RNNを備えた2体のロボット間における身体性に基づいた動的コミュニケーションモデル

○日下航† 尾形哲也† 小嶋秀樹‡ 高橋徹† 奥乃博†
(†京都大学大学院情報学研究科 ‡宮城大学事業構想学部)

Model of Dynamic Communication dependent on Body Dynamics between two Robots with RNN

*Wataru Hinoshita†, Tetsuya Ogata†, Hideki Kozima‡,
Toru Takahashi†, and Hiroshi G. Okuno†
(†Graduate School of Informatics, Kyoto Univ. ‡School of Project Design, Miyagi Univ.)

Abstract— We propose a model of evolutionary communication with voice signs and motion signs between two robots. In our model, a robot recognizes other's action reflecting its self body dynamics by a Multiple Timescale Recurrent Neural Network (MTRNN). Then the robot interprets the action as a sign by their own hierarchical Neural Network (NN). Each of them modifies their interpretation by re-training the NN to adapt the other's interpretation throughout interaction between them. As a result of the experiment, we found that the communication kept evolving through repeating miscommunication and re-adaptation alternately, and induced the emergence of diverse new signs that depend on the robots' body dynamics through the generalization capability of MTRNN.

Key Words: Communication, Body Dynamics, Robot, Neural Network, Sign, Cognition

1. はじめに

社会性を持ち自在にコミュニケーションするロボットの実現は, ロボティクスの究極的な目標の一つである. それには, コミュニケーションの本質に関する洞察と, ロボットへの実装という工学的視点, これら両面からの取り組みが必要不可欠である.

コミュニケーションとはサイン (ジェスチャ, 目くばせ, 言語など) を用いて, 他者とやり取りすることである. 記号論の創始者パースはサインを, 表現と指示内容の静的な対応関係でなく, 表現が解釈され指示内容に結びつく動的な過程（記号過程）として定義している [1]. また, 社会学の流派であるシンボリック相互作用論では, 個々人は独自の解釈を持ち, それは他者との相互作用の過程から発生するとしている [2]. こうした観点から見たとき, コミュニケーションとは, 他者とのサインのやり取りであると同時に, サインを生み出しその解釈を共有しようという社会的な営みであるということになる. また, こうしたコミュニケーションでは, サインは主体間の解釈の齟齬と相互適応作用によって, 崩壊と再自己組織化を繰り返しながら, 力動性を持って変化し続ける.

コミュニケーションを考える上で, 外してはならないもう一つの視点が身体性に基づいた認知である. ミラーニューロンの発見によって, 他者の行為の認知には自己の身体運動を参照していることが明らかになってきた. これは, 他者と自己の行為を対応付ける能力であり, 共感能力ひいてはコミュニケーションの源泉となる機能である.

本研究では, 上に述べたコミュニケーションの本質的特徴をモデル化し, ロボットに実装してその振る舞いを調べる. これは, コミュニケーション能力のロボットへの実装方法の検討であると同時に, 構成論的手法によるコミュニケーションの解明プロセスでもある. コミュニケーションの力動性に着目し, 構成論的な手法を用いて検証した例に, Steels[3] や橋本 [4] らの研究がある. これらは, 既に身体を離れて存在する対象や記号を組み合わせ, 意味付ける過程を主に扱っている. それに対し我々が扱うのは, 身体性に密着した声と動作による, より原初的なコミュニケーションである. ここでは, 意味を持ったサインとしての音声や動作のパターンが, 主体間の相互作用によって, ロボットの身体性から新しく生み出され発展していく過程が対象となる.

2. コミュニケーションの概要とモデル

2·1 コミュニケーション概要

我々が対象とするコミュニケーションは, 2体のロボットが互いに音声パターンを用いて, 意図する動作パターンを相手に伝えようとするものである. 発話者は自分が伝えたい動作を表わす音声を発し, 聞き手はその音声を解釈して動作を返す. この時, 音声は動作を指示するサインとして働いているが, その解釈は2体が独自に獲得してきたものなので, 齟齬が生じる. そこで, 発話側は, 相手が返してきた動作を見ることで, 相手の解釈を推測し, 自己の解釈をそれに合わせて変容させる. この一連のやり取りを役割を交代しながら繰り返す. これは, 相互作用のなかから, 相手と共通の解釈を作り上げようとするシンボリック相互作用論的なコミュニケーションのモデル化と見ることができる.

2·2　モデル概要

Fig.1 にモデルの概要を示す.

各ロボット (Fig.1 太黒枠) は, サインの解釈に用いる 2 個 1 組の 3 層ニューラルネットワーク (解釈 NN, Fig.1“Agent Robot”内白枠) と, 行為の認識生成を行う順逆力学モデル Multiple Timescale Recurrent Neural Networl(MTRNN) を音声用・動作用に各一つずつ持つ (Fig.1“Agent Robot”内青枠, 赤枠).

音声用 MTRNN は, 音声パターンと音声用 MTRNN のパラメータ (音声表象ベクタ) を相互に変換する. 動作用 MTRNN は, 動作パターンと動作用 MTRNN のパラメータ (動作表象ベクタ) を相互に変換する. 解釈 NN は, 音声表象ベクタと動作表象ベクタを相互に変換する.

例えば, 音声を聞いて連想した動作を返す過程はこのモデルでは次のようになる. (1) 認識：音声用 MTRNN で音声を音声表象ベクタに変換する. (2) 解釈：解釈 NN によって, 音声表象ベクタを動作表象ベクタに変換する. (3) 生成：動作用 MTRNN で動作表象ベクタを動作に変換する.

これは, パースの記号論における記号過程のモデル化と見ることができる.

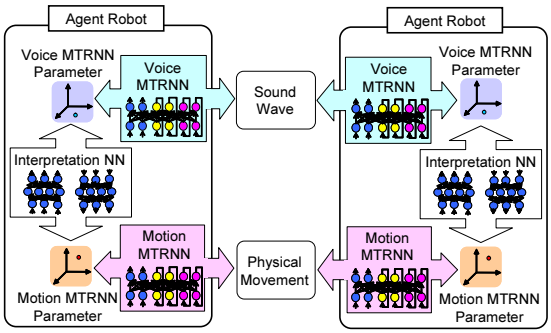

Fig.1 モデル概要図

2·3　MTRNN を用いた行為の認知・生成

コミュニケーションの成立には, 自己の身体性に基づいた他者行為認知（ミラーシステム）が必要である. 本研究では, 行為の知覚情報と運動情報を 1 つの力学系として MTRNN に学習させることで, これを実現する (知覚-運動統合フレームワーク). 知覚情報と運動情報とは, 音声においてはそれぞれ聴覚情報と声道の動かし方に対応し, 動作においてはそれぞれ視覚情報と体の動かし方に対応する.

2·3.1　力学モデル：MTRNN

MTRNN は山下ら [5] によって提案されたモデルであり, 時系列データの学習に優れ, 複数の学習パターンを汎化してパラメータ空間を自己組織化する能力を持っている. MTRNN は, ある時刻の状態 $S(t)$ から次の時刻の状態 $S(t+1)$ を予測するという形で, 時系列データを表現する. MTRNN は入出力部 (IO), 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 (Fig.2). 時定数が大きいほどニューロンの変化は緩やかになり, 抽象度の高い情報を扱うようになる. 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, 従来のモデルより複雑で長い時系列パターンを学習できる (Fig.3). 学習は Back Propagation Through Time (BPTT) によって行う. また, MTRNN は Cs の初期値 (Cs_0) をパラメータとし, これを変更することで異なるパターンを表現できる. Cs_0 のパラメータ空間は学習時にデータ間の力学的相関から自己組織的に獲得される. 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.

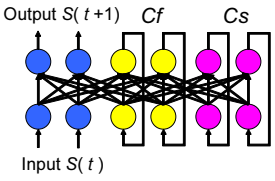

Fig.2 MTRNN の構成

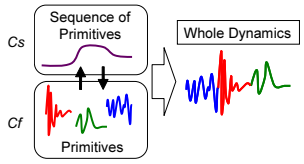

Fig.3 MTRNN による ダイナミクス表現

2·3.2　知覚-運動統合フレームワーク

知覚-運動統合フレームワークにおける行為の学習・認識・生成は以下のように行われる (Fig.4).

1. 知覚-運動統合自己モデルの学習：ロボットは運動情報の時系列データから行為を生成し, その結果をセンサーで知覚する. 運動情報と, 得られた知覚情報を MTRNN に同時に入力し, 1 つの力学系として学習させる. 複数の行為パターンに関してこれを繰り返すことで, 行為パターン間の力学的構造を反映した MTRNN のパラメータ空間が自己組織化される. このパラメータ空間には, 学習経験を汎化する形で, 自己の身体の力学特性を反映した多様な未学習パターンが埋め込まれている.
2. 自己モデルの投影による他者行為認識：他者行為をセンサーで知覚する. 得られた知覚時系列データを自己の MTRNN に入力し, 重み固定の BPTT によって, その行為パターンに対応した MTRNN パラメータ Cs_0 を得る. この時, 運動情報には MTRNN から出力される予測値を入力しておく.
3. 生成：Cs_0 を MTRNN にセットし, 前向き計算をすることでデータを時系列順に生成していく. この時, 各時刻での入力には, 1 ステップ前の出力を用いる. これによって, Cs_0 に対応する運動情報時系列（と知覚情報時系列）が得られる.

2·4　NN の追加学習による解釈の共有

各ロボットは, 解釈 NN を用いて, 動作表象ベクタと音声表象ベクタを相互変換することでサインの解釈を行う. 2 体の間でサインの解釈を共有するためには, 解釈 NN を他者に適応的に変化させなければならない. しかし, 他者の解釈を学習する度に, 解釈 NN の性質が完全に変わってしまうのでは、相互に相手に適応すること自体ができない。つまり, 自己の解釈は学習の前後で全体的な一貫性が保たれる必要がある. そこで, 可能な限り一貫性を保ちながら NN を再学習する手法として, Consolidation Learning を用いる. これは以下の 4 ステップからなる.

1. 声掛けに対する相手の反応動作から, その音声に対する相手の解釈 (音声ー動作ペア) を得る.

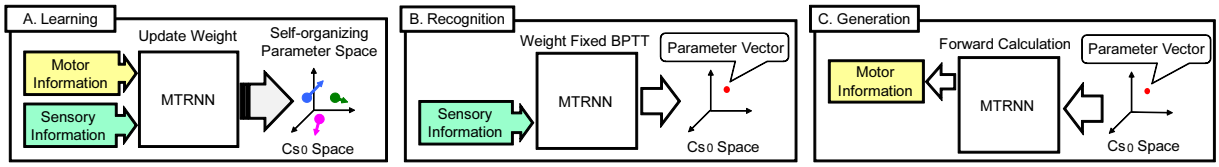

Fig.4 MTRNN による知覚-運動統合フレームワーク

2. 現時点での自分の解釈群を保存する．具体的には，解釈 NN の入力空間に格子点状の入力を行い，その出力とペアにして保存しておく．
3. （2）で得た現在の解釈群（音声-動作ペア）のうち，（1）で得た他者の解釈と競合するものを除く．競合するとは，音声または動作が似ている（時系列データでの 2 乗誤差が小さい）ものを指す．
4. 他者の解釈，及びそれと競合しない自分の解釈群をデータセットとして NN を再学習する．

3. 実験

3·1 実験システム

実験プラットフォームには，小型ぬいぐるみロボット “Keepon”[6] を用いた．Keepon は，眼に小型 CCD カメラを，鼻にはマイクをそれぞれ備えている．また，動作に関しては 4 つの自由度を持つ．実験にはこのうち PAN と TILT の 2 自由度 (Fig.5) のみを用いた．実験システムは，230mm 離して対面させた Keepon 2 体と，Keepon の側面のスピーカから構成した (Fig.6)．

Keepon の声の合成には，人間の物理声道モデルである Maeda モデル [7] を用いた．Maeda モデルは，顎や舌の位置・形状などを表わす 7 つのパラメータによって声道形状を決定し，それに対応する音声を合成することができる．実験では，喉頭位置を表すパラメータを除いた 6 つのパラメータを扱った．Maeda モデルの導入によって声道運動が扱えるようになり，音声モダリティにも知覚-運動統合フレームワークが適応可能となる．

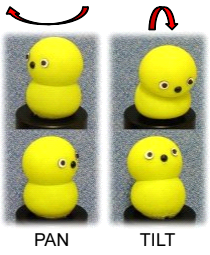

Fig.5 動作軸

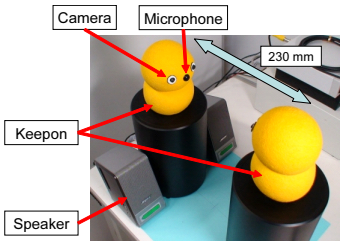

Fig.6 ハードウェア構成

3·2 実験フェーズ

実験は，知覚-運動統合自己モデルの獲得 (Fig.7) と，サインの共有化 (Fig.8) の 2 フェーズで行われる．

3·2.1 フェーズ 1：知覚-運動統合自己モデルの獲得

動作 MTRNN の学習において，運動情報には Keepon のモータ値時系列 (PAN, TILT) を用い，知覚情報には眼部のカメラでとらえた視野画像の特徴量時系列を用いた．画像特徴量には，静止した基準点の視野内座標 (x, y) を用いる．今回はこの基準点を，対面する相手 Keepon の鼻の重心とした．このようにすると，自分が静止していて相手 Keepon が動いた時に得られる画像特徴量時系列データは，自分自身がそれと同じ動きをして相手が静止していた時の時系列データを反転させたものとなる．このため，他者行為を認識するためには，観察して得た時系列データを，MTRNN に入力する前に反転させてやる必要がある．これは，簡易的な視点変換システムになっている．今回この変換自体は既知であるとした．

音声 MTRNN の学習において，運動情報には Maeda モデルのパラメータ時系列 (6 次元) を用い，知覚情報には Maeda モデルから生成された音声の音響特徴量時系列を用いた．音響特徴量には，MFCC(Mel-Frequency Cepstrum Coefficient) の 3〜10 次元目を用いた（サンプリング周波数:16000Hz，フレーム長:25msec，フレームシフト:10msec，フィルタバンク:24）．

3·2.2 フェーズ 2:サインの共有化

Keepon 間の解釈の共有化は以下の手順で行われる．以降，2 体の Keepon をそれぞれ **A**, **B** と表記する．

1. **A** が **B** に話しかける．
2. **B** は音声を認識し，自身の解釈に従って連想した動作を生成して返す．
3. **A** は **B** の返した動作を認識し，**A** の最初の音声と **B** の反応動作が対応づくように，解釈 NN を再学習する．
4. 役割を入れ替えて，1 に戻る．（この時，発話のトピックは前のインタラクションを引き継ぐ）

3·3 実験設定

MTRNN のノード数は以下のとおりである．

- 動作 MTRNN：$IO = 4, C_f = 20, C_s = 3$
- 音声 MTRNN：$IO = 14, C_f = 25, C_s = 3$

また，入出力ノードの時定数は 2，C_f の時定数は 5，C_s の時定数は 10000 とした (音声・動作共通)．

動作 MTRNN の学習用には，PAN と TILT のモータ値時系列を正弦波で表現したデータを作成した．正弦波の周波数や位相を変えることで，24 パターンの運動データを用意し，そのうち 20 パターンを学習させた．

音声 MTRNN の学習用には，Maeda モデルのパラメータ時系列を母音/a/, /i/, /u/, /e/間の遷移パターンとして表現したデータを作成した．16 パターンの声道運動パターンを用意し，12 パターンを学習させた．

未学習パターンは汎化能力の確認に用いた．

4. 結果

実機による自己モデルの獲得後，3·2.2 節のやりとりをコンピュータシミュレーション上で 3000 ターン行った．実験の結果を Fig.9 に示す．図の最上段のグラフは各ターンにおける **A** と **B** の意図伝達誤差を示している．意図伝達誤差は，発話者が発話した際に本来意図していた動作と，聞き手が返してきた動作との誤差を，時

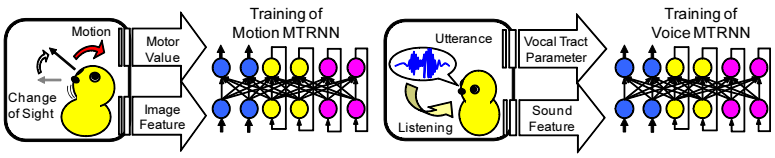

Fig.7 フェーズ 1: 知覚-運動統合自己モデルの獲得

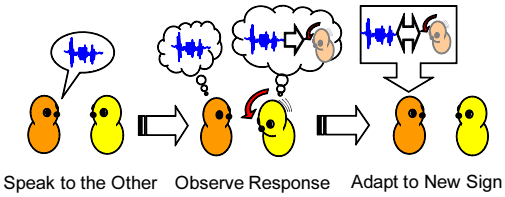

Fig.8 フェーズ 2: サインの共有

系列データ間の平均二乗誤差として評価したものである．その下には, 領域 I, II, III においてやりとりされた音声と動作の時系列パターンをプロットしてある．

この実験結果から, 以下のことが分かる．インタラクションはエラーの小さい安定状態とエラーの大きい不安定状態を繰り返す．安定状態では, 2 体のロボットは解釈の共有により正しく意図を伝達でき, 似たような音声と動作を用いたやりとりが安定して続く (Fig.9 の領域 I, III)．一方, 不安定状態では, 意図の伝達に失敗するため, 不規則な振る舞いが繰り返される (Fig.9 の領域 II)．また同じ安定状態でも領域 I でやりとりされるサイン (音声, 動作のパターン) と, 領域 III でやりとりされるサインは異なっている．さらに, これらのサインは, 知覚-運動統合自己モデル獲得のために用いた学習用パターン (3·3 節参照) とも全く異なっている．これらは, ロボットの身体性と学習経験, およびロボット間の相互作用から創発された新しいサインである．ターンが進むにつれ, 意図伝達誤差は全体として減少する傾向にあるが収束することはなく, コミュニケーションは発展し続ける．

5. 結論

本稿では, サインがロボット間の相互作用から創発され, 発展していく動的なコミュニケーションのモデルを提案した．モデルでは, コミュニケーションの基盤となる身体性に基づいた他者行為理解の枠組み（ミラーシステム）を, MTRNN を用いた知覚-運動統合フレームワークで実現した．また, サインの解釈を NN で表現し, 異なる主体間で相互適応的に解釈 NN の追加学習を繰り返すことで, 他者と解釈を共有しようとする社会的なコミュニケーションをモデル化した．実験の結果, コミュニケーションは全体として成功の度合を高めつつ, 齟齬の発生と解釈の共有を繰り返しながら, 発展し続けることが観察された．この過程で, ロボットは自身の身体性と過去の学習経験から, 新たなサインを次々と生み出していくことも明らかになった．結論として我々のモデルにおいて, コミュニケーションにおける動力学的特性が再現されていると言える．

本研究では, 物理的な現象をそのまま意味づけていくということを行ったが, ある程度以上高度なコミュニケーションを行うには, 現象の分節化と要素の組み合わせという能力が必要不可欠である．こうした能力をいかにモデル化し獲得させるかが今後の課題である．

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．

[1] C. S. Peirce et al, “Collected Papers of Charles Sanders Peirce,” *Harvard Univ. Press*, 1935.
[2] H. Blumer, “Symbolic interactionism: perspective and method, ” *Univ. of California Press.* , 1986.
[3] L. Steels, “The synthetic modeling of language origins,” *Evolution of Communication*, 1997.
[4] T. Hashimoto, “The Constructive Approach to the Dynamic View of Language,” *Simulating the Evolution of Language*, 2001.
[5] Y. Yamashita and J. Tani, “Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: a Humanoid Robot Experiment, ” *PLoS Comput. Biol.* , vol. 4, 2008.
[6] H. Kozima, C. Nakagawa, and H. Yano, “Using robots for the study of human social development, ” *AAAI Spring Symposium on Developmental Robotics*, 2005.
[7] S. Maeda, *Speech production and speech modelling*, Kluwer Academic Publishers, pp. 131-149, 1990.

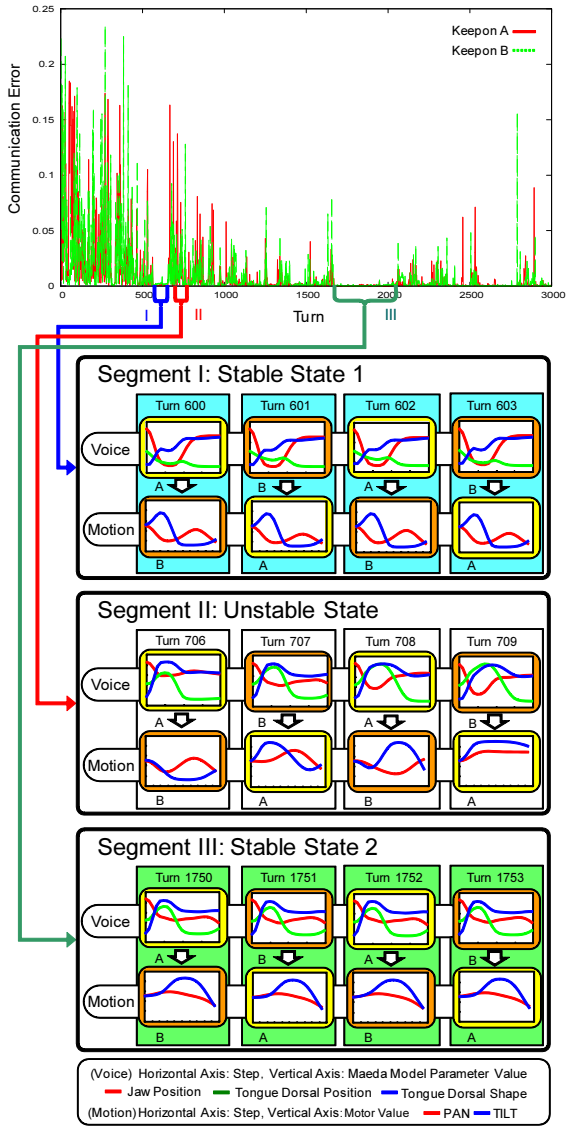

Fig.9 実験結果：グラフは意図伝達誤差を, その下は各領域でやりとりされた音声・動作パターンを示す