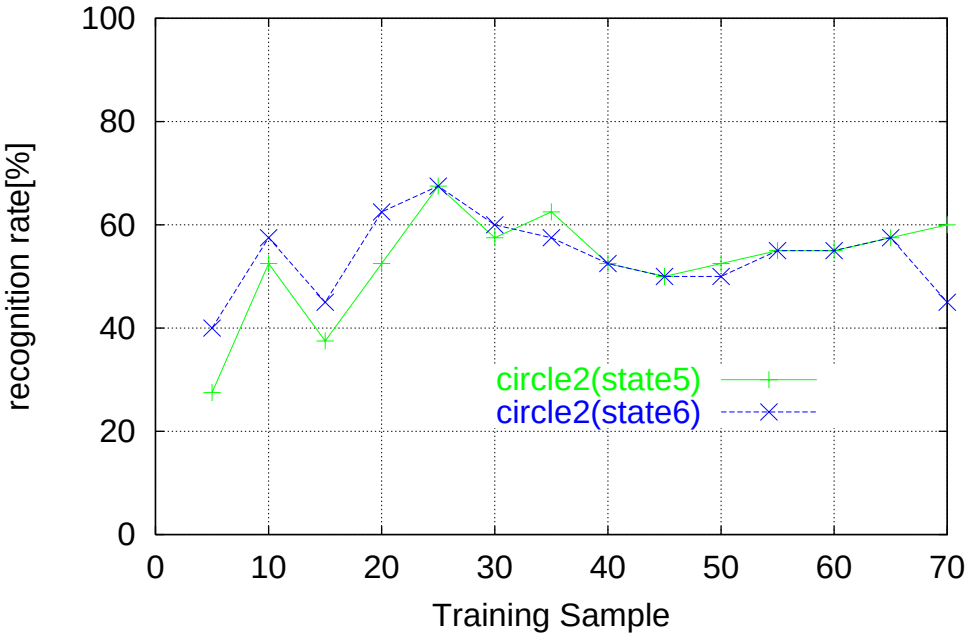


図 6.4: 円（２）の学習サンプル数と認識率の関係

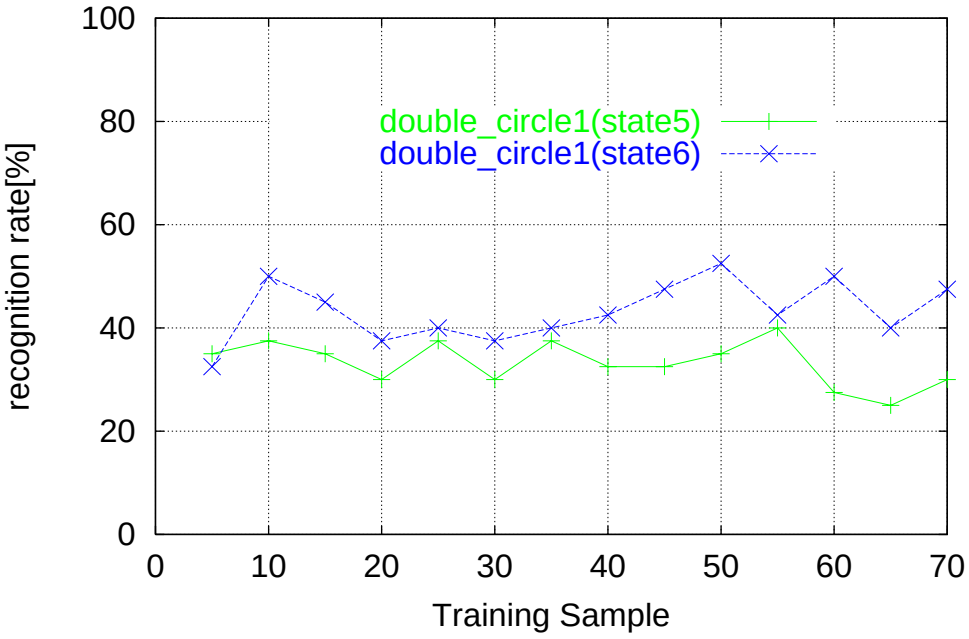


図 6.5: ２重円（１）の学習サンプル数と認識率の関係

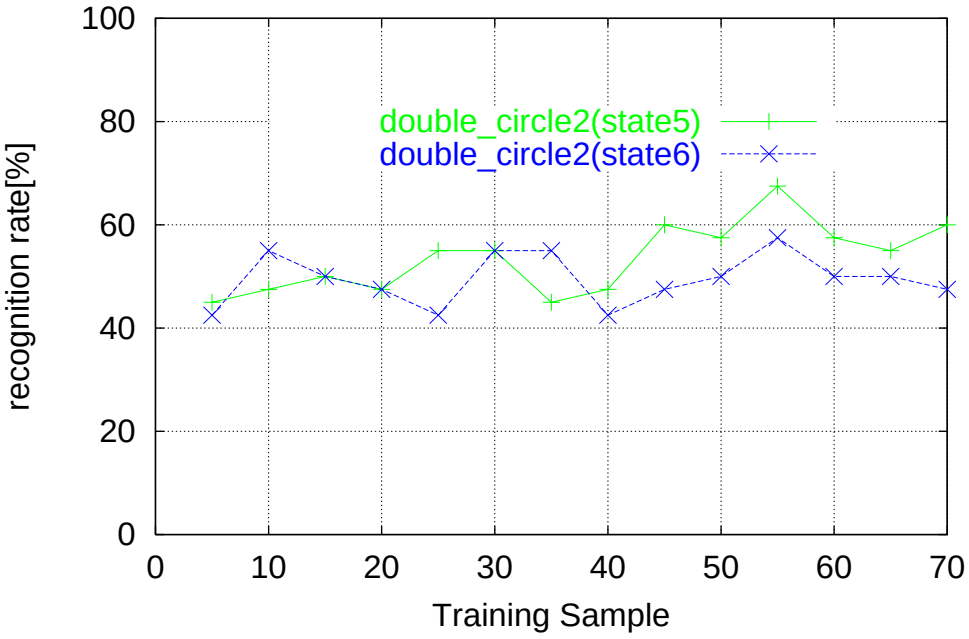


図 6.6: 2 重円（２）の学習サンプル数と認識率の関係

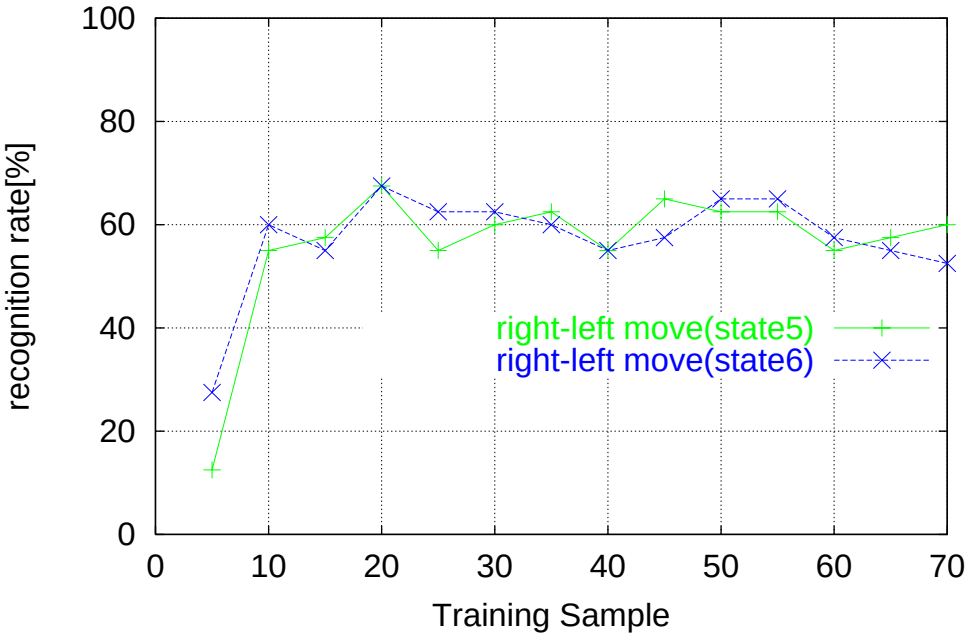


図 6.7: 往復の学習サンプル数と認識率の関係

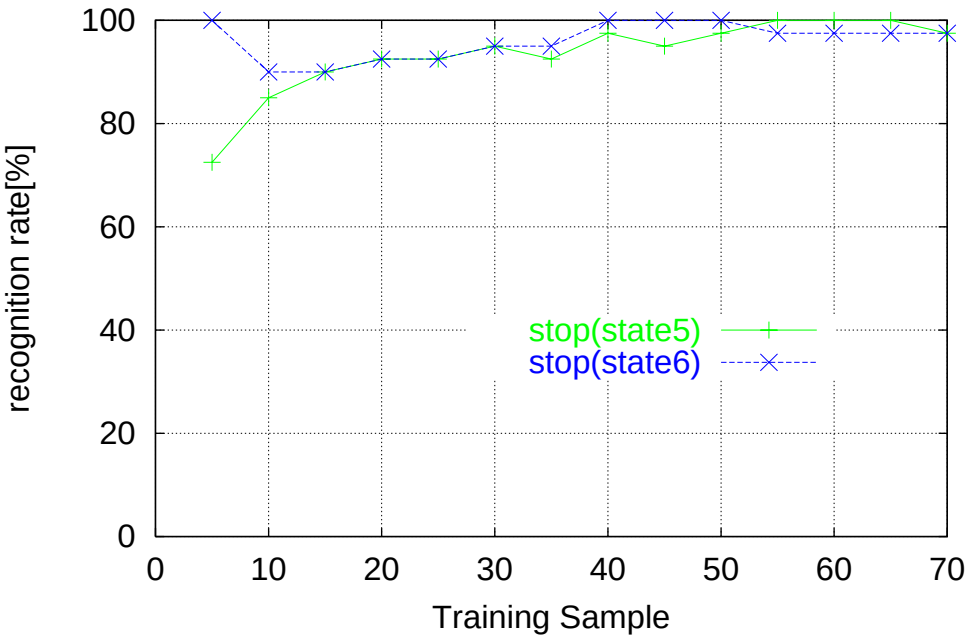


図 6.8: 停止の学習サンプル数と認識率の関係

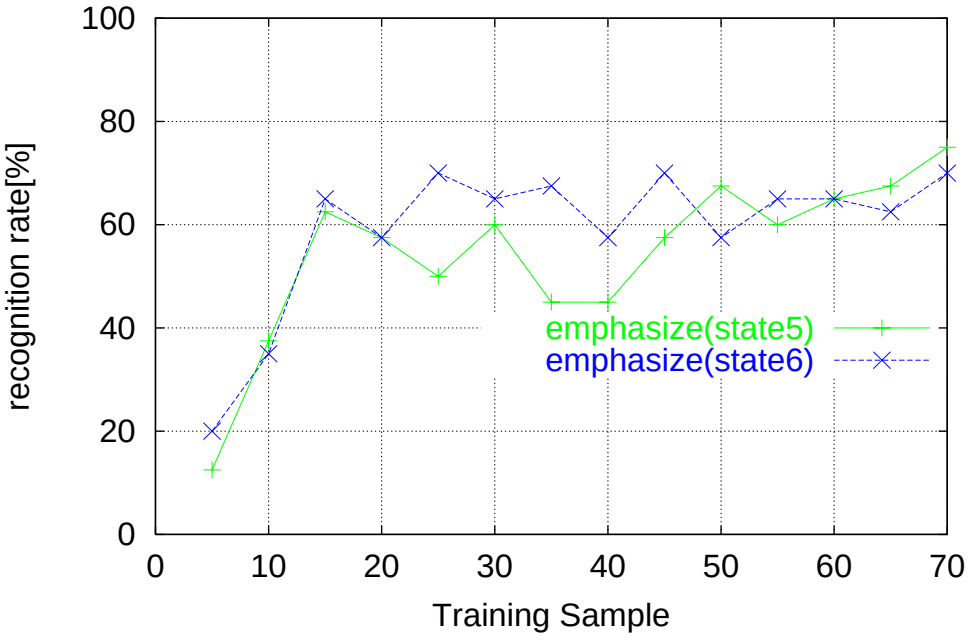


図 6.9: その他の学習サンプル数と認識率の関係

6.2 複数話者学習収束実験結果

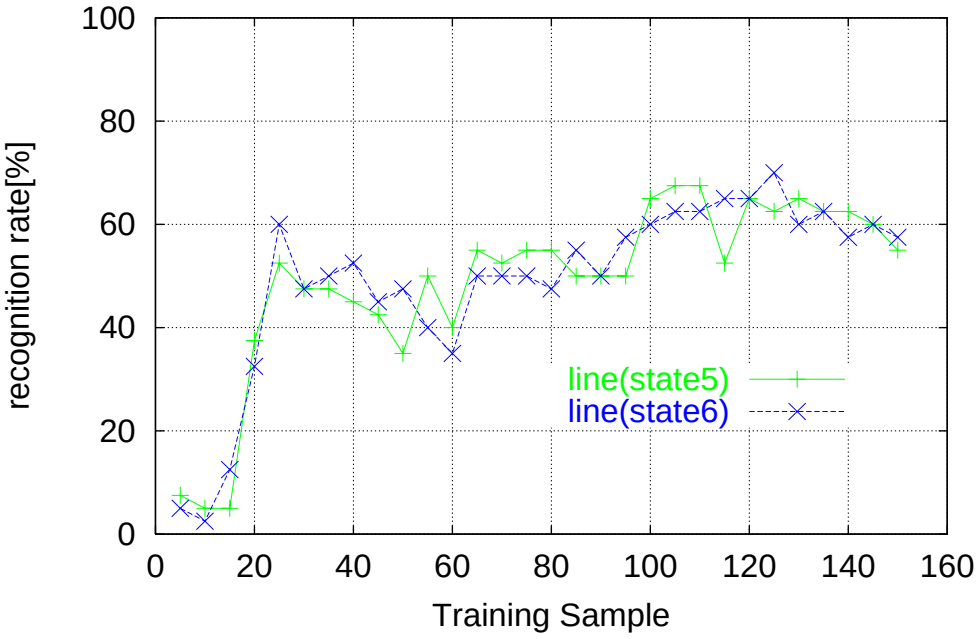


図 6.10: 直線動作の学習サンプル数と認識率の関係

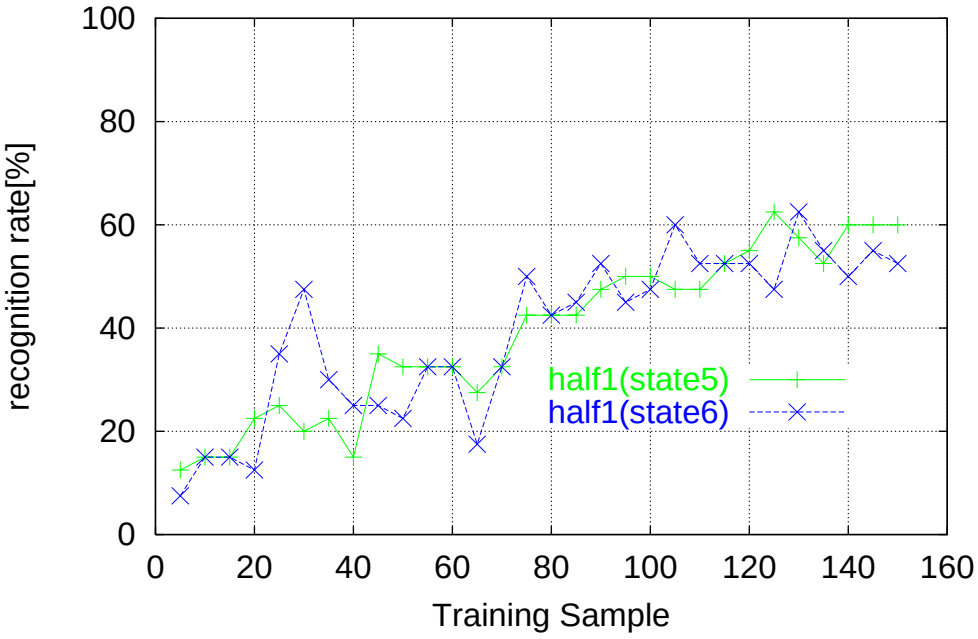


図 6.11: 半円動作の学習サンプル数と認識率の関係

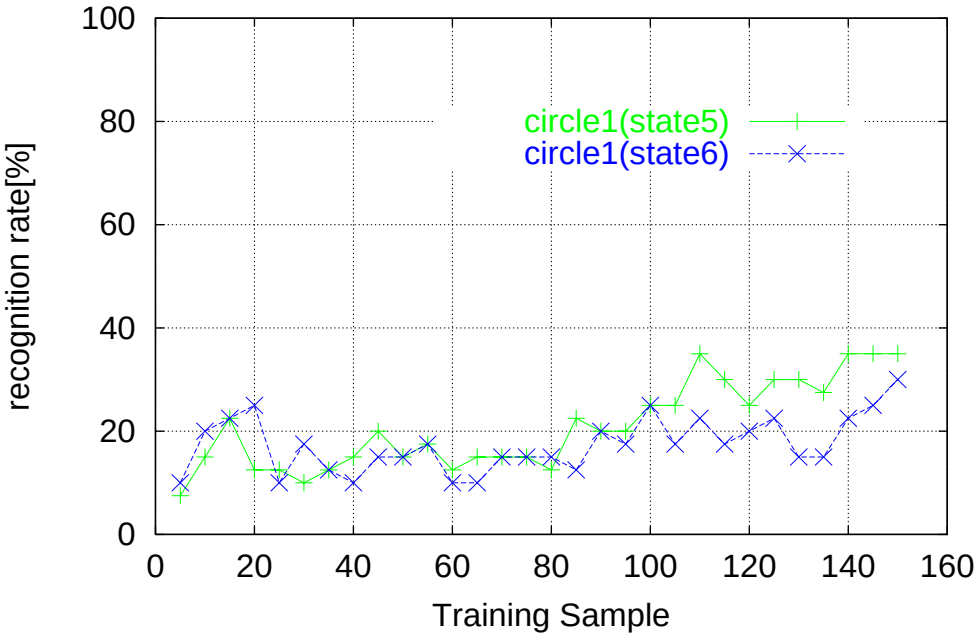


図 6.12: 円（１）の学習サンプル数と認識率の関係

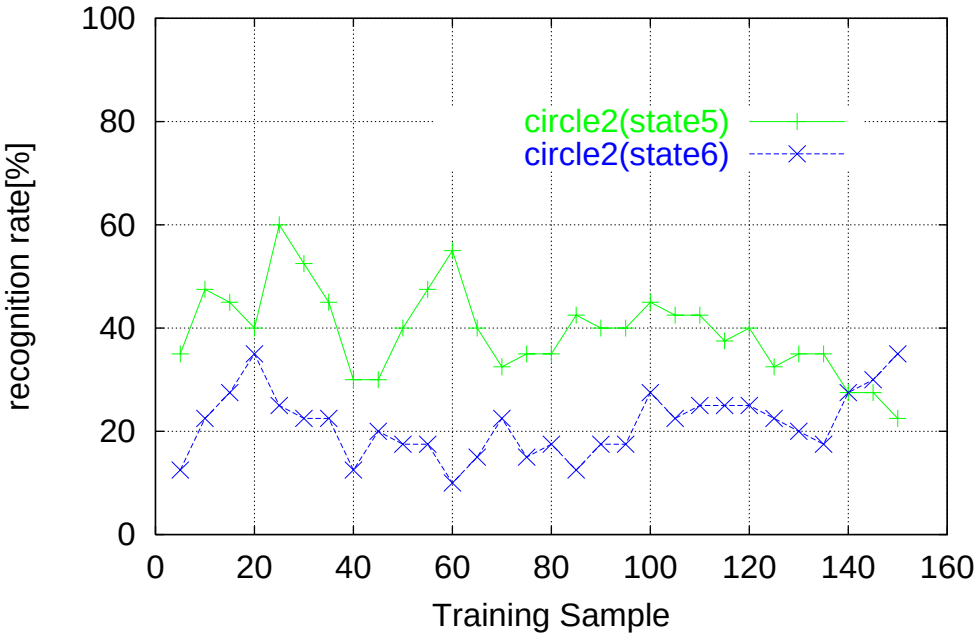


図 6.13: 円（２）の学習サンプル数と認識率の関係

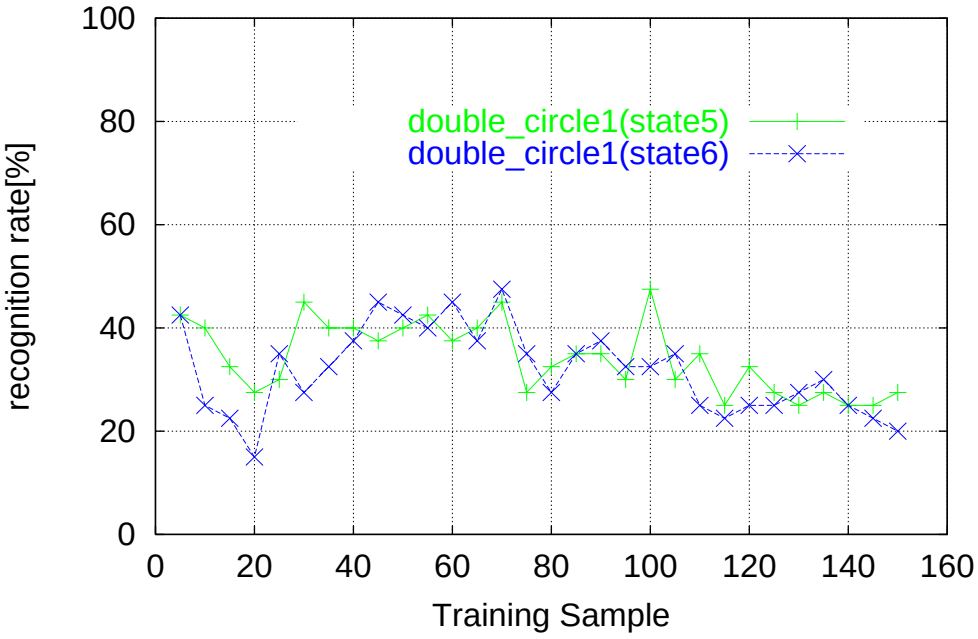


図 6.14: ２重円（１）の学習サンプル数と認識率の関係

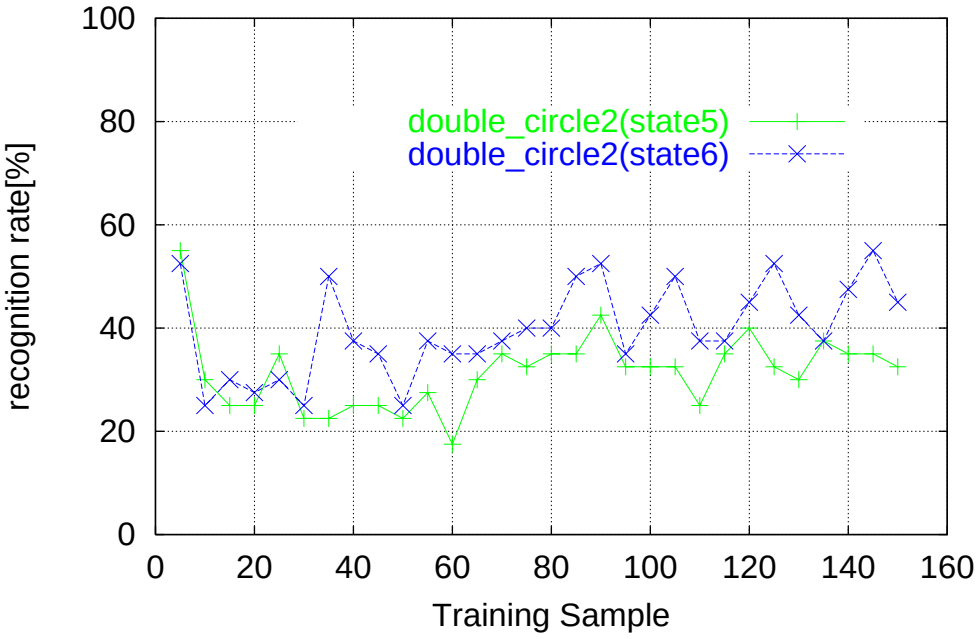


図 6.15: 2 重円（２）の学習サンプル数と認識率の関係

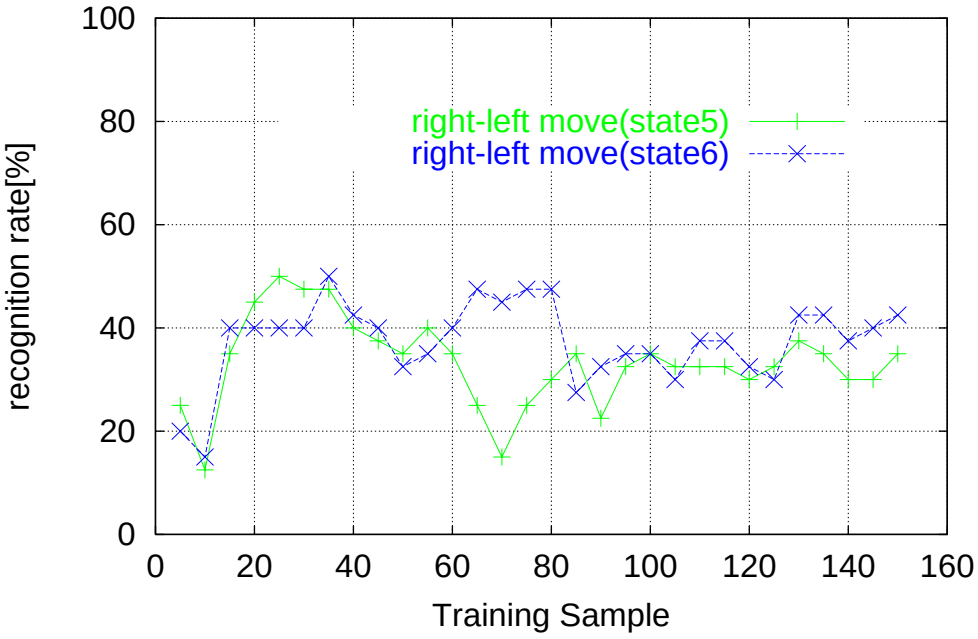


図 6.16: 往復の学習サンプル数と認識率の関係

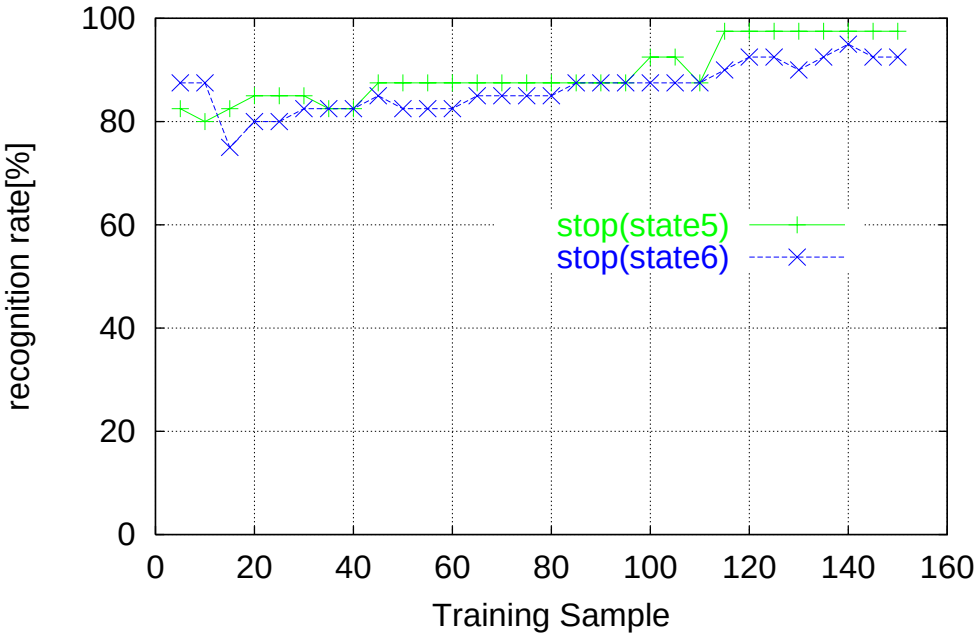


図 6.17: 停止の学習サンプル数と認識率の関係

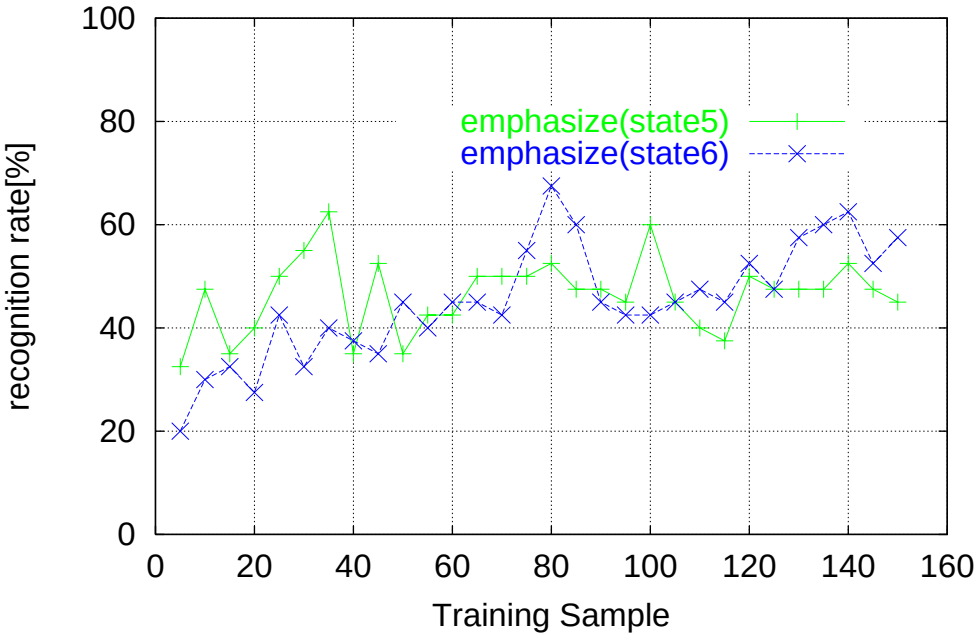


図 6.18: その他の学習サンプル数と認識率の関係

6.3 形状別平均認識率

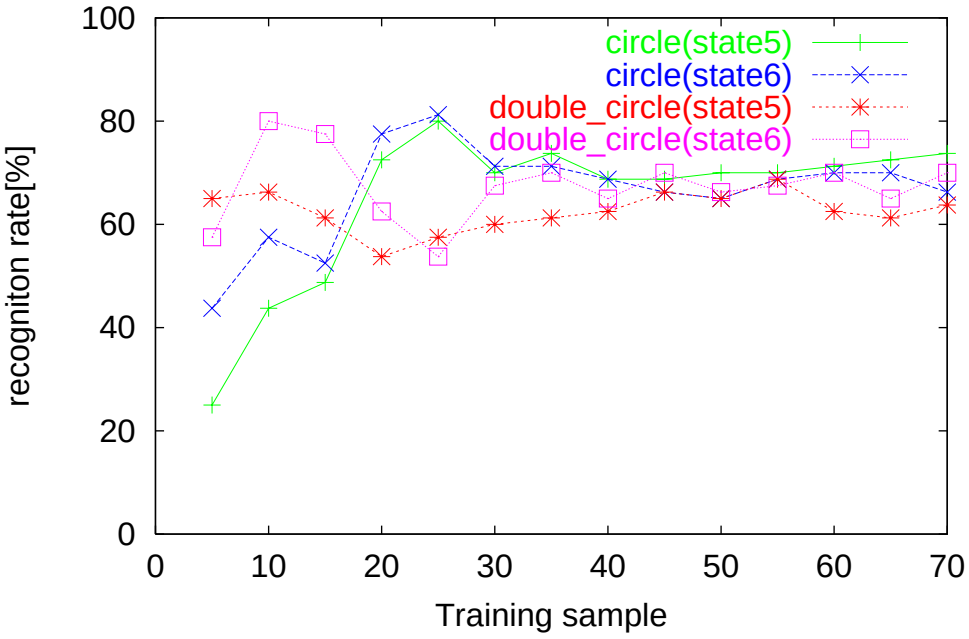


図 6.19: 円・２重円の学習収束結果（話者一名）

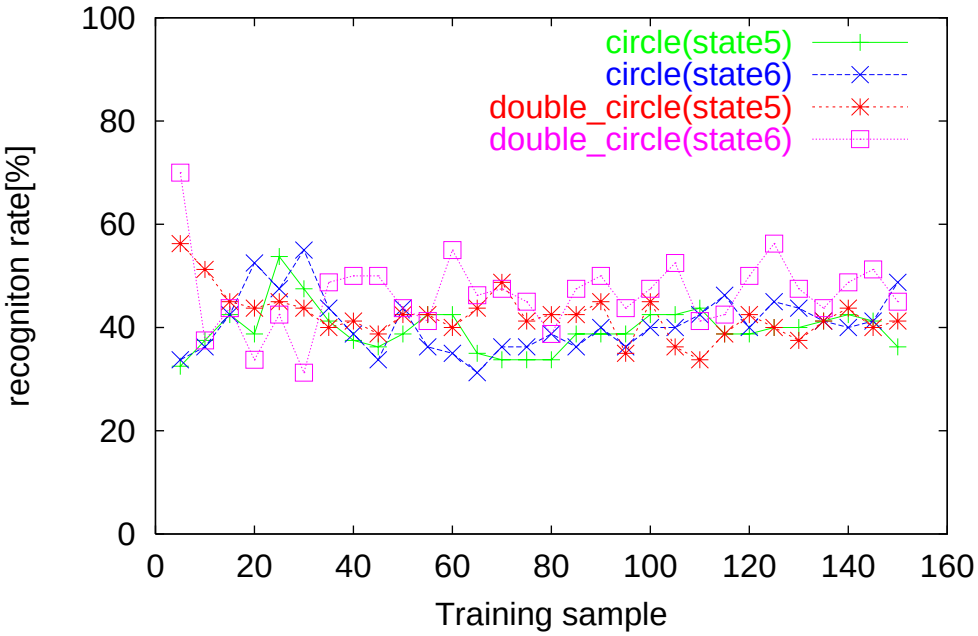


図 6.20: 円・２重円の学習収束結果（話者四名）