C-17　卓球コミュニケーションロボットにおける戦略学習への一考察
An Exploratory Study of Strategy Learning for Ping-Pong Communication Robot

坂本 覚仁 †　　松原 崇充 †　　木戸出 正繼 †

Kakunin SAKAMOTO　Takamitsu MATSUBARA　Masatsugu KIDODE

Abstract: Ultimate goal of our research is to develop a ping-pong communication robot towards natural interaction with human through ping-pong task in real environment. In order to achieve such natural interaction with human, recognizing player's individuality for adapting the strategy of robot would be required, e.g., strategy learning. In this paper, we present an exploratory study of the strategy learning for ping-pong communication robot by the standard linear optimal control framework. In simulation, we first model two kinds of simple human's ping-pong strategies. For both settings, we then show that the standard linear optimal control with the least square system identification technique can be fit for the strategy learning of robot. As a result, it is demonstrated that the optimal feedback controller achieves long term rally for both simple models of human. At the end of this paper, our future work is discussed.

1　はじめに

ヒトと機械とが協調するヒト支援システムでは，ヒトに複雑な操作を強いることなく，ヒトの意図に沿った支援や作業を行うことが求められる．このような知能ロボットはヒトとコミュニケーションを取り，意図を理解するインタラクション指向のロボットであるといえる．これまでもぬいぐるみロボット [1] や対話を行い行動するロボット [2] などインタラクション指向ロボットの実現を目指した研究が行われている．

我々は，ヒトの多様な動作や行動にロボットが適応的に振る舞う機能の実現に向け，卓球タスクを題材とした研究開発用プラットフォーム「卓球コミュニケーションロボットシステム」を開発している（Fig.1）．卓球タスクでは球や卓球台などの環境モデルは不変であるが，ヒトの個性の多様性や，状況に応じて切り替わる戦略モデルなどに対してロボットが適応的に振る舞う機能が求められる．このような機能の実現こそが，ヒトとロボットが共生する社会には必要不可欠であると考えられる [3]．従来，卓球ロボットの開発は知能ロボットの研究 [4, 5] として進められてきたが，ヒトとのコミュニケーションを目的として開発されたものは少ない．

一方，卓球タスクの簡略版といえるエアホッケー課題において観察によるエアホッケーのデータベース化と，獲得したデータベースを環境にあわせて切り替え

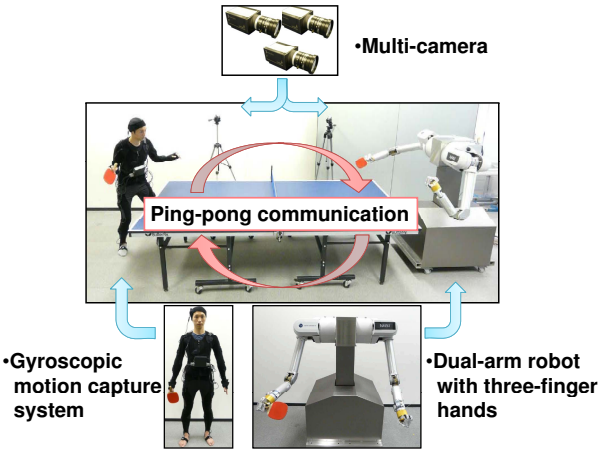

Fig. 1: Ping-pong communication robot system.

ることによる戦略学習を行った Bentivegna らの研究 [6] や，小型ヒューマノイドと他者のインタラクションパターンを設定した上で，身体的インタラクションを線形の力学系モデルで近似し，ロボットの運動から他者にさせたい運動を誘導できることを示した高野らの研究 [7] がある．これらの研究を踏まえ，本研究ではプレイヤー対ロボットの卓球タスクにおいてプレイヤーが取る戦略をモデル化することにより，プレイヤーの特徴に応じたロボットの戦略学習を目指す．状態予測に線形の力学系モデルを，報酬予測に２次モデルを用いる場合，それらに関して最適な制御則を簡便に求めることができるため，卓球タスクにおけるインタラクションが局所的に線形の力学系モデルで表せるようにモデル化し，戦略学習を行う．本論文ではその基礎的

† 奈良先端科学技術大学院大学情報科学研究科，
Graduate School of Information Science, Nara Institute of Science and Technology
E-mail:[kakunin-s, takam-m, kidode]@is.naist.jp

研究として，一般的と考えられる 2 種類の卓球戦略を設定し，プレイヤーがその戦略を用いたときのインタラクションが線形の力学系モデルとして表現でき，制御可能であることをシミュレーションによる実験により示す．

2　卓球タスクについての考察

実システムである卓球コミュニケーションシステムを，シミュレーションで取り扱うために，卓球タスクのモデル化とプレイヤーの行う戦略のモデル化を行う．以下にその詳細を示す．

2.1　卓球タスクのモデル化

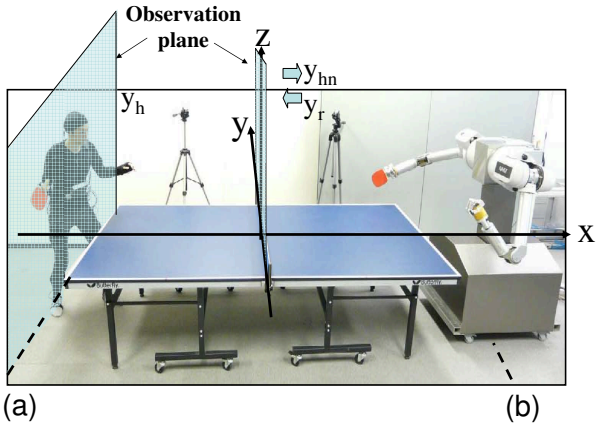

Fig. 2: A diagram of simplified ping-pong communication model.(a) is player's returning position. (b) is robot's returning position.

卓球タスクのモデル化について述べる．まず，そのモデル化の概要を Fig.2 に示す．卓球の球が 1 質点の力学系に従うとする．卓球台の大きさは公式ルールに従い，縦 2.740[m]，横 1.525[m]，高さ 0.760[m] とする．座標軸の原点を卓球台の中心に取る．ロボットとプレイヤーの返球位置の x 座標を固定する．そして，連続時間系の卓球を，観測する $y-z$ 平面を 2 平面指定することにより，離散時間系に変換する．観測平面は，ネットがある面と，プレイヤーの返球位置の 2 平面とする．この 2 平面を通過する球の y 座標を観測し，プレイヤーの返球位置を y_h，プレイヤー側からネット上を通過する位置を y_{hn}，ロボット側からネット上を通過する位置を y_r とする．このとき制御対象の状態 $\boldsymbol{x}(t)$ を

$$\boldsymbol{x}(t) = [y_h(t-1), y_{hn}(t-1)]^T \tag{1}$$

とする．制御対象への制御出力 $\boldsymbol{u}(t)$ はロボットの返球によるネット面での目標通過位置

$$\boldsymbol{u}(t) = [y_r(t), z_r(t)]^T \tag{2}$$

とする．

状態変化は $\boldsymbol{x}(t+1) = f(\boldsymbol{x}(t), \boldsymbol{u}(t))$ に従うとする．また，インタラクションにのみ焦点を当て，一度バウンドした球は，ロボット，プレイヤーともにエラーをせず返球することと仮定する．

2.2　プレイヤーが取る戦略のモデル化

プレイヤーはロボットを y 軸方向に移動させるような戦略を取るとする．その中で今回は 2 種類の戦略について考える．

戦略 **1**：ロボットからの打球を，飛んで来た位置と反対方向へ返球する

戦略 **2**：前回プレイヤーが打った位置の反対側へ返球する

戦略 **1** は y_r に来た球を $y_{hn} = -y_r$ になるように打つ戦略である．戦略 **2** は $y_{hn}(t-1)$ に球を打ったとき $y_{hn}(t) = -y_{hn}(t-1)$ になるように打つ戦略である．ここで t はラリー回数である．以上の 2 戦略を用いてモデル同定と戦略学習を行う．

3　ロボットの戦略学習

ロボットの戦略学習は鮫島らの研究 [8, 9] を参考に行う．プレイヤーの戦略を線形の力学系モデルとして近似し，Linear Quadratic Controller(LQC) により最適制御則を求めることでプレイヤーの状態を制御する戦略を獲得する．本論文では，プレイヤーの位置を $y_h = 0$ に留め，長期的なラリーを行うためのコスト関数を設定し，戦略を求める．以下にプレイヤーの戦略モデル同定方法とロボットの戦略学習方法を示す．

3.1　LQC に基づく戦略学習

ロボットの戦略は，LQC を用いて評価関数 V が最小になるように求める．評価関数はコスト関数の無限時間までの累積値である．状態予測モデルとコスト予測モデルはそれぞれ

$$\hat{\boldsymbol{x}}(t+1) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \tag{3}$$

$$\hat{c}(t) = \boldsymbol{x}(t)^T Q \boldsymbol{x}(t) + \boldsymbol{u}(t)^T R \boldsymbol{u}(t) \tag{4}$$

とする．$\hat{\boldsymbol{x}}$,$\hat{c}$ は予測値を表す．各パラメータの学習には逐次最小二乗法を用る．ここで状態変数 $\boldsymbol{x}$ を D_s 次元のベクトル，制御変数 $\boldsymbol{u}$ を D_c 次元のベクトルとすると，$\boldsymbol{A}$ は $D_s \times D_s$ 次元，$\boldsymbol{B}$ は $D_s \times D_c$ 次元，$\boldsymbol{Q}$ は $D_s \times D_s$ 次元，$\boldsymbol{R}$ は $D_c \times D_c$ 次元のパラメータ行列である．評価関数 V は $D_s \times D_s$ 次元のパラメータ行列 $\boldsymbol{P}$ を用いて

$$V = \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x} \tag{5}$$

のように 2 次形式で表される．$\boldsymbol{P}$ はリッカチ方程式

$$0 = \frac{1}{\tau}\boldsymbol{P} - \boldsymbol{P}\boldsymbol{A} - \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{B}\boldsymbol{R}^{-\boldsymbol{1}}\boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{P} - \boldsymbol{Q} \tag{6}$$

の解である．このとき，V を最小にする最適制御出力は

$$\boldsymbol{u}(t) = -\boldsymbol{K}\boldsymbol{x}(\boldsymbol{t}) \tag{7}$$

により求められる．$\boldsymbol{K}$ は最適フィードバックゲインであり

$$\boldsymbol{K} = \boldsymbol{R}^{-1}\boldsymbol{B}^T\boldsymbol{P} \tag{8}$$

で求められる．以上の制御法により適切なモデルやコストを仮定することで，戦略学習を実現する．次節においてその同定法について述べる．

3.2 プレイヤーの戦略モデル同定方法

状態予測モデルとコスト予測モデルの．パラメータ $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{Q}, \boldsymbol{R}$ の更新則は，入力ベクトルを $\boldsymbol{z}(t)$ としたとき，目標値 $\boldsymbol{y}(t)$ への線形回帰モデルを $\hat{\boldsymbol{y}}(t) = \boldsymbol{\theta}\boldsymbol{z}(t)$ とすると，

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}(t-1) + \frac{\boldsymbol{S}(t-1)\boldsymbol{z}(t)\left(\boldsymbol{y}(t) - \boldsymbol{z}^T(t)\boldsymbol{\theta}(t-1)\right)}{\rho + \boldsymbol{z}(t)^T\boldsymbol{S}(t-1)\boldsymbol{z}(t)} \tag{9}$$

となる．ここで $\boldsymbol{S}$ は入力の逆共分散行列の推定値であり，

$$\boldsymbol{S}(t) = \frac{1}{\rho}\left(\boldsymbol{S}(t-1) - \frac{\boldsymbol{S}(t-1)\boldsymbol{z}(t)\boldsymbol{z}^T(t)\boldsymbol{S}(t-1)}{\rho + \boldsymbol{z}(t)^T\boldsymbol{S}(t-1)\boldsymbol{z}(t)}\right) \tag{10}$$

である．ρ は忘却係数である．状態予測モデルの場合には状態 $\boldsymbol{x}$，行動 $\boldsymbol{u}$ を含めて $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{u}^T)$，コスト予測モデルの場合には各パラメータについて展開し状態の 2 次の成分を含んだ $\boldsymbol{z} = (\boldsymbol{x}^T, x_1\boldsymbol{x}^T, x_2\boldsymbol{x}^T, \cdots, x_n\boldsymbol{x}^T)^T$ を入力ベクトルとする．

4 シミュレーションによる実験結果

上記手法の有効性を検証するため，以下の設定でプレイヤーの 2 種類の戦略についてシミュレーションによる実験を行った．初期状態を，係数行列を全て 0，$\rho = 0.9$ とし，その他のパラメータは 0〜1 のランダムとした．また，自然な卓球を想定して，ラリー中，高さが 1.7$[m]$ 以上になる球は打たないものとした．そして，制御出力の z 成分がシステムに関与しないので，簡便のため

$$u(t) = y_r(t) \tag{11}$$

とした．また，コスト関数を

$$c(t) = 10^5(y_h(t) - 2y_r(t))^2 + 10^{10}y_r(t)^2 \tag{12}$$

と設定した．以下に結果を示す．

4.1 モデル同定結果

Fig.3 において (a)〜(d) はそれぞれ，戦略 1 における $\boldsymbol{x}$ と c，戦略 2 における $\boldsymbol{x}$ と c の，シミュレーションを 5 回行ったときの，実測値と推定値の平均二乗誤差と標準偏差を，それぞれ実線とエラーバーで表している．学習が進み，パラメータが逐次最適化されるに従って誤差が減少した．このことにより，パラメータが収束し，各戦略が線形の力学系モデルで適切に同定されていることが確認できた．設定した 2 種類の戦略を用いたときのインタラクションは線形の力学系モデルで近似できることが確認できた．

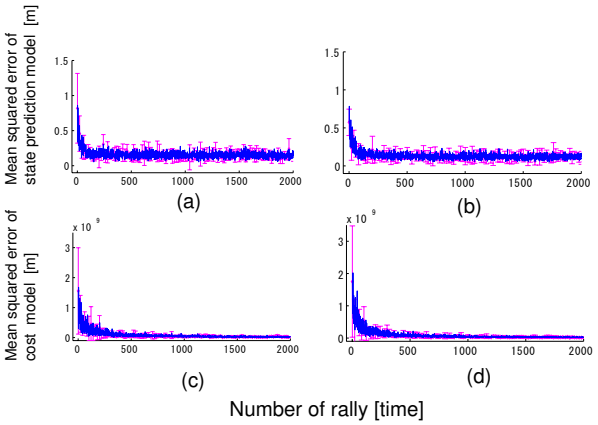

Fig. 3: Learning curve of state and cost prediction model. (a) shows state prediction error with strategy1. (b) shows state prediction error with strategy 2. (c) shows cost prediction error with strategy 1. (d) shows cost prediction error with strategy 2. Solid line is average. Error bar is the standard deviation.

4.2 LQC による制御結果

パラメータ学習と同時に LQC による制御を行った結果を示す．Fig.4 は制御を行ったときと行わなかったときのあるラリーにおけるプレイヤーの位置 y_h の遷移を示す．(a) が戦略 1，(b) が戦略 2 の結果である．青線が制御を行わなかったときの結果であり，赤線が制御を行ったときの結果である．“ X ”の点は，その点でラリーが途切れたことを示す．制御を行っていないときには発散していたプレイヤーの動きが，制御を行うことにより収束し，長期間ラリーが続けられた．このことにより，状態予測が完全に収束せず，誤差がある状態でもロボットがプレイヤーの状態を望みの状態へ遷移させる戦略が獲得できていることが分かった．

5 まとめ

本論文では，卓球コミュニケーションロボットの戦略学習のための基礎的研究として，一般的に想定される卓球戦略が，線形の力学系モデルで同定でき，最適制御の枠組みを用いて，プレイヤーの行動遷移を制御するロボットの戦略学習ができることを示した．今後は，プレイヤーが，複数の戦略を切り替えながらラリーをする場合を考えるなど，卓球タスクの複雑化を

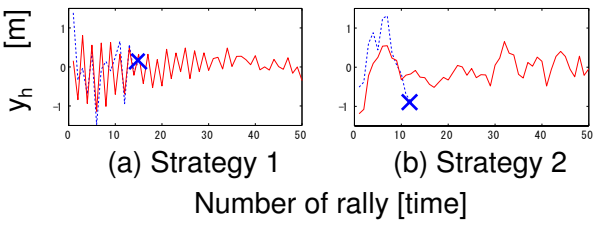

Fig. 4: Player's position trajectory. (a) shows strategy 1. (b) shows strategy 2. Dash line is the result without control. Solid line is the result with control. " X "mark means the end of rally.

行うことで，実際の卓球に近づけていくことを目指す．また，Fig.1 に示した，実システムへの実装についても検討していく．

参考文献

[1] 原田, 佐藤, 森 ,“ 触れ合いロボットによる心理効果ー接触インタラクションによる安心感の演出と痛みの緩和ー ”, 日本ロボット学会誌, vol.16, no.5, pp.698–704, 1998.

[2] 稲邑, 稲葉, 井上,“ PEXIS:統計的経験表現に基づくパーソナルロボットとの適応的インタラクションシステム ”, 電子情報通信学会論文誌, volJ-84-D-I, no.6, pp.867–877, 2001.

[3] 佐藤, 西田, 市川, 畑村, 溝口,“ ロボットによる人間の意図の能動的理解機能 ”, 日本ロボット学会誌, vol.13, no.4, pp.545–552, 1995.

[4] 畑田, 宮本,“ 失敗から学ぶ卓球ロボット ”, 電子情報通信学会技術報告, vol.102, no.430, pp.31-35, 2002.

[5] Matsushima, Hashimoto, Miyazaki, “ Learning to the Robot Table Tennis Task —Ball Control & Rally with a Human— ”, 2003 IEEE Int. Conf. on Systems, Man & Cybernetics, pp. 2962–2969, 2003.

[6] D. C. Bentivegna, C. G. Atkeson, A. Ude, G. Chang, “ Learning to Act from Observation and Pracctice ”, International Journal of Humanoid Robotics, vol.1, no.4, pp.585–611, 2004.

[7] 高野, D. Kulic, 中村,“ インタラクションの推定・制御に基づく身体的コミュニケーション ”, 日本ロボット学会学術講演会, 2007.

[8] 鮫島, 片桐, 銅谷, 川人,“ 複数の予測モデルを用いた強化学習による非線形制御 ”, 電子情報通信学会論文誌, vol.J84-D-II, no.9, pp.2092–2106, 2001.

[9] 杉本, 鮫島, 銅谷, 川人,“ 複数の状態予測と報酬予測モデルによる強化学習と行動目標の推定 ”, 電子情報通信学会論文誌, vol.J87-D-II, no.2, pp.683–694, 2004.