RNNを備えた2体の小型ロボット間の首振り動作と音声によるインタラクションにおける共有シンボルの創発

日下 航† 神田 尚‡ 尾形 哲也‡ 小嶋 秀樹§ 奥乃 博‡
†京都大学工学部情報学科‡京都大学大学院情報学研究科知能情報学専攻§宮城大学事業構想学部デザイン情報学科

1. はじめに

本研究の目的は，実ロボットインタラクションにおいて，形式的記号体系を与えず主に経験による学習から，実世界の動作を指し示す音声シンボルを創発させることである．

ロボットが複雑な実世界でインタラクションを行うには，トップダウンな設計ではなく，環境に合わせてボトムアップに記号を獲得することが必要である．また，記号過程の解明には，記号そのものだけでなく，それが作用し，発展していくような系全体の検証が重要となる．

2体のロボット間で音声と動作の対応関係を共有し，音声によって相手に動作を伝達する時，ロボット間においてこの音声は動作を参照するシンボルとして機能しているとみなせる．これを学習によって実現することを，共有シンボルの創発と位置づける．

それには以下の能力が必要となる．

(1) 実世界の複雑性に対応した，相手の行為の認知能力

(2) 動作と音声のモダリティを対応づけ，その対応を相手に合わせて修正することで共有する能力

(1) で認知した相手の音声を (2) で共有した対応によって動作に変換することで，音声を用いた動作の伝達が可能となり，共有シンボルの創発が実現される．(2) の対応関係修正能力は，(1) を前提とする．

それぞれについて以下のアプローチを採る．

(1) 行動経験を汎化学習して獲得した自己モデルを投影することによって，相手の行為を認知する．

(2) 動作と音声の自己モデル上表現を相互変換することでモダリティの対応づけを行い，この変換構造をインタラクションを通して学習する．

(1) は『他者行為を自己モデル上で認識する』ことであり，『模倣能力を持つ』と言い換えることが出来る．これまで，我々は音声・動作の自己モデルを学習させ，それぞれ模倣という形で (1) の能力の獲得を確認した．

2. 音声・動作インタラクションシステム

本システムのモデルを図1に示す．2体ロボット間で身体動作・音声シンボルの認知・解釈・生成を繰り返し，共有シンボルを創発する．各ロボットは動作理解モジュール，音声理解モジュール，解釈モジュールから成る．

(1) 動作理解モジュール
動作-視覚の自己モデルを核とし，実際の動作時系列とモデル上表現の変換 (動作の認知・生成) を行う．

(2) 音声理解モジュール
声道-聴覚の自己モデルを核とし，実際の音声時系列とモデル上表現の変換 (音声の認知・生成) を行う．

Co-evolution of Symbols for Interaction using RNN with Two Robots' Pan-Tilt Motion and Voice : *Wataru Hinoshita (Kyoto Univ.), Hisashi Kanda (Kyoto Univ.), Tetsuya Ogata (Kyoto Univ.), Hideki Kojima (Miyagi Univ.), and Hiroshi G. Okuno (Kyoto Univ.)*

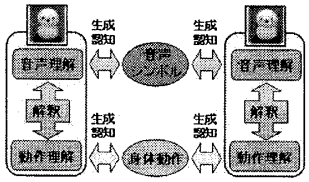

図 1: インタラクションモデル

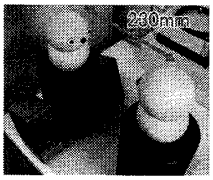

図 2: システム構成

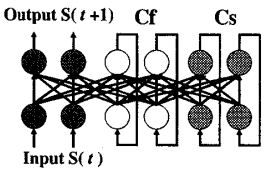

図 3: MTRNN 構成

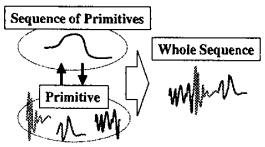

図 4: MTRNN によるダイナミクス表現

(3) 解釈モジュール
動作と音声の自己モデル上表現の相互変換を行う．

本稿では，動作理解・音声理解モジュールの実装を行い，模倣実験によって，これらのモジュールの認知・生成能力を検証した．

2.1 学習器 MTRNN

動作理解，音声理解モジュールの学習には，時系列データの学習・汎化能力とパラメータ空間の自己組織化という観点から，谷らによって提案された Multiple Timescale Recurrent Neural Network (MTRNN) [1] を用いた．MTRNN は複数の非線形時系列パターンを学習・汎化することのできる学習器である．MTRNN は入出力部，Fast Context (Cf)，Slow Context (Cs) と呼ばれる時定数の異なるニューロン群から成る (図3)．Cf が入出力ダイナミクスのプリミティブを表現し，Cs がプリミティブのシーケンスを表現することで，通常の RNN より複雑で長い時系列パターンを学習できる (図4)．学習は Back Propagation Through Time (BPTT) によって行う．また Cs の初期値 (Cs_0) を変更することで異なるパターンを表現でき，Cs_0 のパラメータ空間をデータ間の相関から自己組織的に獲得できる．この Cs_0 が対応する時系列ダイナミクスのモデル上表現となる．認識器として用いる場合，結合重みを固定した BPTT により，Cs_0 のみを更新する．

2.2 システム実装

情報通信研究機構で開発された小型ロボット "Keepon" を用いて，実験システムを構築した．Keepon は眼部に小型カメラ，鼻部にマイクを備えており，4自由度の動作機能を持つ．実験では4自由度の内，PAN 軸 (左右首振り) と TILT 軸 (上下首振り) の2軸を用いた．Keepon 2体を 230mm 離して対面させ，音声再生用に外付けスピーカーを設置した (図2)．

表 1: 実験条件

	動作		音声	
入力次元数	4	モータ 2 視覚 2	14	声道 6 音響 8
入力ステップ数	45 (15step/sec)		51(15step/sec)	
Cf ノード数	30		40	
Cs ノード数	5		7	
学習データ数	70		110	

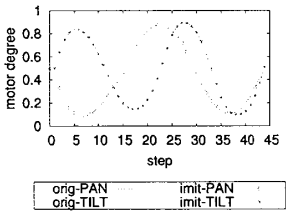

図 5: 動作模倣結果

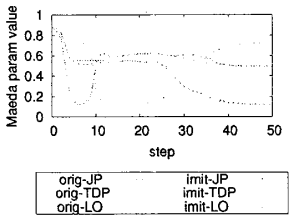

図 6: 音声模倣結果

3. 動作，音声理解モジュールによる模倣実験

本実験で，各モジュールの動作・音声模倣能力を検証した．模倣手順の 3 フェーズ (学習・認識・生成) を以下に示す (以下，キーポン 2 体を A，B で表す).

(1) 学習 (自己身体モデル獲得)
A は自己動作 (モータ/声道) とそれに対して引き起こされる変化 (視野画像/音声) を対応づけて自己モデル MTRNN の結合重み，Cs_0 を学習する．

(2) 認識 (自己身体モデル変換利用による他者理解)
A の自己モデルに，B の動作 (モータ/声道) から得られる情報 (視野画像/音声) を変換入力し，対応する Cs_0 を得る．

(3) 生成 (他者理解結果に基づく動作 / 音声模倣)
認識で得られた Cs_0 から，動作・音声の生成を行う．

3.1 実験 1：動作模倣実験

3.1.1 実験条件 1

実験条件を表 1 左に示す．モータ情報 2 次元 (PAN, TILT)，視覚情報 2 次元 (対面 Keepon の鼻の視野内 x-y 座標) を，それぞれ平滑化・正規化して用いた．本実験では，動作軸の限定と Keepon 位置の固定により，視点変換は A が得た視覚情報の上下左右反転で近似できる．この視点変換は既知である条件の下で認識を行った§．PAN, TILT の値それぞれを正弦関数として，その位相や周波数の組み合わせで学習データ候補の動作を 80 パターン生成した．視覚情報は各動作パターンについて動作中に取得した．これらの動作-視覚データのうち 70 パターンを学習データとした．

3.1.2 実験結果 1

学習済み MTRNN を備えた Keepon に実際に動作を見せ模倣させた．そして元データとの 2 乗誤差を入力次元・ステップ数に関して平均したものを生成エラーとし，模倣精度を評価した．未学習の 10 パターンを含む候補データ全てに関して評価を行ったところ，その平均生成エラーは 0.00153 であった．図 5 に生成エラー 0.00156 の模倣例を示す (4 次元のうちモータ値のみを図示した).

また，獲得されたパラメータ空間には学習候補の動作群とは異なり単純な正弦関数では表せないモータ値推移を持つ多様なパターンが組織化された．例として，(1) 減衰振動，(2) 縦方向に平行移動した正弦波，(3) 直線パターン などが実際に確認された．

3.2 実験 2：音声模倣実験

3.2.1 実験条件 2

本実験では物理声道モデルを用いることで，生物的な拘束条件を付加した．物理声道モデルには Maeda モデル [2] を利用した．実験条件を表 1 右に示す．声道情報は Maeda パラメータの Larynx Position を除く 6 次元，音響情報は MFCC 特徴量の 8 次元 (声道長の影響が大きい低次の項を除く) を，それぞれ平滑化・正規化して用いた．本実験では，A，B の声道モデルを同一としたため，認識フェーズにおける音声変換は行わない．Maeda モデルの母音 4 種 (/a/, /i/, /u/, /e/) から 3 つを連続的に発声させ，母音の順列・変化タイミングの組み合わせで，学習データ候補を 124 パターン生成した．これらのうち，110 パターンを学習データとした．

3.2.2 実験結果 2

学習済みの MTRNN を備えた Keepon に実際に音声を聴かせて模倣させ，動作模倣と同様に生成エラーによって，模倣精度を評価した．未学習の 14 パターンを含む候補データ 124 パターン全てについて評価を行ったところ，その平均生成エラーは 0.00156 であった．図 6 に生成エラー 0.00162 の模倣例を示す (14 次元のうち Maeda パラメータの 'Jaw Position', 'Tongue Dorsal Position', 'Lip Opening' の 3 次元について図示した).

また，獲得されたパラメータ空間には，学習候補の音声群とは異なる特徴を持つ多様なパターンが組織化された．例として，(1) 4，5，6 母音の連続発声に聞こえるパターン，(2) 母音のように聞こえないパターンなどが実際に確認された．

4. おわりに

本稿では，経験による動作・音声の自己モデル獲得，およびそれを用いた模倣を報告した．実機での実験により，動作・音声の認識，模倣におけるモデルの有効性を確認した．さらに MTRNN によって獲得された自己モデルのパラメータ空間には，未経験パターンを含む多様なダイナミクスの土壌が形成されることが確認された．

今後，解釈モジュールの実装とその学習を行う．学習は，一体の Keepon が相手に声を掛け，その反応として返ってきた動作と元の音声を対応づけることで行う．これによって，Keepon 間で音声と動作の対応関係を共有する．このインタラクションの過程では，相手からの刺激がトリガーとなって，パラメータ空間中に埋め込まれた未経験の動作や音声が新たに発現することや，実世界の揺らぎから生じる誤差により，解釈構造の変化と再構築を繰り返す発展的インタラクション創発が期待される．

謝辞　本研究は，科研費学術創成研究，基盤研究 (S)，及びグローバル COE の支援を受けた．

参考文献

[1] Yuichi Yamashita, Jun Tani, "Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: a Humanoid Robot Experiment," PLoS Computational Biology, Vol.4, pp.1-18, 2008.

[2] S.Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," Speech production and speech modeling, Kluwer Academic Publishers, pp.131-149, 1990.

[3] 横矢，尾形ら，"ロボットの順逆モデルの変換による他者行為予測と模倣，" 第 70 回情処全大，pp.413-414，2008.

§視点変換学習を含む模倣手法は横矢らにより提案されている [3]