神経回路モデルによる動作・言語変換を利用した人間ロボット音声協調

張 陽† 尾形 哲也‡ 谷 淳‡‡ 村瀬 昌満‡ 駒谷 和範‡ 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 ‡‡理化学研究所 脳科学総合研究センター

1. はじめに

本研究の目的は，ロボット自身に言語と動作の対応関係を獲得させ，人間と音声を介した自律的協調を行うことである．言語は使われる文脈によって大きく意味を変えうる非常に強力なシンボルであるが，この性質によって生じる記号接地問題のために，ロボットが言語をそのまま扱うのは非常に困難である [1]．杉田らは，ロボットの動作用と言語用の二つの神経回路モジュールを用い，動作と言語の対応を学習している [2]．我々はこの杉田のモデルをベースに，動作シーケンスを予測誤差に基づいて分節化し，対応する文章数に自動的に割り当てる手法を提案した [3]．この手法では，神経回路モデルの汎化特性を利用し，1. 未知文章からの動作生成，2. 未知動作からの複数文章生成を実現した．しかし，これらの検証はシミュレーションのみで行っており，現実のインタラクションにおける有用性の検討は十分に行われていなかった．

本稿では，人間ロボット間の音声協調を設計し，実ロボットに実装し，行った協調実験の例を紹介する．具体的には，提案モデルを実装したロボットが人間との協調実験において 1. 未学習の動作生成，2. 動作中の動作更新，などに対応可能であることを確認した．これは「曖昧な言語表現を許容した人間ロボットの音声協調」のための基盤技術といえる．

2. 人間ロボット音声協調システム

2.1 概要

本システムでは人間とロボットが音声を介して協調作業を行う．ロボットが自ら環境内で試験的に動作する（バブリング）ことで，タスク遂行に必要な身体と環境の構造の記号化を行い，その記号を言語と対応づける．本システムは**神経回路モジュール**を通じて，言語 (動作) 列を認識し，対応する動作 (言語) 列を生成する．この神経回路を通し，4 つのモジュール (動作生成，音声認識，動作追跡，音声合成) を統合している (図 1)．各モジュールの概要は以下の通りである．

音声認識モジュール 音声（日本語）を認識し，言語（英語）及びビット列に変換する．

動作生成モジュール 動作出力をモータの制御信号に変換する..

動作追跡モジュール 頭部カメラで手先を追跡し，画像とモータ角度情報から動作軌道情報を得る．

音声合成モジュール 神経回路モジュールが生成した言語出力を音声に変換する．

本稿では，神経回路モジュール，音声認識モジュール，動作生成モジュールを統合し，人間の音声入力に対してロボットが動作を生成するシステムの動作例を紹介する．

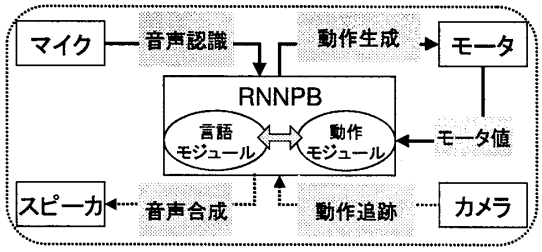

図 1: システムの構成図

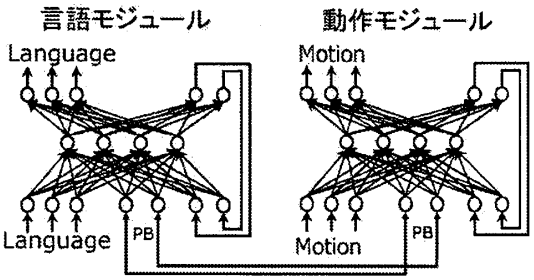

図 2: PB 層がバインディングされた RNNPB の構造

2.1.1 神経回路モジュール：RNNPB

神経回路モジュールには，Recurrent Neural Network with Parametric Bias (RNNPB)[5] を用いた．RNNPB は，現状態を入力とし，次状態を出力する予測器である．RNNPB は PB 層と呼ばれる層を入力層に持ち，PB 層への入力値 (PB 値) を変更することで，異なるシーケンスが生成可能である．また RNNPB を認識に用いることで，認識対象の時系列データを生成するような PB 値を得ることができる．

言語列と動作列の対応関係を学習するために，動作モジュールと言語モジュールの二つの RNNPB を用意し，これらの PB 層をバインディングしたものを用いる (図 2)．これにより，与えられた言語列を言語モジュールで認識し，得られた PB 値を動作モジュールに入力することで，その言語列に対応する動作列を生成することが可能となる．同様に，与えられた動作列を動作モジュールで認識し，得られた PB 値を言語モジュールに入力することで，対応する言語列を生成することが可能となる．

本システムでは，各動作の分節単位を明示的には与えず，動作モジュールの RNNPB が常にモータに制御信号を入力し続ける．

2.1.2 音声認識モジュール

音声認識モジュールには，Julian [4] を用いた．Julian で日本語音声を認識し，それを英語に翻訳，Elman ら [6] と同様に，認識された英単語をビット表現したものに変換する．Julian の認識語彙サイズは 10 であり，使用した音響モデルは性別非依存 PTM モデルである．

2.2 音声協調システム実行手順

本システムは事前に言語列と動作列の対応関係を RNNPB により学習する．学習した動作パターンを表 1 に示す．「黄 → 赤 → 青 → 白 (遅い)」の動作から得られたセンサ・モータデータを動作モジュールに入力する際，この

Human-Robot Voice Co-operation based on Motion-Language Transformation using Neural Network Model *Yang Zhang, Tetsuya Ogata (Kyoto Univ.), Jun Tani (RIKEN), Masamitsu Murase (Kyoto Univ. , presently with Matsushita Electric Industrial Co. Ltd.), Kazunori Komatani, and Hiroshi G. Okuno (Kyoto Univ.)*

表 1: 学習用動作パターン

(1) 黄 → 赤 → 青 → 白
(2) 黄 → 赤 → 青 → 赤
(3) 黄 → 赤 → 黄 → 赤
(4) 黄 → 赤 → 白 → 赤
(5) 黄 → 青 → 赤 → 黄
(6) 黄 → 青 → 白 → 黄
(7) 黄 → 青 → 黄 → 赤
(8) 黄 → 青 → 白 → 赤

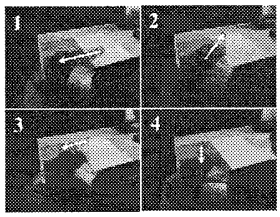

図 3:実際の動き

動作に対応する文章 ("Move to red slowly. Move to blue slowly. Move to white slowly.") を言語モジュールに入力し学習を行った. 音声協調手順は以下の通りである.

1) 人間の発話（日本語）を音声認識する.
2) 認識結果から言語モジュール用 RNNPB で PB 値を求め, 動作モジュール用 RNNPB で対応する動作列を出力する.
3) 動作出力をモータ指令値に変換・動作する.
4) 1) から 3) を繰り返す.
5) 音声入力で "終わり" 受理すれば終了.

システム実行中は, 音声認識モジュールは常に言語の受理を行い, 新しい発話が入力されると対応する PB 値を求め, 動作モジュール用 RNNPB の PB 値を更新する.

3. 音声協調実験

3.1 実験条件

実験には, 人間型ロボット Robovie-IIs を用いた. 対象のタスクは卓上 4 色領域間の腕移動とし, 頭部 3 自由度のうち pitch 軸と yaw 軸, 左腕 4 自由度のうち, 肩の roll 軸と肘の pitch 軸, 及び頭部 CCD カメラを用いた. 動作モジュール用 RNNPB への入力として, 用いた関節の角度 (4 次元), CCD カメラ画像中の各色の割合 (4 次元) を 0.1 秒毎に取得し, [0,1] に正規化したものを用いた.

動作モジュールの RNNPB の各層の数は, 入出力層 8, 中間層 50, 文脈層 8, PB 層は 4, 言語モジュールの RNNPB の各層の数は, 入出力層 9, 中間層 20, 文脈層 9, PB 層は 4 とした.

3.2 実験

3.2.1 実験 1 : 未学習の動作の生成

未学習の動作・言語関係に対し, その関係を推測可能かどうかを検証する. 本実験では, 文章 1) move to red slowly, 2) move to white slowly, 3) move to blue fast, 4) move to red slowly, を順に発話した. 1), 4) は 2 つの動作速度を両方学習済, 2) は速い場合のみ, 3) は両方ともに学習したことがない動作である.

3.2.2 実験 2 : 動作中の動作更新

ロボットが動作している途中で, 新たな言語を与えられたときに, それに対応して動作更新可能かどうかを検証する. 動作中に新たな発話に対し, 対応する PB 値を求め, 動作モジュール用 RNNPB の PB 値を切り替える. 本実験では, 文章 1) move to red slowly, 2) move to blue slowly, 3) move to white fast, 4) move to red fast, を順に発話した.

3.3 実験結果・考察

3.3.1 実験 1

実際ロボット動きの写真は図 3 に示す. 軌道のグラフを図 5 に示す. 未学習の 2) 及び 3) の言語から正しい動作を生成できていることがわかる. 1, 4 番目の PB 値を

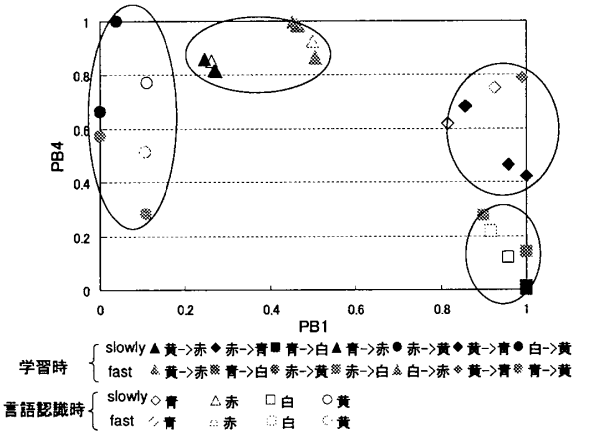

図 4: PB 空間

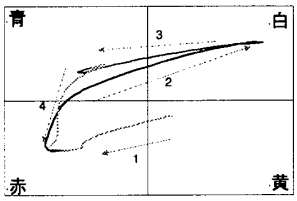

図 5: 実験 1 の動作軌道

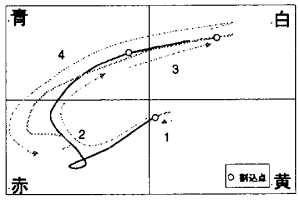

図 6: 実験 2 の動作軌道

動作の目標色で分類した状況を図 4 に示す図 4 から, PB 値は目標色によって分類していることがわかる. "move to blue" の PB 値は, 他の PB 値と完全に分離されており, 青色までの動作は, より正確に生成しやすいと考えられる. 一方, 赤と黄の PB 値 (slowly) は近接しており, 動作生成した際に誤動作を引き起こすことが確認された.

3.3.2 実験 2

結果は図 6 に示す. 動作の途中で, 発話割り込みによって新しい動作が入力された場合も, 正しい目標色まで移動できることが確認された. 動作 1) は 7step, 動作 2) は 17step, 動作 3) は 7step で, 発話割り込みを行った.

4. おわりに

我々が提案した RNNPB による動作・言語変換モデルを実ロボットシステムに実装し, 音声協調システムのプロトタイプを開発した. 基礎実験の結果, 開発システムでは未学習の動作・言語関係を, RNNPB の汎化能力によって補間でき, 動作中の発話の割り込みにも柔軟に対応可能であることを確認した. 今後はより複雑なタスクに発展させていく予定である.

謝辞 本研究の一部は, 科研費, グローバル COE の支援を受けた.

参考文献

[1] S. Hamad.: "The symbol grounding problem", *Physica D*, Vol. 42, pp.35–346, 1990.
[2] Y. Sugita, *et al.*: "Learning Semantic Combinatoriality from the Interaction between Linguistic and Behavioral Processes", *Adaptive Behavior*, 13, 1, pp.33–52, 2005.
[3] T. Ogata, *et al.*: "Two-way Translation of Compound Sentences and Arm Motions by Recurrent Neural Networks", *Proc. IEEE/RSJ IROS*, pp.1858-1863, 2007.
[4] 河原達也, 李晃伸: "連続音声認識ソフトウエア Julius", 人工知能学会誌, Vol. 20, No. 1, pp.41–49, 2005.
[5] J. Tani, *et al.*: "Self-Organization of Behavioral Primitives as Multiple Attractor Dynamics: A Robot Experiment", *IEEE Trans. on SMC PartA: Systems and Humans*, Vol. 33, No. 4, pp.481–488, 2003.
[6] J. L. Elman: "Distributed representations, simple recurrent networks, and grammatical structure", *Machine Learning*, 7, pp.195–225, 1991.