ロボットの音声単語獲得における前言語期学習のモデル化*

藤田 祐貴 (琉球大)　　高良 富夫 (琉球大)

1　まえがき

現在、ロボット工学の進展により、小型の人型ロボットは一般の人でも手に入れられるようになった。人型ロボットをより人間的にするためには、どのように人間の言葉を認識させ、動作させるべきかを考える必要がある。その手初めとして、ここではロボットによる前言語期学習から音声単語の獲得までを検討する。

本研究では音声認識方法として主流の隠れマルコフモデル (HMM) を使用する [1]。

前言語期学習とは、言語獲得期以前の準備期における言語学習である。人間の乳児は、この時期に意味が分からないまま、日常の音声にさらされている。しかし、意味が分からなくても、音の違いは聞き分けれるようになっていく。この前言語期学習は、人間にとって非常に重要で、この時期に、獲得される言語が決定されると考えられている [2]。また、前言語期学習があることにより、その後の言語獲得が高速に行われると考えられる。この前言語期学習をここでは、意味がまだ付与されていない HMM の訓練としてモデル化する。これは時系列パターンのクラスタリング、すなわち、教師なし学習として実行する。この訓練により、ロボットも音声言語の獲得をスムーズに行うことができるようになる。

音声単語の獲得は、ここでは、ロボットができる動作に対する自動音声ラベリングとしてとらえる。すなわち、ロボットが自分で動作単語を選択してその HMM を訓練することである。今回、ロボットは「よし」という単語だけは認識できるものとする。「よし」と言われると、その前に言われた単語と動作とをロボットが対応付ける。この能力だけを与えておけば、ロボットは自分のいくつかの動作に対応する音声単語を自動獲得できると考えられる。

以上のようなロボットは、まるで乳児が任意の言語を獲得できるように、任意に発音された音声単語を獲得できる。

2　音声単語の獲得

初めに、あらかじめロボットには右手を上げる、走るなどのいくつかの動作が行えるようにしておく。またロボットは HMM を使用して「よし」という単語だけは認識できるようにしておく。

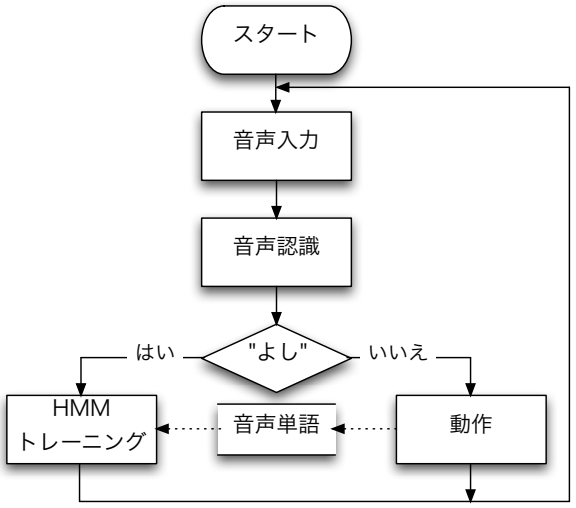

図 1　音声単語獲得の流れ

図 1 のような流れにより、音声単語を獲得する。初めに、ロボットに音声を入力する。ロボットはそれを認識して、自分ができる動作のうちひとつをランダムに選択し、実行する。さらに、入力された音声を一時蓄えておく。入力した音声に対して正しくない動作なら、人間がまた同じ単語を音声入力する。も

*”Modeling of Learning Pre-linguistic in Period for Acquisition of Spoken Word by a Robot”, by Yutaka FUJITA and Tomio TAKARA (University of the Ryukyus).

し、正しい動作をしたら「よし」と言ってやる。「よし」と言われたらロボットは、直前に入力された単語音声を用いて動作単語の HMM を訓練する。すると次からはこの単語を音声入力するとその単語にあった動作をするようになる。

3　前言語期学習

前言語期学習では、単語音声を動作のラベルとする前に、入力音声をクラスタリングする。入力する音声データは「右手」、「左手」、「走れ」、「万歳」、「後ろ」の 5 種類である。この 5 種類を話者一人が各 5 回発声した計 25 個の音声データを入力する。

3.1　静的閾値を用いた HMM クラスタリング

閾値を固定して 25 個の音声データをランダムに入力する。初めに入力された音声データの HMM を作成し、クラスタの代表点を形成する。二つ目の音声データの入力からは、各クラスタの HMM により尤度を計算する。そして、閾値を超え、かつ最も尤度が高いクラスタを選択し、入力データを新たにそのメンバとし、HMM を更新する。もし閾値を超える十分な尤度のクラスタがなければ、HMM を作成し、新たなクラスタを形成する。この流れをデータの数だけ繰り返して、HMM を代表点とするクラスタリングを行う。

閾値を変えてこの一連の流れを行ったクラスタリングの結果を図 2 に示す。

図 2 から分かるように、閾値を低くするに従って 1 個のメンバから成るクラスタの数が減っていく。しかし、クラスタのメンバを見ると、同じ単語同士でまとまってはいるが、別の単語も共にクラスタを形成している。閾値 -300 では「走れ」と「後ろ」が別のクラスタを形成せずに、2 単語で一つのクラスタを形成している。

3.2　動的閾値を用いた HMM クラスタリング

動的閾値を用いた HMM クラスタリングの流れを図 3 に示す。

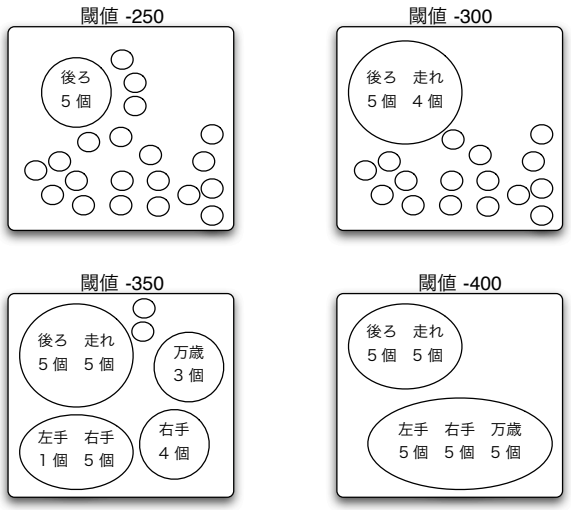

図 2　静的閾値による HMM クラスタリングの結果

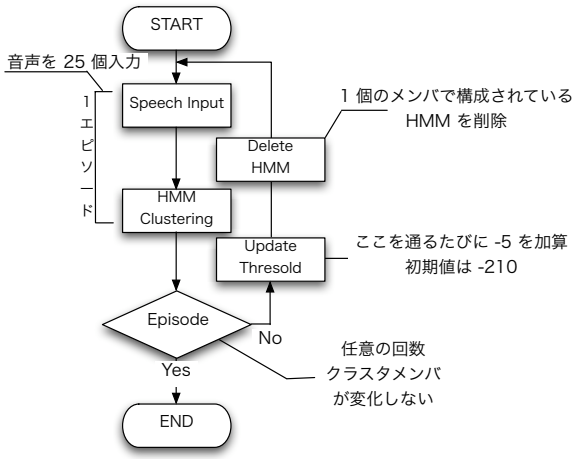

図 3　動的閾値による HMM クラスタリング

初めに、前節の静的閾値同様に音声データを入力し、HMM を代表点とするクラスタリングを行う。この処理を 1 エピソードとする。1 エピソードを終えるたびに閾値を更新し、1 個のメンバで構成されているクラスタを削除する。このエピソードは、すべての音声データがクラスタのメンバになり、かつクラスタのメンバが変化しなくなるまで繰り返えされる。この処理の結果を図 4 に示す。

図 4 から分かるようにエピソードを繰り返していくうちに同じ単語によりクラスタが形成されている。エピソード 12 では「走れ」と「後ろ」がうまく別々にクラスタを形成している。また、エピソード 29 では入力単語をすべて用いて 5 個のクラスタが形成されている。このクラスタリング手法は最終的な

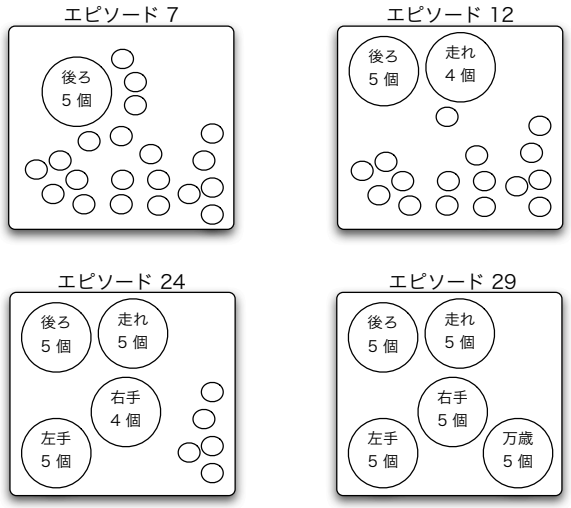

図 4　動的閾値による HMM クラスタリングの結果

クラスタ個数を指定せずに自動的にクラスタ個数が決まっている。今回、我々はこのクラスタリングを前言語期学習のモデルとする。

4　前言語期学習を用いた音声単語の獲得

前言語期学習を用いた音声単語の獲得のモデルを図 5 に示す。

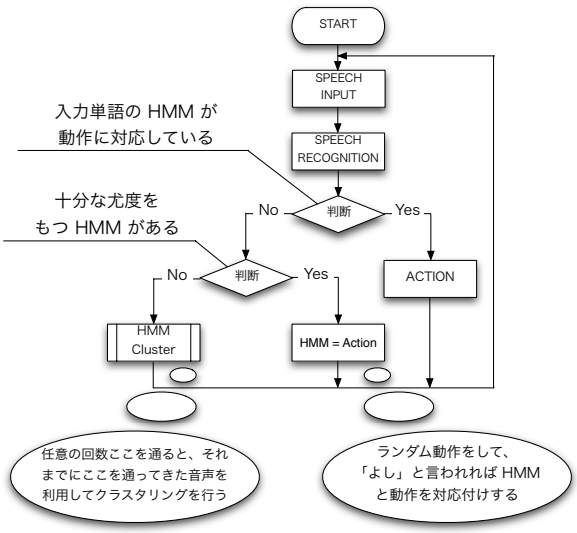

図 5　前言語期学習を用いた音声単語獲得の流れ

まず、前節と同じ音声データ 25 個をランダムに入力する。そして、前節の動的閾値による HMM クラスタリングを行う。それによって作成された HMM を使用する。作成された HMM はこの段階では動作と対応付けがされていない。

しかし、HMM と動作を対応付けすることができる。まず、再び音声データを入力していき、「よし」と言われれば、HMM と動作を対応付ける。つまり、ロボットは入力音声により動作にラベリングができたことになる。

入力音声と動作が対応していれば正しく行動でき、まだ対応していなければ「よし」と言われるまで音声入力を続ける。また、HMM クラスタリング時に HMM を十分に訓練しているので、前言語期学習を用いない方法より少ない音声入力回数で音声単語の獲得ができる。

しかしながら、ロボットの動作はランダムに選択されているので音声入力回数に無駄がある。そこで、次節で No-List を用いた動作選択を提案する。

5　No-List を用いた動作選択

No-List は過去の間違った動作を単語ごとに記憶したリストである。このリストを用いることによって同じ間違った動作を選択しないようにすることができる。例えば、「右手」と音声認識したときに、左手を上げて「よし」と言われなかったとする。次から「右手」と音声認識したときには左手を上げる以外の動作を選択させる。これによりランダム法による動作選択より音声入力回数を少なくすることが可能である。

6　実験

ロボットによる音声単語の獲得における音声入力回数の検討のため、ロボットを使用しない以下の 3 つのシミュレーション実験を行った。

- 前言語期学習なし
- 前言語期学習あり
- 前言語期学習あり + No-List

6.1　実験方法

入力としては話者一人による「右手」、「左手」、「走れ」、「万歳」、「後ろ」の 5 単語を 5 回発話した 25 個の音声データを用いる。音声の入力順序はランダムとし、すべての実

験で同じである。ただし、音声データ 25 個をすべて入力し終わるまで、同じデータは入力されないものとする。ロボットは初めから「よし」だけ認識できるものとし、入力単語に対応した 5 つの動作を持つ。また、ロボットが正しい動作を行ったら必ず「よし」と言われるものとする。

各実験の終了条件は 5 単語を連続して正しく認識したときとし、途中に認識誤りがないものとする。この条件を基に各実験を 10 回行い、それぞれの平均音声入力回数を求める。

6.2 実験結果

実験の結果を表 1、2、3 に示す。

表 1 と表 2 を比較すると、平均音声入力回数は、「前言語期学習あり」では半分以下になっている。「前言語期学習なし」では一度の音声単語と動作の対応付けで HMM を作成しても十分な尤度がないため正しく認識できないことがある。すなわち、音声単語と動作の対応付けの際に何度か HMM を訓練しないと同じ単語でも正しく認識できない。そのため「前言語期学習あり」より音声入力回数が多くなっている。「前言語期学習あり」では HMM クラスタリングの際に HMM に十分な尤度があるため、認識誤りが生じない。そのため音声入力回数が少なくなっている。

表 1　前言語期学習なし

	平均音声入力回数
5 連続認識	152.6

表 2　前言語期学習あり

	平均音声入力回数
5 連続認識	74.2

表 3　前言語期学習あり + No-List

	平均音声入力回数
5 連続認識	51.6

表 2 と表 3 を比較すると「前言語期学習」に No-List を加えたものが音声入力回数が少ないことが分かる。これはロボットが動作を選択する際に No-List を有効利用した結果である。つまり、一度間違えた動作を再び選択しないことにより、正しい動作を選択する確率が増したためである。

7 むすび

ロボットが「よし」だけを認識できることを仮定して、音声単語を獲得できることが示された。また、前言語期学習は、まだ意味が付与されていない HMM の訓練と定義した。これは、音声単語獲得がスムーズに行えるための要因となった。動作選択の際には、ただ単にランダムではなく、過去の情報を利用することにより、音声単語獲得を少ない入力回数で行うことができることが示された。

今後の課題としては、前言語期学習後の音声認識を行う際の閾値の決め方、No-List だけではなく、既に対応付けが終わっている動作を利用した動作選択等が挙げられる。また、前言語期学習時の HMM クラスタリングのロバスト性を検証する必要がある。

参考文献

[1] 古井 貞熙: ”音声情報処理”, pp.96-99, 森北出版株式会社, 2002

[2] 酒井 邦嘉: ”言語の脳科学”,pp.283-284, 中公新書,2002