6L-8

RNNPB を用いたモダリティ間マッピングによるロボットの動作生成

服部 佑哉 † 駒谷 和範 † 尾形 哲也 † 小嶋 秀樹 ‡ 奥乃 博 †
† 京都大学大学院情報学研究科 知能情報学専攻 ‡ 情報通信研究機構けいはんな情報通信融合研究センター

1. 序　　論

情報処理技術の急速な発展により，近い将来ロボットが一般に広く用いられるようになることが期待されている．それに伴い，実環境で動くロボットも人間と同様に多様なモダリティを用いてインタラクションを行うことが求められている．また，作りこみの動作ではなく，実環境の事象に即した反応を取ることが必要である．しかし，ロボット動作の設計は難しく非常に手間がかかるため，今までは単一のモダリティを扱うことで精一杯であった．

本研究では，複数モダリティを同時に扱うことで，逆に動作の簡便な生成を可能とすることを目的とする．本研究ではこれらの情報を，**モダリティ間マッピング**によって対応付け，表現を生成する．

モダリティ間マッピングの定義

人は普段，現実世界の事象を複数モダリティにおける刺激として知覚している．また，それらの刺激に対する表現を各々のモダリティで表すことができる．例えば，ボールが自分の身体に衝突した事象に対して人間は，以下のような表現が可能である．

- 衝突音を声（擬音語，口真似）で表現する
- ボールの動きを手の動作（ジェスチャー）で表現する
- 衝突の触覚刺激を，相手を触ることで表現する

実環境下では常にすべてのモダリティから適切に情報が得られるとは限らない．そのような場合にも，人間は欠けた情報を連想することができる．このような，得られた刺激から他のモダリティに対する刺激を得るようなマッピングをモダリティ間マッピングと定義する．

2. モダリティ間マッピングによる表現生成

本研究では実環境の情報でも特に，環境音，すなわちドアの開く音やガラスの割れる音といった身の回りの様々な音に着目し，環境音の鳴る事象における視聴覚情報を同時に扱う．自分の身体で物体の動きを表現するすることは，観察者視点のジェスチャー[1] として知られている．また，動きを表現する音は，擬態語として知られている．

2.1 提案するインタラクションモデル

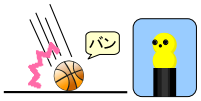


図 1 学習フェイズ

図 2 インタラクションフェイズ (1) 音から動きの連想

システムは学習とインタラクションの 2 つのフェイズを持つ．学習においては，音の発生を感知すると，同時にカメラから物体の動きを抽出し，両者を組として学習する．インタラクションにおいては (1)「音を伝達するために，対象音から適切な動きを連想し自身の身体で模倣し表現する」か，または，(2)「動きを伝達するため，対象の動きから適切な音を連想し，スピーカーから発音する」ことを行う．なお，両フェイズの切り替えは手動で行われる．

Robot Motion Generation with Inter-modality Mapping using RNNPB *by Yuya Hattori, Kazunori Komatani, Tetsuya Ogata (Kyoto Univ.), Hideki Kozima (NICT), and Hiroshi G. Okuno (Kyoto Univ.)*

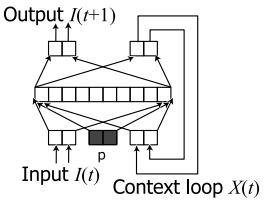


図 3 RNNPB

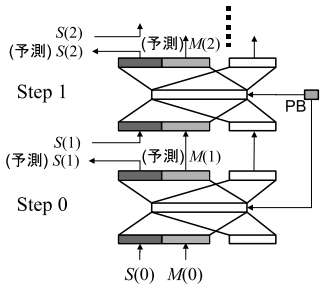


図 4 RNNPB による生成 ($S(t)$ から $M(t)$ を生成)

2.2 学習モデル

実世界に環境音は無数に存在し，それらを全て教示することはできない．したがって少ない学習データからいかにふさわしい動作を生成するかが重要となる．そこで我々は，少ない学習からの汎化能力という点から，谷らの提唱した Recurrent Neural Network with Parametric Bias (RNNPB)[2] を学習器として用いる．

RNNPB は，Jordan 型の RNN に Parametric Bias (以下 PB) を付加した構造をしている（図 3）．PB 値の変更によって 1 つの RNNPB に複数パターンを埋め込むことが容易になり，あるパターンに対応する PB の値を入力することで，希望のパターンを復元させることができる．

PB の内部値 ρ_i はニューロンの重み閾値と同様に，時刻 t 毎に出力誤差から学習信号 δ_t を求め，それらを後ろ向きに伝播させることによって計算される．この内部値はシグモイド関数を通して $p_i = \mathrm{sigmoid}(\rho_i)$ として出力される．本研究においては PB の修正量は式 $\Delta\rho = \epsilon \cdot \sum_t \delta_{i,t}$ によって求められ，学習パターン 1 つにつき全時刻共通の PB 値を持つ．なお，ϵ は学習定数である．

2.3 生成モデル

学習後の RNNPB を予測器として用いることで生成を行う（図 4）．まず，音情報のみ，または動き情報のみのデータを与え，重みを固定した状態で BPTT を適用し，与えられたパターンの予測誤差を最小化するような PB 値を求める．この時，与えられなかった入力次元に対する予測誤差は全て 0 とする．次に，求まった PB 値と与えられたパターン（音情報のみまたは動き情報のみ）を入力とし，入力に用いられなかった次元の値を予測する．

3. システムの実装

3.1 テストベッドロボット

本研究ではロボット “Keepon” を用いて実装，実験を行う．Keepon は 情報通信研究機構で開発された体長約 12 cm のぬいぐるみ型ロボットであり，図 5 に示す 4 自由度からなる動作が可能である．

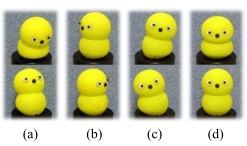
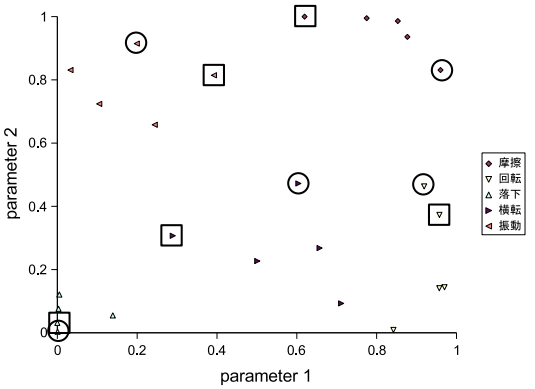


図 5 Keepon の自由度

3.2 学習データ

学習パターンは，以下のとおりにして映像と音から自動的



（○印は音のみからの認識，□印は動きのみからの認識）

図 6 学習データと認識結果の PB 値分布

に切り出される．閾値以上のパワーを持つ音が鳴ると，閾値以上である区間の映像と音が切り出される．その後一定時間以内に再び閾値以上のパワーの音が鳴れば，間の区間を含めて切り出し区間とする．一定時間内に鳴らなければ，そこで音の終わりとする．この手法は，連続した音の発生を全体として一つの音とみなすものである．

RNNPB の入力信号として，音からはメルフィルタバンクの出力を 4 次元を用いる．メルフィルタバンクは人間の聴覚特性に合わせて設計された特徴量であり，本研究に適切であると考える．映像からは対象物体の上下左右端の座標を用いる．なお，物体位置は色を追従して得た．各特徴値は，50 ms ごとのデータであり，区間 $[0.05, 0.95]$ に正規化される．

3.3 生成信号から表現への変換

RNNPB により生成されるのは，抽出された特徴量に対する予測であるため，これを実際の表現とするには，特徴量抽出の逆の変換が必要となる．この変換は以下のように行った．

動き 動きの特徴量は，各時刻 t における物体の上下左右端位置 $(x_{1t}, x_{2t}, y_{1t}, y_{2t})$ である．これを表現にするために，ロボットの動作自由度から適切な 2 動作軸を選び，

$$(\theta_t^x, \theta_t^y) = \big(C_x(x_{1t} + x_{2t}),\ C_y(y_{1t} + y_{2t})\big)$$

となる軌道を描くように動かす（C_x, C_y は係数）．本研究では図 5 (a), (b) の 2 自由度を用いた．

音 音の特徴量は，4 次元のメルフィルタバンク出力である．4 つの値の平均値をパワーとみなし，ホワイトノイズに掛け合わせることで音を生成する．

4. 動作生成実験

4.1 学習させる事象

学習に用いた RNNPB のノード数は，入出力層 8，文脈層 45，中間層 60，パラメータ層 2 である．学習には青い直方体の箱を用い，「箱を擦り付ける」，「箱を倒す」，「箱を落とす」，「箱を振動させる」，「箱を回転させる」様子の映像と音を用いた．各事象について 3 サンプルずつを学習させる．

4.2 学 習 結 果

学習結果の PB 値の分布を図 6 に示す．ただし，図 6 中，丸や四角で囲われていない点が学習結果である．同種のデータに対する PB 値が近接して分布していることが見て取れる．PB 値の分布を見ると，上部に継続的に鳴る音が，下部に単発の減衰音が分布している．

4.2.1 音から動きの生成

学習した RNNPB で，5 種類の事象について，別の試行における音のデータのみを与えて PB 値を計算した．結果の PB 空間を図 6 中の丸印で囲われた点として示す．いずれも学習データの近くに認識されていることがわかる．

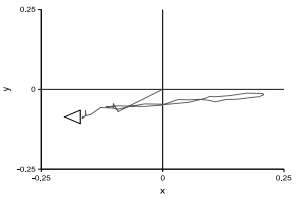


図 7 摩擦音に対して生成された動き

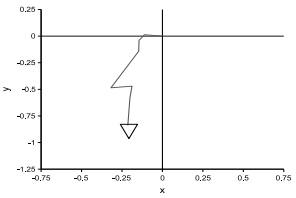


図 8 落下音に対して生成された動き

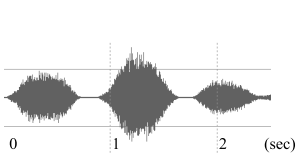


図 9 摩擦の動きに対して生成された音

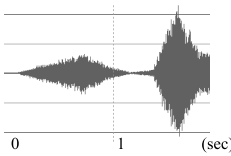


図 10 落下の動きに対して生成された音

また，これらの音に対して動きを生成させた．ここでは一例として，擦り付けた音と落下した音に対して生成された動きを図 7，8 に示す．前者は左右の往復，後者は下方向への移動という元の動きの特徴が再現されている．また，別の音に対して生成された動きも，学習時の動きの特徴を再現するものであった．

4.2.2 動きから音の生成

4.2.1 と同様に，学習した RNNPB で，5 種類の事象について別試行の動きのデータのみを与えて PB 値を計算した．結果の PB 空間を図 6 中に四角で囲った点としてに示す．いずれも学習データの近くに認識されていることがわかる．

また，これらの動きに対して音を生成させた．同様に一例として，擦り付けた動きに対して生成された音を図 9 に示す．これは，継続音が複数回鳴るという学習時の音の特徴を再現している．

結果のうち，落下の音に対して生成された音（図 10）は，学習時とは異なるものであった．後半部に実際には存在しない音が含まれている．これは，落下においては試行ごとに軌道の誤差が大きかったため，正しく連想が行われなかったものと考えられる．

5. 結　　論

本稿では，複数モダリティにおける刺激間のマッピングによる動作生成を提案し，入力音を動きによって，または入力された動きを音によって表現するシステムをロボット Keepon 上で実現した．

今後の課題を以下に挙げる．生成された特徴量から動作への変換は，現在，決め打ちであるが，どの自由度を用いるかを発達的に獲得できるのが望ましい．また，被験者実験などが必要である．

謝 辞 本研究の一部は学振科研費，NICT，21 世紀 COE プログラムの支援を受けた．

参 考 文 献

1) D. McNeill: “HAND AND MIND: What Gestures Reveal about Thought,” University of Cicago Pr., 1992.
2) Jun Tani and M. Ito: “Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment,” IEEE Trans. Sys., Vol.33, No.4, pp.481–488, 2003.