3B2-01

手をとって遊んでもらうことにより学習するぬいぐるみロボット

A Teddy-Bear Robot That Learns from Being Taken Its Hands

中伏木 新[*1] 岡 夏樹[*1] 伊藤 慶明[*2]
Arata Nakafushiki Natsuki Oka Yoshiaki Itoh

[*1]京都工芸繊維大学大学院 工芸科学研究科 電子情報工学専攻
Kyoto Institute of Technology

[*2]岩手県立大学 ソフトウェア情報学部
Iwate Prefectural University

What is a good way for robots or agents to interact smoothly with a person? When we interact with a person, we deal with much information from the person. It should be the same for robots or agents. In this paper, we develop a system that learns relationship between voice and motion from continuous input. This system finds recurrent patterns in voice and motion using Continuous DP, and learns relationship between voice and motion from the co-occurrence of the both inputs.

1. はじめに

ロボットやエージェントが人と自然なインタラクションを行えるようにするためにはどうしたらよいのだろうか。人と人の間でインタラクションを行う場合，いろいろな情報を感知して相手に応じたふるまいを行っている。つまり，ロボットやエージェントも同様にして、状況に応じたふるまいができなければならない。近年音声認識や表情の認識など個々の認識技術は発展し，様々なロボットが商品として出されているがそれらはあらかじめ組み込まれた行動を持っているのであって，人とのインタラクションで考えられうるような未知の状況に上手く対応できるようなすべを持っているわけではない。そこで乳児が生まれ持った能力から様々な状況から情報を感知して育っていくのを参考に，ロボットなどに最低限必要な能力を持たせ人間が学習させるよう，さらに学習させる人間に負担がかからないよう楽しみながら出来ればよいのではないだろうか。

今回はロボットと遊んでいるうちに，具体的には声をかけながらロボットの手を取り動作を教えると学習を行うようにする。この場合，自然な発話など入力が連続であるので，対応付けるべき音声や動作の中で繰り返し出現しているものを連続 DP を用いてパターンとして切り出し，次に同時に生起するパターンを関連付けることとする。これは乳児が音声中において繰り返し出現するパターンに基づいて単語の区切りを見つける能力を持つという実験報告 [Saffran 96] を元とし，また例えば「ばいばい」と言いながら手を相手に向かって振るというように動作と音声が同時に生起する状況を想定した。

関連研究として Deb Roy は，本研究の音声と動作の対応付けとは違い，視覚と音声を対応付けて言葉を学習する CELL[Roy 99] を開発している。CELL では複数の角度から写された静止画像を使用しており，今回の研究とは音声，動作ともに時系列データとして扱っているという点で異なっている。

2. 構築するシステム

2.1 RobotPHONE

RobotPHONE（図.1）とは RUI（Robot User Interface）の一つの形として，ロボットの持つ身体性に着目し，ロボットを入出力インタフェースとして使用するコミュニケーションデバイスである [Sekiguchi 01]。ここで RUI とはロボットが人間に対する有効なインタフェースになるという考えに立って、それを実現するインターフェースを GUI（Graphical User Interface）になぞらえたものである。RobotPHONE は東京大学 館研究室のプロジェクトの一つで，現在はイワヤ株式会社で「IP RobotPHONE」として製品化されている [Iwaya 05]。

図 1: IP RobotPHONE

RobotPHONE はマイクとスピーカーを１つずつ備えており、さらに首に２自由度、肩に２自由度の駆動軸がある。今回はこのロボットを用いてシステムを構築する。

2.2 システムの構成

今回構築するシステムの構成としては，人間が動作を入力するロボットを一体，音声入力用のマイクを一つ用意し，そしてそれらを PC に接続し PC 上のプログラムで制御を行う。プログラムは，大きく分けて音声や動作の入力制御部，その入力からのパターン切り出しを行う部分，切り出された音声及び動作のパターンを関連付ける部分，どのパターンが関連付けられているかを記憶する記憶部に分かれる。

このシステムの動作は，学習フェーズと，音声応答フェーズ，の２つのフェーズに分かれる。

システムに対する入力として，あらかじめ対応付けを行いたい音声パターンと動作パターンの組み合わせをいくつか決定しておき，学習フェーズではそれらを同時に入力することと定

連絡先: 中伏木 新 京都工芸繊維大学大学院パターン情報処理研究室 e-mail:arata@vox.dj.kitac.jp

める。また音声応答フェーズでは学習フェーズで入力した音声パターンと同じものを入力するとする。学習フェーズでの動作は以下のようになる。

1. 音声はマイクを通して，動作はロボットの手や頭を直接動かして，入力を行う。
2. 入力からそれぞれ連続 DP を用いて，繰り返し出現するパターンを音声と動作の標準パターンとして切り出す。
3. 検出された音声パターン，動作パターンで同じ時間に入力されたと思われるものの間に関連付けを行い，それを記憶する。

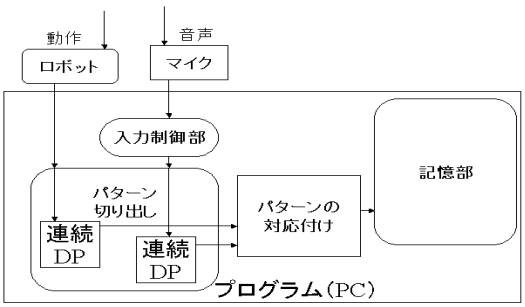

図 2: 学習フェーズの動作

また，システムが音声に応答するときは以下のような動きになる。

1. システムに対して音声を入力。
2. 音声パターンを記憶部から取り出し，連続 DP で入力音声と比較を行う。
3. 入力音声に音声パターンと同じパターンであると判断された部分がある場合，その音声パターンと関連付けられた動作パターンを取り出す。
4. 動作パターンを用いてロボットを動かす。

図 2，3 にそれぞれのフェーズの動作の概要図を表す。

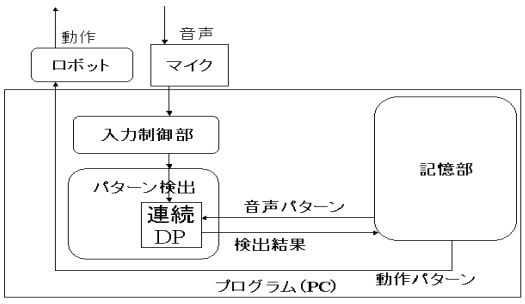

図 3: 音声応答フェーズの動作

3. 連続 DP

3.1 連続 DP とは

連続 DP とは非線形に伸縮しうる 2 つの時系列データを比較する DP マッチング（Dynamic Programming：動的計画法）を、連続音声認識などに応用したものである [aoshima 92]。連続 DP では事前に用意されている標準パターンを入力された時系列データの始端から 1 フレームずつずらしながら時系列データの部分区間と DP マッチングを行う。そして全標準パターンと入力の部分区間のマッチング結果（距離）中に閾値以下になるものが存在するとき、その時系列の部分区間に標準パターンが存在すると判定する。

3.2 IRIFCDP

連続 DP を用いてパターン認識を行う場合，標準パターンは通常事前に準備しなければならない。しかし，どのような時系列データが入力されるかがあらかじめわからない場合，事前に適切な標準パターンを用意するのは困難である。この欠点を解消し，標準パターンを入力された時系列から自動的に獲得していく手法として，IRIFCDP（Incremental Reference Interval-free Continuous DP）[koyama 96] がある。この手法は 2 時系列データ間の共通区間検出手法である RIFCDP（Reference Interval-free Continuous DP）[itoh 96] の拡張であり，入力時系列データに同期して標準パターンを動的に更新することで，時系列データの入力と共通区間対検出の同時進行を実現している。今回は入力時系列データから標準パターンの切り出しを行う方法として IRIFCDP の概要を述べる。IRIFCDP では，無限に続く入力時系列データを考える。このとき，$t_1 \leq t_2 < \tau_1 \leq \tau_2$ となる，区間 $[t_1, t_2]$ と $[\tau_1, \tau_2]$ があるとする。簡単に言えば，この 2 つの区間で RIFCDP を行うことにより共通区間を検出し標準パターンを得ることとなる。また最新の入力フレームがデータの末尾に付け加えられていくため常に最新の区間が共通区間探索の対象となる。よって標準パターンの自動的な獲得を行うことができる。

3.3 ShiftCDP

ShiftCDP とは，RIFCDP の最大の問題点である必要とする計算資源が大きいことを解消し，なおかつ RIFCDP の性能を維持することを目標としたものである [itoh 03]。ShiftCDP の概念図を図.4 に示す。ShiftCDP では単位標準パターンを先頭から N_{Shift} フレームずつシフトしながら標準パターンを構成していく。単位標準パターンは固定フレーム数 N_{CDP} からなる。シフトするフレーム数は任意の設定が可能である。そして，この各単位標準パターンに対し連続 DP を実施し，その結果を連結及び統合するという，単純な構造となっている。

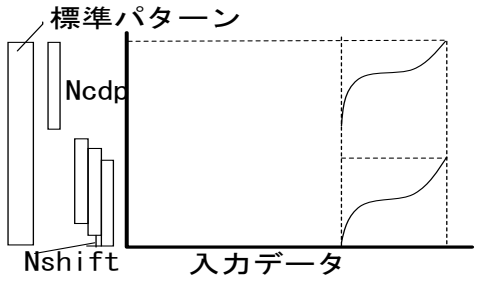

図 4: ShiftCDP の概念

4. 対応付けの学習

4.1 標準パターンの取得

時系列データより標準パターンを取得するアルゴリズムはIRIFCDP を参考にして以下のようにした。図 5 に概要図を示す。

1. 入力時系列データの先頭から決まったフレーム数を切り取り，それを標準パターンの候補とする。フレーム数を固定としてもいいのは，ShiftCDP によってフレーム数以上の区間を検出できるからである。
2. 標準パターンの候補と入力時系列データのうち標準パターン候補出現以前のデータを除いたものとの間で ShiftCDP を実施する。
3. ShiftCDP により検出された区間の距離が閾値以下なら，その検出区間を標準パターンとして採用する。
4. 入力時系列データの先頭から切り取ったフレーム数の分移動し，そこから 1 と同様に候補を切り出す。
5. 入力時系列データの終端になるまで，2 から 4 を繰り返す。

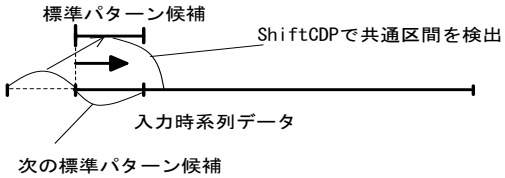

図 5: 標準パターンの取得

4.2 音声パターンと動作パターンの対応付け

4.1 のアルゴリズムを行うと同時に音声・動作パターンの対応付けを行う。今回は同時に生起されているパターンを対応付けることにする。

一般的に音声や動作の特徴量を取得するときのパラメータ，例えば音声におけるサンプリング周波数など，はその対象によって異なり，その影響で音声の標準パターンと動作の標準パターンのフレーム数は異なる。

そこでまずある時刻が入力の終端であるとする。このとき図 6 のように始端から終端までの音声の時系列データで，あらかじめ決定したフレーム数によって始端から区切っていくと標準パターンの候補が最大いくつ取れるかを計算し，動作の標準パターン候補もそれと同数となるように動作の時系列データを区切っていく。そして音声標準パターンの候補を出現時刻が早い順に番号を割り当てていく。同様にして動作の標準パターン候補も番号を割り当てる。ある時間内で音声，動作ともに同じ数の標準パターンの候補が存在するので，同じ番号の標準パターン候補同士を同時に生起されたパターンとして仮に記憶する。

そして音声標準パターン候補を用いて 4.1 のアルゴリズムを行う。ある番号の音声標準パターン候補が採用されなければ同時に動作の標準パターン候補も採用されないことになる。また，音声の標準パターン候補が採用されても動作標準パターン候補を用いた 4.1 のアルゴリズムの結果で採用されなければ，その番号の音声標準パターン候補は採用しないようにする。

このように絞込みを行い，最終的に残った音声・動作標準パターンを対応付けられたものとして記憶部に登録する。

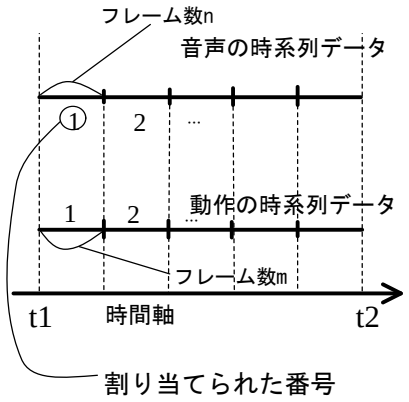

図 6: 時系列データの区切り

4.3 音声の認識と動作パターンの決定

音声に対応した動作を行わせたい場合，まず入力された音声と保持されている全ての音声標準パターンとの間で ShiftCDP を行う。基本的には ShiftCDP によって音声標準パターンと入力音声との間で類似した区間が検出されれば，その類似区間が出現する順を動作を行う順番として，音声標準パターンに対応付けられた動作標準パターンを採用し，ロボットの動作とする。

このとき，図 7 のように検出区間が別の音声標準パターンによる検出区間と重なる可能性がある。その場合，現在の標準パターンによる ShiftCDP による結果の距離が別の標準パターンによる ShiftCDP の結果の距離と比べて小さいという条件を満たせば，現在の音声標準パターンのほうが入力音声に近いものとして，対応付けられている動作標準パターンを採用する。

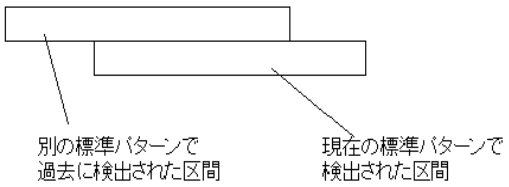

図 7: 検出区間の重なり

5. システムの実装と動作確認

5.1 実装

今回の実装における開発環境は，CPU Pentium 700MHz の Microsoft Windows2000 SP4，Visual C++ .NET 2003 を使用した。

RobotPHONE の動作は RobotPHONE SDK により操作を行う。SDK により，動作データの保存，各駆動軸の角度データの取得，設定などを行うことができる [Motion 05]。

まず，学習フェーズにおける入力部分と音声パターンと動作パターンの関連付けを行う部分を分割して実装を行った。したがって，学習データ入力フェーズ，パターン関連付けフェーズ，音声応答フェーズの３段階に分ける。理由としては，パターン間の対応付けを行う処理やデータ入力の処理，ShiftCDP による共通区間検出の処理が重なることにより，今回の環境ではリアルタイム性が維持できない可能性がある，また分割することによりプログラム上の複雑さを避けられること，が上げられる。音声の特徴量はサンプリング周波数 16KHz，サンプルビット 16bit の音声データを直接バイナリファイルとして保存し，HTK（HMM Tool Kit）[htk 05] を使用し 12 次の MFCC（Mel Frequecy Cepstrum Coefficent），その 1 次微分 12 次、2 次微分 12 次，パワー 1 次を抽出した。このときハミング窓を用い，窓のサイズ 25msec，フレーム間隔 10msec で抽出した。また RobotPHONE の各駆動軸の角度データは 30msec 毎に取得したものを動作の特徴量とした。ShiftCDP のシフト数は 5 とし，音声パターンのフレーム長は 60 とした。

音声に対応した動作を行う場合，学習時の入力と動作はほぼ同じだが，２秒間分の音声データを保存してから MFCC を抽出し ShiftCDP で共通区間を検出，4.3 で述べたアルゴリズムを用いて動作の決定を行った。

5.2 動作確認

以上のようにして構築したシステムの動作確認を行った。

まず学習フェーズでは RobotPHONE の右手を 1 回上げて下げながら「右手」と音声を入力する。次に左手も同様にして行い，これを交互に何回か繰り返し行い入力した。次に関連付けを行い，システムに対して音声で「右手」もしくは「左手」と入力し，それに対する応答を見た。

現時点の結果としては，まだ安定した動作は実現できていない。例えば,「右手」「左手」と続けて入力したとき，連続して右手だけを，または左手だけを上げることが多くあった。

6. おわりに

今回は同時に生起されるパターン間で対応付けを行い，音声に対応した動作を行うシステムを構築したが安定した動作はまだ実現できていない。これから考えられうることは，1) 正しく標準パターンが取得できていない，2) 対応付けが上手く行われていない，などがある。

例えば対応付けの学習を行うアルゴリズムについては，様々な改善すべき点がある。標準パターンを取得するアルゴリズムの場合，標準パターン候補の選び方では今回考えなかった 2 つの標準パターン候補をまたぐようなパターンがあったときには検出漏れがおこる可能性がある。

またこれに関連して，音声応答フェーズにおいてどの動作標準パターンを選択するか，ということにも再考の余地がある。今回は，図 7 のように音声標準パターンと入力音声の共通区間が過去に他の音声標準パターンで検出された共通区間と重なる場合，4.3 の方法を用いて動作パターンを決定していった。しかし，これはその共通区間の重なってない部分についての考慮がなされていないということでもある。これを改善するには，標準パターンを取得する部分にさかのぼって，2 つの標準パターン候補をまたぐようなパターンを新たに標準パターンの候補として採用することが考えられる。

ほかには，今回それぞれ取得した音声標準パターン間どうし，または動作標準パターン間どうしの関係については何も考慮に入れていない。たとえばある音声パターンの次にはこの音声パターンがくる確率が高いというようなことだ。今回は音声パターンと動作パターンのことのみ注目していたので，このことも考える必要がある。

さらには今回用いた特徴量や距離尺度が問題に対して妥当であったかどうか，実装面ではリアルタイム入力からの学習の実現などが上げられる。

謝辞

東京大学の関口大陸氏には RobotPHONE SDK の提供をしていただいた。またイワヤ（株）の中野殖夫氏には RobotPHONE に関して様々な相談にのっていただいた。本研究の一部は、国立情報学研究所との共同研究「人間とエージェントの適応のためのインタラクション設計」によるものである。

参考文献

[Saffran 96] Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport:“Statistical Learning by 8-Month-Old Infants” Science, 274, 1926-1928. (1996).

[Roy 99] Deb Roy：“Learning from Sights and Sounds: A Computational Model” Ph.D. Thesis, MIT Media Laboratory.（1999）.

[Sekiguchi 01] D. Sekiguchi, M.Inami, S. Tachi: “RobotPHONE: RUI for Interpersonal Communication”, CHI2001 Extended Abstracts, pp.277-278（2001）.

[Iwaya 05] イワヤ株式会社：“http://www.iwaya.co.jp/index.html”,（2005）.

[aoshima 92] 青島 伸治, 小畑 秀文, 南谷 崇: “電子情報工学入門シリーズ２ 音響・音声工学”，pp183-185，pp.197-198，近代科学社（1992）.

[koyama 96] 木山 次郎，伊藤 慶明，岡 隆一: “Incremental Reference Interval-free 連続 DP による任意話題音声の要約と話題境界検出”，電子情報通信学会論文誌，D-II Vol.J79-D-II No.9，pp.1464-1473（1996）.

[itoh 96] 伊藤 慶明, 木山 次郎, 小島 浩, 関 進, 岡 隆一:“時系列標準パターンの任意空間によるスポッティングのための Reference Interval-free 連続 DP（RIFCDP)”，電子情報通信学会論文誌，D-II Vol.J79-D-II No.9，pp.1474-1483（1996）.

[itoh 03] 伊藤 慶明: “時系列パターンの任意部分区間の高速マッチング手法 ShiftCDP 法”，電子情報通信学会論文誌，D-II Vol.J86-D-II No.9，pp.1267-1277（2003）.

[Motion 05] 計測自動制御学会 SI 部門 モーションメディア調査研究会 ダウンロード：“ http://www.star.t.u-tokyo.ac.jp/~dairoku/mm/index.php?Contest%2FDownloads”,（2005）.

[htk 05] HTK Speech Recognition Toolkit：“http://htk.eng.cam.ac.uk/”,（2005）.