ニューラルネットと強化学習を用いた音声によるコミュニケーションの自律学習

大分大学 ○笹原冬月　柴田克成

Autonomous learning of sound communication by reinforcement learning with a neural network

Kazuki Sasahara, Katsunari Shibata Oita University

Abstract. In order to develop a robot that can communicate with a human naturally, autonomous learning of communications must be required. Based on the idea, the authors have examined whether communications emerge through reinforcement learning with a neural network in a task. A transmitter and a receiver of communication signals are put in a two dimensional space. The transmitter can perceive the position of the receiver, and can send some communication signals to the receiver. The receiver can move and receive the communication signals that are sent by the transmitter, but cannot know the relative position of the transmitter. The transmitter does not know what signal should be sent, and the receiver does not know what action should do according to the received signals. What we give them is only a reward. In this paper, in a discrete space task with more states than previous, it was examined that they could learn to communicate appropriately using a real sound communication system with a speaker and microphone. In a continuous space task, it was examined after learning that the receiver's location on two-dimensional space could be transmitted using two continuous communication signals.

1　まえがき

近年、コミュニケーションを行うロボットが出現しているが、融通が利かないことや、受け答えの無機質さから、まだ、人間社会には受け入れにくい存在と言える。これは、全てを人間がプログラムして、ロボットの受け答えを基本的に与えてきたことが原因の一つであろう。融通が利く柔軟なコミュニケーションを実現するためには、テーブルルックアップ的な受け答えではなく、会話の内容を理解することが必要と考える。そのため、人間からプログラムを与えられるのではなく、ロボット自身が自律的に学習を行っていき、柔軟なコミュニケーションを獲得していくことが重要になってくるであろう。

強化学習は、報酬や罰による自律的な学習則であるが、従来の強化学習は、一般的に行動やプランニングのための学習ととらえられる傾向が強い。しかし、適切なコミュニケーションを行うことで報酬をもらえるとすれば、コミュニケーションも報酬を得るための行動の一種と考えられるので、強化学習で自律的に学習できる可能性がある。しかし、単にテーブルルックアップをベースとした強化学習では、型にはまったコミュニケーションしかできないと考えられる。柔軟なコミュニケーションを実現するためには、センサ信号や相手からの信号をもとに自分なりに状況を判断し、信号を出力していく必要があると考えられる。そのためにはニューラルネットの導入が有効であると考える。

そこで、筆者らは以前より強化学習とニューラルネットを組み合わせることによって、報酬や罰の情報を元にコミュニケーションが創発するかどうかの研究を行ってきた。そして WERNER[1] らの文献を参照して、動けない雌がコミュニケーション信号を発し、目の見えない雄がその信号を受けて行動し、両者が出会ったら報酬が得られるというタスクで学習能力を検証してきた。先行研究[2] では、ニューラルネットと強化学習を組み合わせ、離散空間のタスクにおいて学習を行うと、出会うために雌はどういう情報を伝達すれば良いのか、雄は受け取った情報をどういう行動に反映させれば良いか、ということを自ら獲得していくことが確認されている。さらに、別の先行研究[3] では、ノイズのある環境でのコミュニケーション信号の離散化を検証するために、一次元の連続空間で、雌の信号と雄の行動を連続的にして、ニューラルネットワークと強化学習で学習することで、正しくコミュニケーションが行われ、さらにノイズによって、信号が離散化することを確認している。このように、強化学習とニューラルネットを組み合わせることで、コミュニケーションを自律的に獲得できる。ただし、両者の先行研究[2][3] とも信号伝達は同一コンピュータのプログラムの中で仮想的に行われており、タスクも簡単であった。

本研究では、Fig.1 のように、独立したニューラルネットで構成されている信号の送信者と受信者を配置し、コミュニケーションに実際の音声を使用しても、また、タスクを少し難しくしてもコミュニケーションを獲得できるのか検証した。離散空間タスクでは実際に信号のやりとりにマイクとスピーカーを用いて、音の大きさの強弱ではなく、周波数による情報の伝達によってコミュニケーションを獲得できるのか、さらに、空間を広げ状態数を大きくしても、コミュニケーションを獲得できるのかを検証した。連続空間では、空間を 2 次元に広げ、2 次元の移動を 2 つの信号で表現することが可能か検証した。なお、

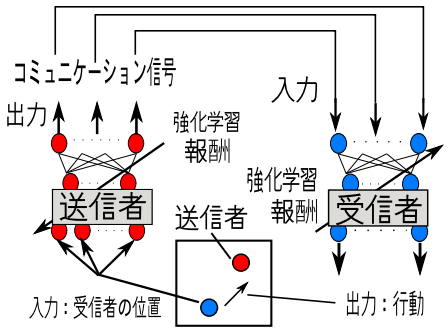


Fig. 1: System architecture and signal flow of the communication learning system used in this paper

先行研究[2)3)]にならい、離散空間タスクでは、Q-learning[4)]を、連続空間タスクでは Actor-Critic[4)] を用いた

2　離散空間タスク

2.1　タスク設定と学習方法

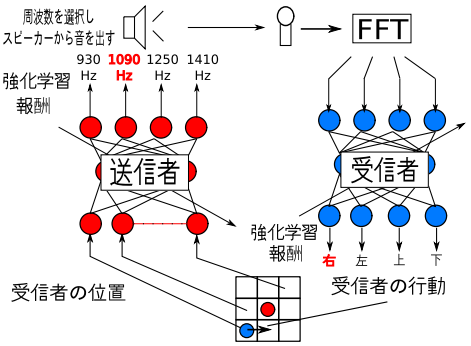


Fig. 2: Discrete space task

離散空間タスクの信号の流れを Fig.2 に示す。Fig.5(a) のような外側に壁がある 7×7 の格子状の離散空間において、コミュニケーション信号の送信者を真ん中に、受信者を毎回ランダムな位置に配置した。両者は別々のニューラルネットを持ち、送信者の発する信号を元に受信者が行動し、両者が接触したらゴールとして両者ともに報酬をもらえるものとした。

送信者は、7×7 の離散空間の状態数 49 と同数の入力があり、受信者がいる位置に該当する入力のみを 1、それ以外を 0 とした。出力は 4 個用意し、それぞれ 930Hz, 1090Hz, 1250Hz, 1410Hz の周波数に対応させる。その出力を Q 値として扱い、ε-greedy[4)] で周波数を選択し、選択された周波数の音を実際にスピーカーから出力した。

一方、受信者は、マイクを使って音を受信し、FFT に通して得られたスペクトルについて、920〜940Hz, 1080〜1100Hz, 1240Hz〜1260Hz, 1400Hz〜1420Hz の 4 つの各周波数帯の平均を最大値で正規化した値をニューラルネットへの入力とした。受信者は上下左右の 4 つの行動のどれかをとることができる。したがって、各行動に対応する 4 つの出力を設け、その出力を Q 値として、コミュニケーション信号の場合と同様 ε-greedy[4)] で行動選択を行った。先行研究での知見を参考にし、受信者は早い段階から greedy 選択に近く、送信者はランダム行動を少し残すように、送信者は式 (1)、受信者は式 (2) にしたがって ε を指数関数的に減少させた。

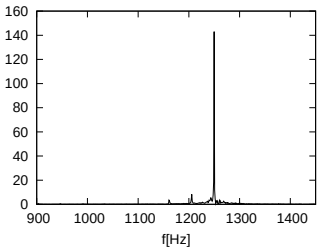


Fig. 3: An example of spectrum distribution of the sound received by the microphon when the 1250Hz sound is transmitted from the speaker

$$\varepsilon = 0.01 + \exp\left(-\frac{episode}{600}\right) \tag{1}$$

$$\varepsilon = 0.01 + 0.2\exp\left(-\frac{episode}{300}\right) \tag{2}$$

ニューラルネットはともに 3 層構造で、各層のニューロン数は、受信者の方が 49-20-4、送信者の方が 4-10-4 とした。各ニューロンの出力関数は 0.0 から 1.0 を値域とするシグモイド関数とした。

学習は、送信者が音を出すことも行動の一種と考え、両者とも離散行動の学習に適した Q 学習を用いた。ニューラルネットに対し、実際に行った行動に該当するニューロンに対する教師信号を

$$T_a(S_{t+1}) = r_{t+1} + \gamma \max_a Q_a(s_{t+1}) \tag{3}$$

と与える。ただし、ゴールしたときの報酬は $r = 0.9$、壁にぶつかったときは $r = -0.1$ とし、試行終了後は次のステップの Q 値は計算できないため Q 値の最大値は 0 とした。また、Q 値とニューラルネットの出力については、-0.1 から 0.9 の範囲の Q 値が 0.1 から 0.9 のニューラルネットの出力になるように、互いに線形変換して用いた。したがって、実際の教師信号は Eq.(3) を 0.8 倍して 0.18 を足す必要がある。また、教師信号は 0.1 から 0.9 の範囲を越えないようにした。

学習の流れは以下のようになる。

1. 送信者は受信者の状態 s を観測し、自身のニューラルネットに入力する。
2. 送信者のニューラルネットの出力を計算し、コミュニケーション信号を ε-greedy に従って選択する。
3. 選択された周波数の sin 波をスピーカーから再生する。

4. 同時にマイクで録音を行う。
5. 録音した音を FFT に通し、各周波数帯のスペクトルの平均を最大値で正規化した値を受信者のニューラルネットに入力する。
6. 受信者のニューラルネットの出力を計算し、行動を ε-greedy に従って選択した行動を実行する。
7. 両者が接触したら報酬を両者に与え、壁にぶつかったら、その場所に留まるとし、両者に罰を与える。
8. 送信者は状態遷移後の状態 s_{t+1} を観測し、ニューラルネットの計算を行い、最大 Q 値の周波数の音を再生する。受信者はその音を受けて、ニューラルネットの計算を行い、そこから送信者、受信者の状態 s_t での教師信号を生成し、BP 法で教師あり学習をさせる。
9. 両者が接触したら試行終了とし、それ以外は時間ステップを t から $t+1$ へ進めて 1 に戻る。

2.2 離散空間タスクの結果

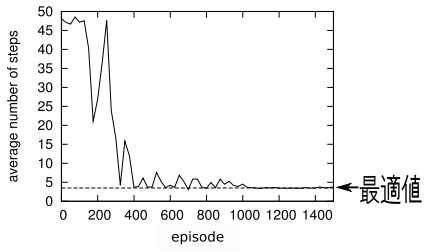
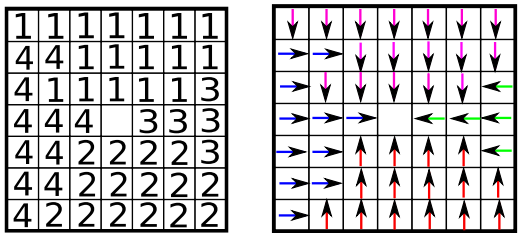


Fig. 4: Learning curve in the discrete space task with real sounds. The vertical axis indicates the number of average steps.

Fig.4 に、離散空間で、実際にスピーカーとマイクを用いて 1500 回試行させたときの試行回数に対する平均ステップ数の変化を示す。ここで、平均は、25 回試行するごとに学習を中断して、ランダム行動をなくし ($\varepsilon = 0$)、100 回試行させて、ゴールするまでにかかったステップ数の平均である。ただし、50 ステップ以上かかった場合そこで試行終了とし、次の試行へ移るとした。なお最適な平均ステップ数は、およそ $\frac{1\times4+2\times8+3\times12+4\times12+5\times8+6\times4}{48} = 3.5$ となる。試行回数に対して、平均ステップ数が最適な平均ステップ数に近づいていることが分かる。Fig.5(a) は、学習終了後の受信者の位置における、送信者の出す各周波数の音を番号で示したものである。930Hz は 1、1090Hz は 2、1250Hz は 3、1410Hz は 4 で示している。Fig.5(b) は、Fig.5(a) の信号を受けての受信者がとる行動を示している。Fig.5 と学習終了後の Table 1 から、送信者は、受信者がとるべき行動に関する情報を実際の音で表現し、受信者は、930Hz の音のときは下に行動、1090Hz の音のときには上に行動、1250Hz の音のときには左に行動、1410Hz の音のときには右に行動、というように、各音に対して行動を確立させていることが確認できた。



(a)Communication signal (b)Receiver's motion

Fig. 5: (a) The communication signals generated by the transmitter and (b)the receiver's action generated from the communication signals in the discrete space task

Table 1: Q-value for each frequency-action pair in the receiver after learning

	右	左	上	下
930Hz	0.66	0.63	0.59	**0.74**
1090Hz	0.65	0.61	**0.69**	0.63
1250Hz	0.65	**0.80**	0.57	0.65
1410Hz	**0.79**	0.64	0.62	0.67

3 連続空間タスク

3.1 タスク設定と学習方法

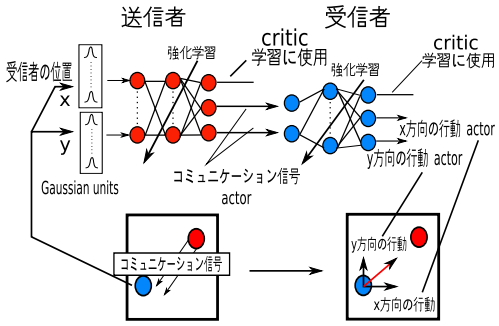


Fig. 6: Continuous space task

連続空間タスクでは、二次元の連続空間を用意し、送信者の位置、つまりゴールの位置を、中心 (0.5,0.5) の半径 0.25 の円内とした。なお、送信者は x, y 方向それぞれに最大 0.6 動くことができ、空間のどこからでも 1 ステップでゴールすることができ、ゴールしたら、報酬 $r = 0.9$ を送信者、受信者の両者に与える。また、空間の外側の黒い線は壁であり、壁にぶつかったら、ぶつかった場所に止まるとし、罰を送信者、受信者の両者に $r = -0.1$ を与えた。受信者は中心に向かうだけでなく、行き過ぎないように中心付近では小さく、中心から離れているところでは大きく移動することが求められる。ニューラルネットはともに 3 層構造で、各層のニューロン数は、受信者の方が 60-20-3、送信者の方が 2-10-2 とした。学習は、連続的な動作の学習に適した Actor-Critic を用いた。

送信者の入力には、受信者の x 座標と、y 座標をそれぞれ Eq.(4) と Eq.(5) によって表される 30 個の gaussian

で局所化し合計 60 個の信号を、ニューラルネットに入力した。これによってニューラルネットの計算で求められる非線形性を少なくし、受信者の位置の変化に対するニューラルネットの出力の急激な変化を送信者が容易に学習できるようにした。

$$GS_i(x) = exp\left(-\frac{1}{2}(29x - i)^2\right) \tag{4}$$

$$GS_i(y) = exp\left(-\frac{1}{2}(29y - i)^2\right) \tag{5}$$

ここで、x, y は受信者の位置の座標 (x, y) を表し、0.0 から 1.0 の値をとる。i は gaussian の番号で $i = 0, 1, 2....29$ とする。送信者も受信者も出力層のニューロン数は 3 つとし、1 つを critic に、残り 2 つを actor とした。中間層および出力層の各ニューロンの出力関数は、値域 -0.5〜0.5 のシグモイド関数を用いた。学習はまず TD 誤差を

$$\hat{r}_t = r_t + \gamma P_t - P_{t-1} \tag{6}$$

と計算する。ここで r_t は報酬、P_t は Critic の出力を表し、実際のニューラルネットの出力に 0.5 を足したものを用いた。また γ は割引率である。Critic の出力に対しては、教師信号を

$$P_{s,t-1} = P_{t-1} + \hat{r}_t = r_t + \gamma P_t \tag{7}$$

と計算し、BP 法で教師あり学習をさせる。出力ベクトル $\boldsymbol{M_t}$ は

$$\boldsymbol{M_t} = \alpha 2.5 \boldsymbol{A_t} + \boldsymbol{rnd_t} \tag{8}$$

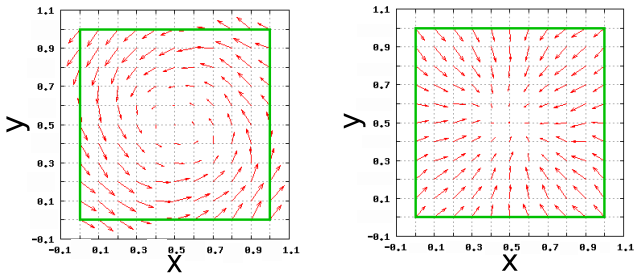
と計算する。ただし、$\boldsymbol{A}$ は Actor の出力ベクトルでニューラルネットの出力の 2 つを割り当てる。$\boldsymbol{rnd}$ は試行錯誤のための乱数ベクトルで、値域-0.9〜0.9 の正負対称の一様乱数を 3 乗した値を用いた。α は定数であり、送信者は $\boldsymbol{M_t}$ の値域が-1.0〜1.0 となるように $\alpha = 2.5$、受信者は 1 ステップでゴールできるように $\alpha = 1.0$ とした。この出力は、送信者にとっては、コミュニケーション信号であり、受信者にとっては、動作信号である。受信者には、送信者の出力する 2 つの Actor の出力値を入力した。Actor の出力に対しては、教師信号を

$$\boldsymbol{A_{s,t-1}} = \boldsymbol{A_{t-1}} + 0.5\hat{r}_t \boldsymbol{rnd_{t-1}} \tag{9}$$

と計算し、教師あり学習をさせた。教師信号は最大で 0.4、最小で-0.4 とした。

3.2 連続空間タスクの結果

Fig.7 には、400000 回試行させた後の送信者の 1 つ目の信号を x 成分に、2 つ目の信号を y 成分としてベクトルで示したものを (a) に、(b) には、(a) の信号を受け取って受信者がとった行動をベクトルで示している。ただし、(a) のベクトルは 0.1 倍、(b) のベクトルは受信者が行動



(a)Communication signal (b)Receiver's motion

Fig. 7: (a)The communication signals generated by the transmitter and (b)the receiver's motion command generated from the communication signals in the continuous space task

できる最大の大きさ $0.6\sqrt{2}$ で割って、0.1 倍して表現している。Fig.7(b) のベクトルがゴール方向を差し、ゴールに近づくにつれて小さくなっていることから、受信者はコミュニケーション信号を受けてゴールするために最適な行動を学習することができたことが分かる。送信者は、60 個の入力信号を総合して、1 つ目の信号が受信者の $-y$ 座標を、2 つめの信号が x 座標をおおよそ表しており、2 つのコミュニケーション信号をつかって受信者の位置の情報を送ることができるようになっていることが分かる。一方、受信者もその信号を受けて、正しい方向に、そして、行きすぎることなくゴールに到達できた。

4 まとめ

本研究では、ニューラルネットと強化学習を用いたコミュニケーションの自律学習において、以下の 2 点を確認した。まず、Q-learning を用いた離散空間でのタスクにおいて、スピーカーとマイクを用いて周波数が違う実際の音声で信号伝達を行っても、学習できることを確認した。また、離散領域を 7×7 にし、状態数を大きくしても学習ができることが確認できた。

次に、Actor-Critic を用いた連続空間でのタスクでは、2 次元の行動を 2 つの信号で表現し、コミュニケーションが学習できることを確認した。

謝辞 本研究は、日本学術振興会科学研究費補助金 ♯19300070 によって補助された。ここに謝意を表す。

参考文献

1) WERNER G.M., Evolution of Communication in Artificial Organisms, Artificial Life II, 1991.
2) 仲西賢展, 柴田克成, Q 学習に基づく一方向コミュニケーション学習における学習効率化手法の提案, システム・情報部門学術講演会講演論文集, pp157-162, 2004.
3) 柴田克成, コミュニケーションの強化学習におけるノイズ付加による連続値信号の離散化, 電子情報通信学会技術研究報告, Vol.103, No.734, NC2003-203, pp.55-60, 2004.
4) Richard S.Sutton, Andrew G.Barto, 三上 貞芳, 皆川 雅章 共訳, 強化学習, 森北出版, 2000.