社団法人　人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B002-3

言語・非言語情報を統合した指示パターンに対応するロボットの行動則獲得

Learning Robot Action Controller Corresponding to Direction by Verbal and Non-verbal Information

岡田　将吾，伊豆蔵　拓也，名渕　博人，高橋　徹，西田　豊明

Shogo Okada, Takuya Izukura, Hiroto Nabuchi, Toru Takahashi, Toyoaki Nishida

京都大学　情報学研究科　知能情報学専攻

Dept. of Intelligence Science and Technology, Kyoto University

okada_s@i.kyoto-u.ac.jp

Abstract

Human-Robot Interaction using free hand gestures and speaking word is more importance for humans which are operating robots in home or office environments. In this paper, we propose a novel technique for learning gesture command and spoken language command, action corresponding to these command by just observing interaction behavior of user with robot operated by a human operator. The main contribution of this paper is the introduction of a novel algorithm to segment and cluster patterns in its perceived signals. Proposed algorithm find gesture patterns and action patterns by using information of speech unit. The Experimental result shows that gesture patterns and action patterns are able to discovered with 85.0% ,88.0% respectively by using proposed pattern discovery algorithm.

1　はじめに

ユーザが自由なジェスチャ(非言語) と言語を用いてロボットに指示出来るインターフェイスはヒューマンロボットインタラクションにとって重要な機能の 1 つである．ロボットは聴覚とユーザの表出する自由なジェスチャを認識出来るセンサ双方を統合してユーザの指示行為を観測し，その観測結果の認識に基づき行動する必要がある．

本研究では，ユーザの言語による指示パターンと自由なジェスチャによる指示パターン，それらに対応するロボットの駆動パターンの対をインタラクションの履歴データからボトムアップに学習・獲得するシステムを提案する．まず提案システムでは人間とロボットの間で行われる言語および非言語による指示によりロボットをナビゲーションする，一連のインタラクション活動を観察する．観察から得られるユーザの発話・ジェスチャとロボットの駆動履歴に関する時系列データを，ロボットのための訓練データとして取得する．この訓練データから，ロボットに対する指示のパターンとそれに対応する適切な動作のパターンを学習により獲得する．

言語情報は，音声区間検出および音声単語認識により，発話区間および単語のシンボル情報として抽出する．ジェスチャは，光学式モーションキャプチャにより取得した位置座標データ系列として抽出する．この時系列データに対し，モチーフ発見アルゴリズムを用いることによりジェスチャパターンを抽出する．このアルゴリズムをロボットの駆動履歴系列データにも適用し，ロボットの駆動パターン（右に進む，左に回転する）を抽出し，抽出したパターン群に対しクラスタリングを行い，ジェスチャ・動作のシンボル化を行う．

本論文では，上記のシステムの内，ユーザの音声発話区間をジェスチャパターンおよびロボットの駆動パターン発見のための制約として用いる，制約付きパターン発見手法を提案し，その評価について報告する．音声認識を用いた言語パターンの獲得と，獲得したジェスチャパターン・駆動パターンの認識・生成モデルの構築については今後の課題とする．

提案するパターン発見手法の基盤には[Arita 02]で提案された方法を用いる．この手法では時系列データをその極値を与える点で区切ったのち，点同士のユークリッド距離に従ってノイズを除去した結果として得られるデータのセグメントを Motion unit として保持し，クラスタリングすることによってパターンを発見する．本研究では，Motion unit を抽出する手法にジェスチャの連続性を加味したルールを加えることで拡張し、ノイズに対して頑健なパターン発見手法を提案する．またクラスタリング部には，HMM 同士の kullback leibler 距離を用いた階層的クラスタリングを用いた．評価実験では，狭路を含む迷路をユーザが発話とジェスチャを用いてナビゲーションするタスクを行い，ジェスチャ・駆動パターンの抽出を行った．

2　関連研究

提案システムでは，一連のインタラクション活動を観測し，そのインタラクションデータを訓練データとして利用してロボットの行動制御則を獲得するという観点から，提案システムにおけるロボットの学習方法は，事例からの学習 (Robot learning from demonstration) [Argall 09]の一種と見なすことが出来る．

一連のインタラクション活動はセンサーを通じて連続的な時系列データに変換される．システム側はこの連続的なインタラクションデータから学習対象動作（ロボットの駆動パターンや，人間のジェスチャ）を分節化し，これをモデル化する必要がある．しかし従来多くの動作・運動学習に関する研究では，学習対象の動作が人間の手によって予め分節化されており，学習対象のカテゴリ数も予め与えられていることが指摘されている[Breazeal 02]．

これらの研究に対し[Kulic 09]では, 動作のモデル化にモデルの複雑度を入力データによって変化可能な Factorial-HMM を新規に提案し，階層的クラスタリングと HMM に基づく分節化手法と併用することで，連続動作データから動作の認識・生成モデルを教師無し学習により獲得する手法を提案している．この研究では動作の階層的構造も同時に獲得可能である．

[Kulic 09]の研究では，連続時系列データからの動作パターンの学習・獲得を実現しているが，本研究の目的とするユーザの指示とそれに対するロボットの動作の対の学習・獲得には着目していない．これに対し[Mohammad 09]では，変化点検知アルゴリズムに基づく制約付きパターン発見アルゴリズムを新規に提案し，さらに発見されたユーザのジェスチャ指示パターンとロボットの駆動パターンの時間的因果関係を grager causality analysis を用いて発見する手法を提案した．上記の時系列マイニング手法を，ユーザの自由なジェスチャを用いたナビゲーションタスクに利用し，ユーザ個人に依存するジェスチャによる指示パターンとロボットの駆動パターンの組み合わせを教師無し学習により獲得した．[Mohammad 09]では，インタラクションにおける指示をジェスチャに限定しており，発話による言語指示を用いていない．これに対し本研究では，言語指示パターン，非言語パターンと，それに対するロボットの駆動パターンをインタラクションデータから獲得する手法を提案する．本研究ではジェスチャパターンとロボットの駆動パターン，それらの組み合わせパターンをボトムアップに獲得することを目指している．言語パターンについての情報 (辞書・文法) は，人手により与えるものとする．

3　問題設定

本論文ではロボットのナビゲーションタスクを想定し，以下のように問題を設定した．

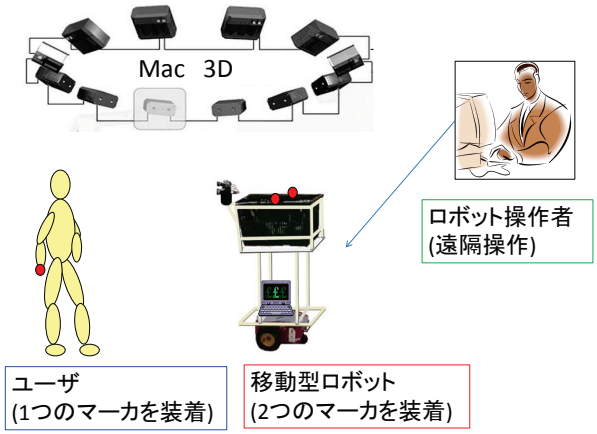

Figure 1: Environment for experimentation

3.1　環境設定

本研究ではユーザのジェスチャを観測するため，Motion Analysis 社のリアルタイム光学式モーションキャプチャである,「MAC3D System」を用いた．このシステムではマーカーが反射する光を複数のカメラで測定することにより，マーカーの三次元座標を 1 秒間に 120 フレーム取得することができる．このマーカーを指示者であるユーザの右腕に 1 つと動作を行うロボットに 2 つ取り付ける．また，指示者に従って動作するロボットとして，Mobilerobots 社の Pioneer3 を基盤とした移動型ロボットを使用する．ロボットが駆動可能な自由度は，向きの変更（左回転と右回転）と前進，後退の三つである．

3.2　インタラクションデータの取得

データ取得環境を Figure 1 に示す．本実験環境ではロボットに指示をするユーザ (以下ユーザと呼称する．) とロボット，またユーザに見えない場所でロボットを操作する操作者の三者によりデータ取得が行われる．ユーザは自由にジェスチャと言語を用いて指示を行い，ロボットを誘導する．ユーザの指示に対して，操作者は遠隔からロボットを操作する．ロボットのナビゲーションタスクを通じて，訓練データはユーザとロボットに取り付けられたマーカの位置座標の多次元時系列データとして収集される．この連続的な時系列データ (インタラクションの履歴データ) から，時系列マイニングを用いてジェスチャ・ロボットの駆動パターンをそれぞれ抽出する．

3.3　本研究で行う音声処理と言語処理

如何なる環境でも音声認識が行えるよう，ロボットにマイクなどの聴覚モジュールを保持させることが最終的に重要であるが，現段階ではシステム内部のアルゴリズムの評価に重点を置くため，出来る限りユーザのクリアな音声を取得することを目指しユーザに接話マイクを装着した．振幅と零交差に基づく入力検知を用いて音声区間の

Table 1: The word set used in the experiment

目的語句	副詞句	動詞句
右に	はやく	いって
左に	ゆっくり	きて
前に	たくさん	まわって
後ろに	すこし	とまって
こっちに	すぐ	さがって

検出を行った．発話された音声認識に関しては今後の課題とし，手動で音声データにアノテーションを付加する．以下の Table 1 に，本実験において使用される単語の一覧を示す．各品詞の単語数種類の組み合わせで数十程度のオーダーとなる．

4 提案システム

本節では観測された連続時系列データよりジェスチャおよび駆動パターンを発見するための方法を述べる．提案するパターン発見手法では，検出された発話区間に基づきセグメンテーション，およびクラスタリングを行いパターンをシンボル化する．

4.1 システムの概要

システムは，連続時系列データからジェスチャおよびロボットの駆動パターンを抽出する．セグメンテーション部分と，抽出されたパターンをクラスタリングする部分とから構成されている．Figure 2 にシステムの流れを示す．

4.2 セグメンテーション部分の実装

まずはシステムのセグメンテーション部について，入力データ，手法，そして全体での処理の流れの順で説明する．

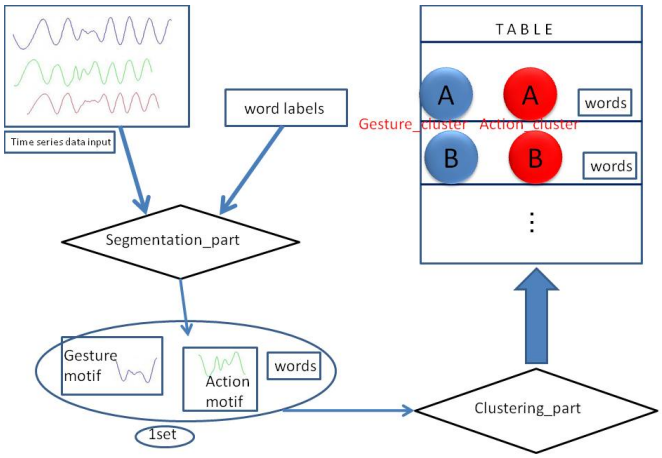

Figure 2: System flow

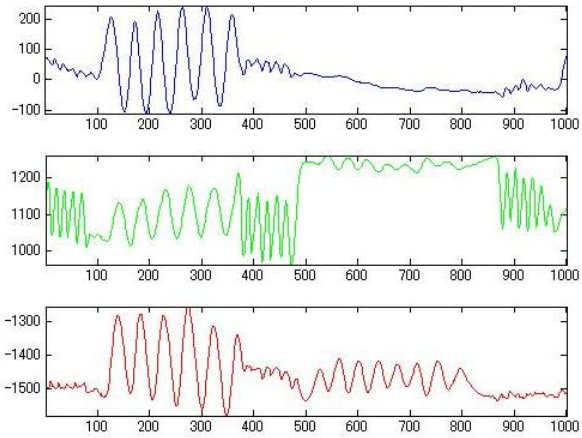

Figure 3: An example of input data

4.2.1 入力データ

システムへの入力は，MAC3D によって取得したマーカの三次元位置座標の時系列データと，言語ラベルである．ここで，ロボットの向きとはロボットに取り付けられたマーカから算出した，ロボットの中心から向いている方向へのベクトルの値のことである．また，言語ラベルとは，言語アノテーションと音声データから自動検出した発話区間である．例として，ジェスチャの三次元時系列データを Figure 3 に示す．上から順に，マーカの x 座標，y 座標，z 座標の軌跡である．

4.2.2 パターンの抽出

本研究では[Arita 02]で提案された手法の一部を基盤として新規の手法を提案した．基盤手法では，時系列データをその極値を与える点で区切ったのち，点同士の距離や値の変化量などの尺度に従って適度にノイズを除去した結果として得られるデータのセグメントを Motion unit と呼称し，これらを Nearest Neighbor 法によってクラスタリングしている．また，Motion unit 同士の類似度は DP マッチングによって計算される．本研究ではセグメンテーションの第一段階として，有田らの手法の一部分である，Motion unit を得る手法を改良したものを用いる．与えられた入力時系列に対して，まずその極値を与える点を求めたのちノイズや微小な振動などを取り除く作業を行う．このとき点同士の距離が近い点と値の変化量の小さい点を単純に取り除いていくだけではノイズが増えたときに対処しきれず，除去すべき点が残ってしまったり除去すべきでない点が除去されてしまったりすることがある．Figure 4 に具体的な事例を示す．

したがって基盤手法を用いた場合，後に Motif の抽出を行う際に影響が出てしまうため，本研究では，ノイズ除去の手法を以下のように改良した．

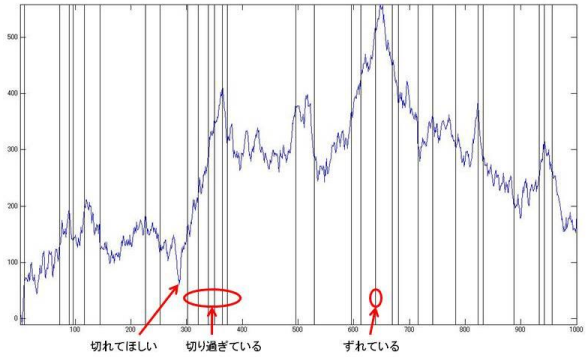

Figure 4: An example of failure case of segmentation

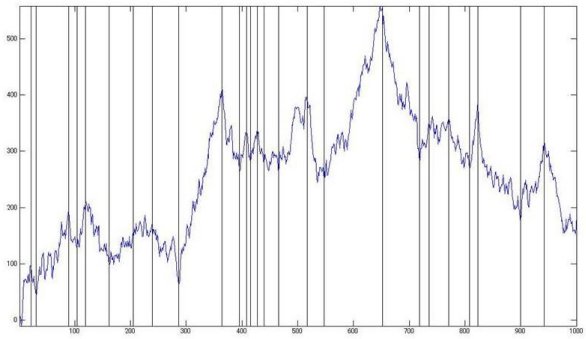

Figure 5: A segmentation result by improved method

- セグメント候補点（すなわち時系列データの極値を与える点）を端から順次見てその場ですぐに破棄するか残すかを判断するのではなく，バッファを用いて候補点を一定量保持する．
- ある点から見て次のセグメント候補点における値が十分に大きく変化しているかどうかの判断を，バッファ内に保持してあるデータでその点に十分近いものと比較する．
- 値の変化量の大きい点を探索の結果，見つけた場合，変化が正方向か負方向かを記憶しておき，さらにその次の大きな値の変化が先の変化とは逆方向であるときのみ，バッファから最小値かあるいは最大値を与える点をセグメント点として選択する．
- 一定時間以上値の大きな変化が観測されない状態が続いたのちに大きな変化に行き当たった場合は，前回の大きな変化点（変化後の点）と今回の変化点（変化前の点）とをセグメント点として選択する．
- セグメント点を選択した時点でバッファを破棄し，再びセグメント点から観測を行う．

以上のジェスチャおよび駆動パターンの抽出に特化した改良を加えることにより，時系列データの分割をさらに適切な点で得ることができる．改良した手法によって Figure 4 で用いたデータを分割した結果を Figure 5 に示す．以下では，この手法を用いたパターン抽出方法を述べる．

まず，上述した手法によって本研究においての入力となる時系列データを分割した様子を，Figure 6 に示す．

ここで，Figure 6 において赤い矢印で示された区間は，その始点と終点との差が閾値以上となっているセグメントの列からなる区間である．連続時系列データの中に発見すべきパターンが埋もれている場合，その周辺では必ず何らかの形でデータの動的傾向に変化が生じているはずであり，また逆に，ある次元に閾値以上の変化があり動的傾向に変化が生じているならばその周辺には何かしらの

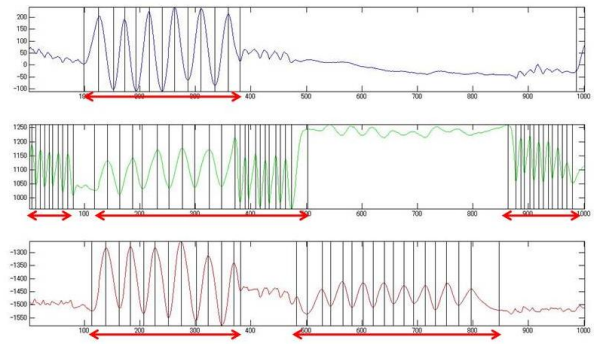

Figure 6: The approach for decision of segmentation unit

モチーフがあると考える．そこで本研究では，Figure 6 の例でいえば赤い矢印で示された区間とそうでない区間で，時系列データを変化の有無に応じて 1 か 0 かの二通りに分け，それらを全次元についてマージした列を考え，それらの共通する区間を一つの Motif であるとして抽出する．Figure 7 に，抽出される区間を状態ごとに分けて示した．

4.2.3 セグメンテーション処理の流れ

一段階目に，本システムでは言語ラベルを受け取るとラベル区間の周辺に何らかのジェスチャパターンが存在すると仮定してラベルの近辺に探索範囲を絞り込んでパ

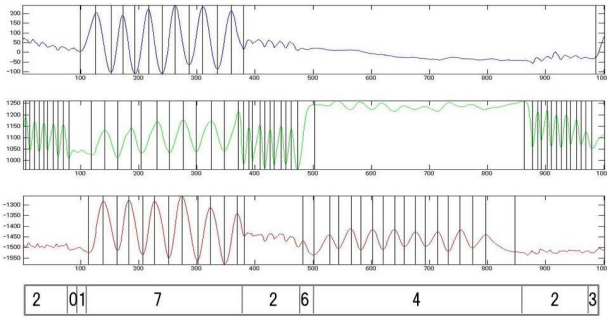

Figure 7: An example of extracted segments

ターンの発見を行う．パターン発見のアルゴリズムの手順を以下に示す．

Step1. 言語ラベルのある区間についてデータの分割を行い，その動的傾向を確認する．

Step2. 傾向に変化がなければデータの分割を行う範囲を左右に広げて Step1 へ．

Step3. 傾向に変化があり，最も長い系列がそうでない系列に挟まれている場合，その系列を抽出する．

Step4. そうでなければ，最も長い系列が伸びている方向へ探索範囲を広げて Step1 へ．

二段階目に，一段階目でジェスチャが抽出されなかった系列データに対して再度上記のアルゴリズムを適用する．

以上のステップを，探索範囲が隣り合う言語ラベルの区間に及ぶまで繰り返す．そうして全ての言語ラベルについてジェスチャのパターンが発見されたら，次にアクションのパターン発見に移る．ジェスチャは言語ラベルのある近辺から探索範囲を左右に少しずつ伸ばしていく方針をとるが，アクションではジェスチャにとっての言語ラベルの役割をジェスチャのある区間に担わせる．また，アクションは初期探索区間から右方向（時間の進む方）へしか探索範囲を広げない．これは指示を受けるよりも早くロボットが適切な動作を取ることは仮定していないためである．構成されるアルゴリズムを以下に示す．

Step1. ジェスチャのある区間についてデータの分割を行い，その動的傾向を確認する．

Step2. 傾向に変化がなければデータの分割を行う範囲を右に広げて Step1 へ．

Step3. 傾向に変化があり，最も長い系列がそうでない系列に挟まれている場合，その系列を抽出する．

Step4. そうでないなら，探索範囲を右へ広げて Step1 へ．

以上のステップを，繰り返し，全てのジェスチャに対し駆動パターンが発見できたら終了する．これを，以降のクラスタリング部分の入力とする．

4.3 クラスタリング部分の実装

4.2.3 で述べたセグメンテーション処理後に得られる，ジェスチャおよび動作の候補である時系列パターン群をクラスタリングする．今回行ったロボットナビゲーションの実験では，上下，左右に手を小刻みに動かすような，ビートジェスチャ[西田 豊明 角 康之 松村 09]が多用された．ビートジェスチャから得られる時系列パターンは繰り返し構造を持つため，セグメントされたジェスチャパターン（時系列パターン）は例えば，五回小刻みに動かしたジェスチャや一回だけ小刻みに動いたジェスチャなど任意の繰り返し構造を持っている．

今回「左に行く」という概念をクラスタとして抽出したいため，これらの繰り返し構造を持つジェスチャパターン間の類似度が高くなるように距離関数を定義したい．本研究ではエルゴード型（全結合型）の HMM をこの問題に対して利用する．エルゴード型の HMM では全状態間の遷移を許すため，上記のような繰り返し構造を持つジェスチャの学習に有用である．

クラスタリングの手法には Ward 法をクラスタ併合の基準とする凝集型階層的クラスタリング法を用いた．以下にセグメントされたジェスチャパターンのクラスタリング手法を述べる．

1. セグメントされた時系列パターン群の数 L だけエルゴード型の HMM を用意する．
2. 各時系列データを各 HMM でそれぞれ学習する．HMM のパラメータ推定には EM アルゴリズムを用いる．
3. L 個の HMM 同士の Kulbuck-Leibler 距離[Rabiner 89]を算出し，$L \times L$ の距離行列 D を作成する．
4. D に基づき階層的クラスタリングを行う．ここでクラス数はパラメータとして事前に設定する．

ロボットの動作パターンに関しても，上記の手法を適用した結果，ジェスチャ同様クラスタリング精度が良好であったため，上記の手法を用いた．階層的クラスタリングにおけるクラスタ数の推定は以下の要領で行う．クラス数 K を 2 から 30 まで変化させた各場合において Ward 法を適用して，Ward 法による距離基準である，クラスタをマージした場合のクラス内分散とマージする前の 2 つのクラスタのクラス内分散の和の差 $\triangle E$ をプロットする．各プロット点を境として対象となるプロット点から前のプロット点集合と後のプロット点集合をそれぞれ線形近似した後，その傾きを求める．この傾きのなす角度が最小となる点のクラス数をクラスタ数として推定する．

5 評価実験

本研究で提案した手法を，実際の実験データに対して用い評価を行う．実験では 1 人のユーザがナビゲーションタスクを 8 分 50 秒間行い，計 63600 フレームの多次元時系列データが得られた．セグメンテーションの結果を Recall, Precision として算出し，Table 2 に示す．ここでセグメンテーションが正解したかどうかの判定について，ビデオ分析とモーションキャプチャのデータから正解区間を決定

し，この区間と抽出したジェスチャ区間が 80% 以上共通していれば正解とした．

Table 2 より，ジェスチャパターンのアクションパターン共に良好な recall, precision の値が得られた．しかしながら発話区間の推定で，息継ぎなどの部分も発話区間として抽出した場合に，その付近にある無意味なパターンを抽出してしまう場合があり，これらを抽出してしまった．

5.1 クラスタリングの評価方法

まず，クラスタリングが正しく行われたかどうかを判断する基準を述べる．

クラスタリング精度の評価には Purity [Manning 07] を用いる．まず Purity は以下の式で算出される，

$$Purity(\Omega, \mathcal{C}) = \frac{1}{N} \sum_{l} \max_{k} |\omega_l \cap c_k| \qquad (1)$$

式 (1) で N は学習データの総数であり，$\Omega = \{\omega_1, \omega_2, ..., \omega_L\}$ はクラスタリング後のクラスタの集合を表し，$\mathcal{C} = \{c_1, c_2, ..., c_K\}$ はジェスチャの正解カテゴリ (クラス) の集合を表す．$|\omega_l \cap c_k|$ は集合 $\omega_l \cap c_k$ に属するデータ数を示す．

5.2 クラスタリング結果

実験によって得られたジェスチャパターンは Come, Forward, Forward2, To left, To right, Stop, Turn left, Turn right の計 7 種類のジェスチャと Forward, Backward, To left, To right, Round clockwise, Round counterclockwise の計 6 種類であった．Figure 2 より，インタラクションデータより無意味なパターンが含まれていた．本システムにおけるクラスタリングの目的は，同じカテゴリのジェスチャ・駆動パターンを異なるクラスタとしてはじくことである．無意味なパターンを実際のジェスチャや駆動パターンと異なるクラスタとしてはじくことである．

以下に階層的クラスタリングした結果を Table 3 に示す．ジェスチャ12 のクラスタの内訳として 7 種類のジェスチャにそれぞれ対応するクラスタが出力された．この内 3 クラスは無意味なパターン集合によるものであったが，一部無意味なパターンが Backward, To left とチャンキングする場面が見られた．これは Backward, To left 内のパターンにセグメント境界が上手く検出できず，無意味な動きとチャンキングしたまま 1 つのパターンとして抽出されたためである．

Table 2: The result of segmentation

	Recall	*Precision*
ジェスチャパターン	0.86	0.90
駆動パターン	0.90	0.95

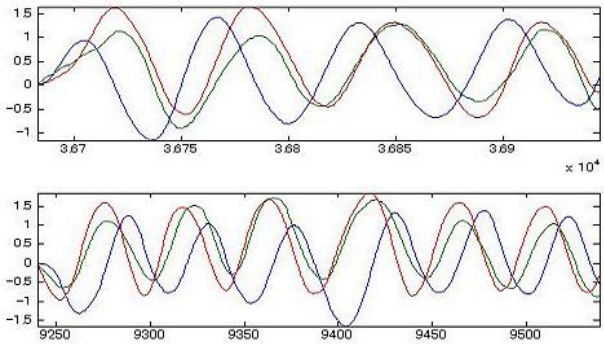

Figure 8: A failure case of clustering (upper shows the gesture pattern which means Turn right, lower shows the gesture pattern which means Turn left)

また Turn left，Turn right に関するクラスタが 2 クラスずつに分かれた上，お互いのクラスタに属するはずのパターンが誤ってチャンキングしてしまうといった場合があった．Turn right クラスのジェスチャと Turn left クラスのジェスチャは似通った傾向を持っている（Figure 8）．Turn right は手を床に平行に時計回りに動かすジェスチャで，turnleft は反時計回りに動かすジェスチャである．このようにセグメントする位置が少しずれるとこれらのパターンについては区別がつかなくなることがわかる．上記の問題に対処するためセグメンテーションの精度およびクラスタリングの精度を改善する必要があるものの，連続動作から発話区間を制約としてパターンを発見する本手法の有効性を示した．

6 結論

本研究ではユーザの言語・非言語指示とロボットの駆動パターンの組み合わせをボトムアップに獲得するため，ユーザの音声発話区間をジェスチャパターンおよびロボットの駆動パターン発見のための制約として用いる，制約付きパターン発見手法を提案した．実験の結果，7 種類のジェスチャを 85% の精度 (purity) で，6 種類の駆動パターンを 88% の精度 (purity) で抽出可能であることを示した．

現段階ではユーザの用いた言語による指示を，人手によりテキストにして書き出しているが，今後この部分を音声認識による自動獲得に移行していく予定である．

Table 3: The result of clustering

ジェスチャ		駆動パターン	
クラスタ数	Purity	クラスタ数	Purity
12	0.85	6	0.88

参考文献

[Argall 09] Argall, B. D., Chernova, S., Veloso, M., and Browning, B.: A survey of robot learning from demonstration, *Robotics and Autonomous Systems*, Vol. 57, No. 5, pp. 469–483 (2009)

[Arita 02] Arita, D., Yoshimatsu, H., and Taniguchi, R.: Frequent motion pattern extraction for motion recognition in real-time human proxy, in *Proceedings of JSAI Workshop on Conversational Informatics*, pp. 25–30 (2002)

[Breazeal 02] Breazeal, C. and Scassellati, B.: Robots that imitate humans, *Trends in Cognitive Sciences*, Vol. 6, No. 11, pp. 481–487 (2002)

[Kulic 09] Kulic, D., Takano, W., and Nakamura, Y.: Online Segmentation and Clustering From Continuous Observation of Whole Body Motions, *IEEE Transactions on Robotics*, Vol. 25, No. 5, pp. 1158–1166 (2009)

[Manning 07] Manning,C.D., Raghavan, P., and Schutze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2007)

[Mohammad 09] Mohammad, Y. F. O., Nishida, T., and Okada, S.: Unsupervised simultaneous learning of gestures, actions and their associations for Human-Robot Interaction, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2537–2544 (2009)

[Rabiner 89] Rabiner, L. R.: A tutorial on hidden markov models and selected applications in speech recognition, in *Proc. IEEE*, pp. 257–286 (1989)

[西田 豊明 角 康之 松村 09] 西田 豊明 角 康之 松村 真宏：社会知デザイン, オーム社 (2009)