Emergence of Purposive and Grounded Communication through Reinforcement Learning

Katsunari Shibata and Kazuki Sasahara

Dept. of Electrical & Electronic Engineering, Oita University,
700 Dannoharu, Oita 870-1192, Japan
shibata@oita-u.ac.jp

Abstract. Communication is not just the manipulation of words, but needs to decide what is communicated considering the surrounding situations and to understand the communicated signals considering how to reflect it on the actions. In this paper, aiming to the emergence of purposive and grounded communication, communication is seamlessly involved in the entire process consisted of one neural network, and no special learning for communication but reinforcement learning is used to train it. A real robot control task was done in which a transmitter agent generates two sounds from 1,785 camera image signals of the robot field, and a receiver agent controls the robot according to the received sounds. After learning, appropriate communication was established to lead the robot to the goal. It was found that, for the learning, the experience of controlling the robot by the transmitter is useful, and the correlation between the communication signals and robot motion is important.

Key words: emergence of communication, grounded communication, reinforcement learning, neural network, robot control task

1 Introduction

Many speaking robots have appeared recently, and interactive talking can be seen in some of them. A robot talking with humans looks intelligent at a glance, but a long interaction with them makes us notice that the partner is not a real life but a robot. One major reason must be that the communication is not grounded, but is just the manipulation of words based on pre-designed rules. Many attempts have been made to solve the "Symbol Grounded Problem"[1] for a long time. In the model of lexicon emergence in [2] or [3], extracted features of a presented object are associated with words or codes. Under the assumption of common observation between two agents, the models have a way of getting the listener's words closer to the speaker's.

They suppose patterns and symbols separately, and focus on bridging between them through specialized learning that is independent of the other learning. Steels himself said in [3], "The experiments discussed in this article all assume that agents are able to play language games, but how do the games

themselves emerge?" The question gets the heart of the problem. Primitive communication observed in animals or ancient people seems purposive such as telling food location or coming dangers. Communication should emerge in the learning in daily life, and the communication learning should not be isolated from the other learning. It is worth noting that, when we see the section of the brain, the language areas are not isolated from the other areas, nor look so different from them. The communication is not generated only by the language areas of the brain, but is generated by the whole brain as a massively parallel and flexible processing system. That enables us to consider many things simultaneously in parallel and to decide flexibly and instantly what we talk, the authors think.

The emergence of purposive communication has been aimed by evolutional approach[4] or reinforcement learning[5]. The author's group has also investigated it through reinforcement learning[6][7][8]. Discretization of the communication signal through reinforcement learning in a noisy environment was also shown[8]. However, in these cases, the environment is very simple, and learning is performed only on computer simulation.

In this paper, using a real camera, speaker, microphone, and robot, a transmitter learns to output two sounds with appropriate frequencies from more than one thousand color image signals from the camera, and a receiver learns to output appropriate motion commands from the received sounds. Each agent uses a neural network to compute the output, and learns it by reinforcement learning only from a reward when the robot reaches a goal state and a small punishment when it is close to a wall. The emergence of symbol is left as a future problem.

There are some communication robots with one or two cameras[9][10][11], but the camera is used for the perception of communication partners or environment or for giving the feeling of being gazed to the partner. The camera image is not reflected to the communication directly, and no organic integration of the camera image and communications can be seen in them.

2 Reinforcement Learning with a Neural Network[12]

Reinforcement learning is autonomous and purposive learning based on trial and errors, and a neural network (NN) is usually used as a non-linear function approximator to avoid the state explosion due to the curse of dimensionality. An author has claimed that by the combination, parallel processing that enables to consider many things simultaneously is learned purposively, seamlessly and in harmony, and as a result, necessary functions such as recognition, memory (when using RNN) emerges to get rewards and to avoid punishments. The flexible and parallel processing is expected to contribute to saying goodbye to the "Functional Modules" approach, in which each functional module is sophisticatedly programed independently and the modules are integrated to develop an intelligent robot. It is also expected to contribute to solving the "Frame Problem".

The system is consisted of one NN whose inputs are sensor signals and whose outputs are actuator commands. Based on reinforcement learning algorithm, training signals are generated autonomously, and supervised learning is applied

using them. This eliminates the need to supply training signals from outside. In this paper, for a continuous input-output mapping, actor-critic[13] is used as a reinforcement learning method. Therefore, the outputs of the NN are divided into a critic output P and actor outputs $\mathbf{a}$. The actor output vector $\mathbf{a}$ is used as motion commands to its actuators after adding a random number vector $\mathbf{rnd}$ as an exploration factor. For learning, TD-error is represented as

$$\hat{r}_{t-1} = r_t + \gamma P(\mathbf{s}_t) - P(\mathbf{s}_{t-1}) \tag{1}$$

where r_t is the reward given at time t, γ is a discount factor, $\mathbf{s}_t$ is the sensor signal vector that is the input of the NN at time t, and $P(\mathbf{s}_t)$ is the critic output when $\mathbf{s}_t$ is the input of the network. The training signal for the critic output is computed as

$$P_{d,t-1} = P(\mathbf{s}_{t-1}) + \hat{r}_{t-1} = r_t + \gamma P(\mathbf{s}_t), \tag{2}$$

and the training signal for the actor output is computed as

$$\mathbf{a}_{d,t-1} = \mathbf{a}(\mathbf{s}_{t-1}) + \hat{r}_{t-1}\mathbf{rnd}_{t-1} \tag{3}$$

where $\mathbf{a}(\mathbf{s}_{t-1})$ is the actor output when $\mathbf{s}_{t-1}$ is the input of the NN, and $\mathbf{rnd}_{t-1}$ is the random number vector that was added to $\mathbf{a}(\mathbf{s}_{t-1})$. Then $P_{d,t-1}$ and $\mathbf{a}_{d,t-1}$ are used as training signals, and the NN with the input $\mathbf{s}_{t-1}$ is trained once according to Error Back Propagation[14]. Here, the sigmoid function whose value ranges from -0.5 to 0.5 is used. Therefore, to adjust the value range of the neural network output to that of the actual critic value, 0.5 is added to the critic output of the neural network in Eq. (1), and 0.5 is subtracted from the derived training signal in Eq. (2). The learning is very simple and general, and as you notice, no special learning for communication or the task is applied.

3 Learning of Purposive and Grounded Communication

3.1 System Architecture and Robot Control Task

Fig. 1 shows the system architecture and performed task. There are a mobile robot (e-puck) in a $30cm \times 30cm$ square field and two communication agents; a transmitter and a receiver. The transmitter has a camera that is fixed and looking down the field from above. It has a neural network (NN), and its input vector $\mathbf{s}$ is the RGB pixel values of the camera image. It also has a speaker and transmits two sounds. The frequencies of two sounds are decided by the sum of the actor output vector $\mathbf{a}$ and an exploration factor $\mathbf{rnd}$ through the linear transformation of each element to the range between 1,000Hz and 1,300Hz. The two sounds are one-second sin-waves, and come out successively with a small interval. Due to a bug in the program, the frequency of the transmitted signal was actually about 20Hz smaller than intended. The receiver has a microphone and catches the two sounds from the transmitter. The receiver also has a NN. Its input vector $\mathbf{s}$ has 60 elements, each of which represents the average spectrum over 10Hz width around its responsible frequency of one of the two sounds and is

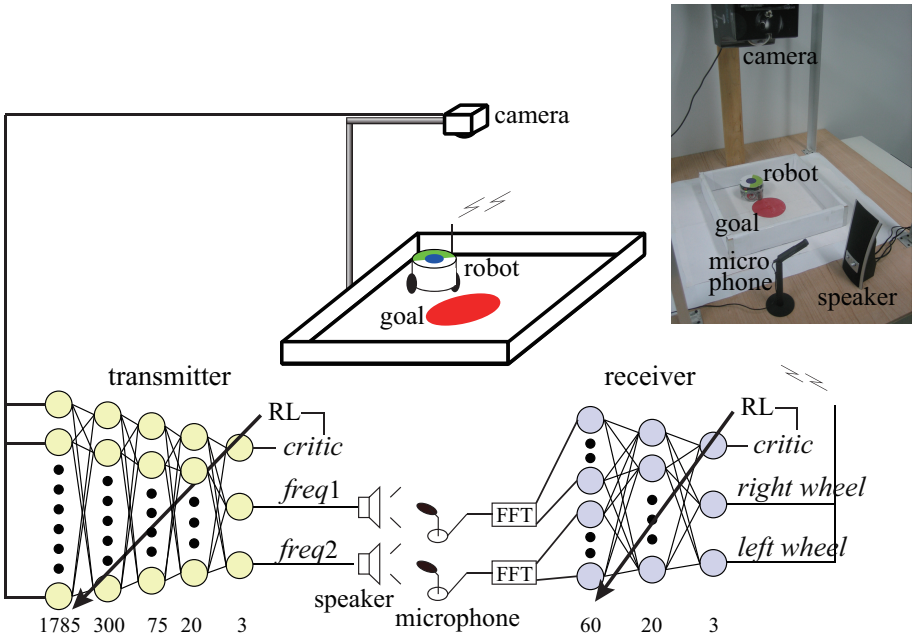


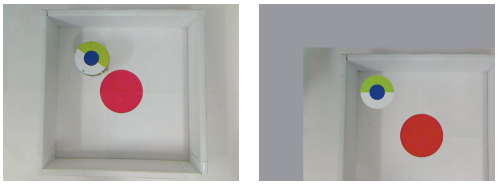
Fig. 1. System architecture and robot control task. In this figure, two speakers and two microphones are drawn, but actually, two sounds come out from one speaker with a small interval and are received by one microphone.

normalized by the maximum value. The receiver generates the control commands for the left and right wheels of the robot in proportion to the sum of its actor output vector **a** and an exploration factor **rnd**, and sends them to the robot through bluetooth.

Learning is very easy, and just proceeds according to the regular reinforcement learning independently in each agent as described in the last section. There is a big red circle in the center of the robot exploration field. When the robot center reaches the circle, the both agents get a reward 0.9 and the episode terminates. When the robot comes close to the wall, it is brought back to the position at the previous time step, and a small punishment -0.01 is imposed.

A sample raw camera image is shown in Fig. 2(a). To reduce the computational time, the image is resized to 26×20. Fig. 3 shows the definition of forward and backward and also relative and absolute orientation of the robot. The green part indicates the front of the robot, and absolute angle θ is the angle from the vertical axis of the image, and relative angle α is the angle from the line connecting to the center of the goal.

In the preliminary learning in which the NN with the input of 26×20 pixels is trained to output the relative distance and orientation $(cos\alpha, sin\alpha)$ for a variety of robot locations by supervised learning, the error for the orientation outputs did not decrease so much. It would be difficult to recognize the relative orientation



(a) Sample camera image (b) Robot-centered image

Fig. 2. Robot-centered image

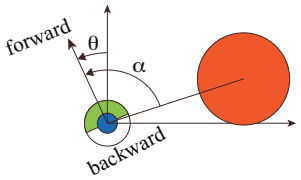


Fig. 3. The definition of forward and backward, and absolute and relative orientation θ and α of the robot.

for every robot location from the image inputs. Therefore, the robot-centered image as shown in Fig.2(b) was introduced. From the viewpoint of autonomous and seamless learning, acquisition of appropriate image shift by camera motion through learning is expected, but here, for simplicity, the image shift was given. The empty area that appears by the shift is filled with gray color as in Fig.2(b). Furthermore, to increase the precision, the resolution of the 5×5 area around the center of the image is doubled. Each pixel color is represented by the three signals for RGB, and 1,785 signals are the input of the NN in total. Each signal is linearly normalized from -0.5 to 0.5 prior to the input.

3.2 Effect of Preparation Learning

In this task, the robot can reach the goal area by going forward or backward after changing its orientation by rotating motions. The rotational direction can be left or right, but for eliminating wasted motion, the optimal one is right for $\alpha \leq 90°$ or $180° < \alpha \leq 270°$, and left for otherwise. Around $\alpha = 90°$ or $\alpha = 270°$, the optimal direction changes drastically by the small difference of α. After learning, the robot could reach the goal successfully. However, the rotational direction was not optimal, but was always the same. That would be because, for the transmitter, the communication signals do not directly influence the robot motion, but indirectly influence it through the receiver.

Then, before the communication learning, the transmitter learns directly to control the robot by reinforcement learning as a single agent learning. After that, using the internal representation of the NN, in other words, after resetting all the connection weights between hidden and output layers to 0.0, it learns the communication signals with the receiver. After the single agent learning, the rotational direction was appropriately chosen depending on the relative orientation α. Also after the following communication learning, the direction was appropriately chosen as shown in the next section. It is interesting that the previous experiments are useful for learning of appropriate communication.

3.3 Correlation between Communication Signals and Motions

One of the reasons of unsuccessful learning found during investigation is little correlation between communication signals and motions. In the receiver's NN,

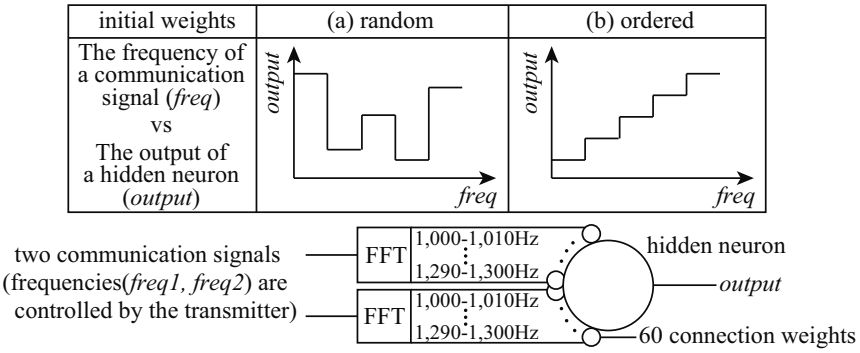


Fig. 4. The loss of the correlation between the frequency of a communication signal and the output of each hidden neuron by random initial weights in the receiver agent.

each hidden neuron had a random initial connection weight to each input signal after FFT. Therefore, the output of the neuron does not change monotonically according to the frequency of a communication signal as shown in Fig. 4(a). Then, the motion commands, which are the receiver's actor output, also have little correlation with the frequency. If the correlation does not exist, it is difficult for the transmitter to know whether the frequency should be increased or decreased to make the robot motion more appropriate. Accordingly, in this research, the weights for the inputs for one communication signal to each hidden neuron increase or decrease gradually as the responsible frequency of input increases as shown in Fig. 4(b). In the same reason, the exploration factor **rnd** that is added to the receiver's actor output is ± 0.1, while the transmitter's exploration factor is ± 1.8. It is reported also in [7] that such setting is useful.

4 Experiment

Parameters in this learning are shown in Table 1. Because of the high-dimensional input, the NN in transmitter has 5 layers, while the receiver has a 3-layer NN. 6,000 episodes of learning were done. The range of initial location of the robot becomes wider gradually as the learning progresses. Fig. 5 shows two sample episodes with no exploration factors after learning. In one of the episodes (a), the robot was located upper-left area and the absolute orientation of the robot was $\theta = 0°$, that means that the green part of the robot was located upper than the white part. In the other episode (b), the robot was located lower-left area and the orientation was also $\theta = 0°$. For each episode, time series of camera image, transmitter's critic and actors (signal frequencies), and receiver's critic and actors (motion commands) are shown. In the first sample, at first, the transmitter sent a high frequency sound followed by a low frequency sound, and the robot went backward rotating anti-clockwise. After that, the transmitter sent high frequency sound and then a little high frequency sound, and the robot went backward, and finally arrived at the goal. In the second sample, at first,

low-frequency sound and then high-frequency sound are sent, and the robot went forward rotating clockwise. After that, the transmitter's second sound became around the middle, and the robot went forward until it arrived at the goal.

Table 1. The parameters used in the learning.

	transmitter	receiver
number of neurons	1785-300-75-20-3	60-20-3
learning rate	0.5	0.3
initial weight (input -> hidden)	weight after preparation learning	orderd (-2.0 - 2.0)
initial weight (hidden -> output)	random [-0.5 - 0.5]	random [-2.0 - 2.0]
exploration factor	random [-1.8 - 1.8]	random [-0.1 - 0.1]
reward	0.9	
penalty	0.01	
discount factor γ	0.96	

Fig. 6(a) shows the two signal frequencies (transmitter's actor outputs) for some combinations of the robot location and absolute orientation θ. The frequencies are generated in the transmitter from the actually captured camera image. It can be seen that the frequencies are different depending on the location or orientation of the robot, but when the relative location of the goal from the robot is the same, the frequencies are similar to each other (e.g. upper left in (a-1) and lower left in (a-2)). Fig. 6(b) shows the motion commands (receiver's actor outputs) for some combinations of the two signal frequencies. To make this figure, actual sin-wave sound were emitted from the speaker, caught by the microphone, and were put into the receiver's NN after FFT. It can be seen that two motion commands change smoothly according to the two signal frequencies. Fig. 6(c) shows the relation between robot state and motion commands. The motion commands were generated from the actually captured image through the transmitter, the speaker, the microphone, FFT, and the receiver. It is shown that through appropriate communications, the robot rotated appropriately depending on the state even though the robot motion was not completely optimal.

The communication signals represent only the motions that the robot should execute, but does not represent the state or action value. Therefore, the receiver cannot represent the critic considering the robot state, but acquires the mapping from the communication signals to the robot motions. That is also shown in [15], and the problem of state confusion in the receiver was pointed in it.

5 Conclusion

It was shown that using a real mobile robot, a camera, a speaker, and a microphone, the communication from the transmitter, who saw the robot's state as

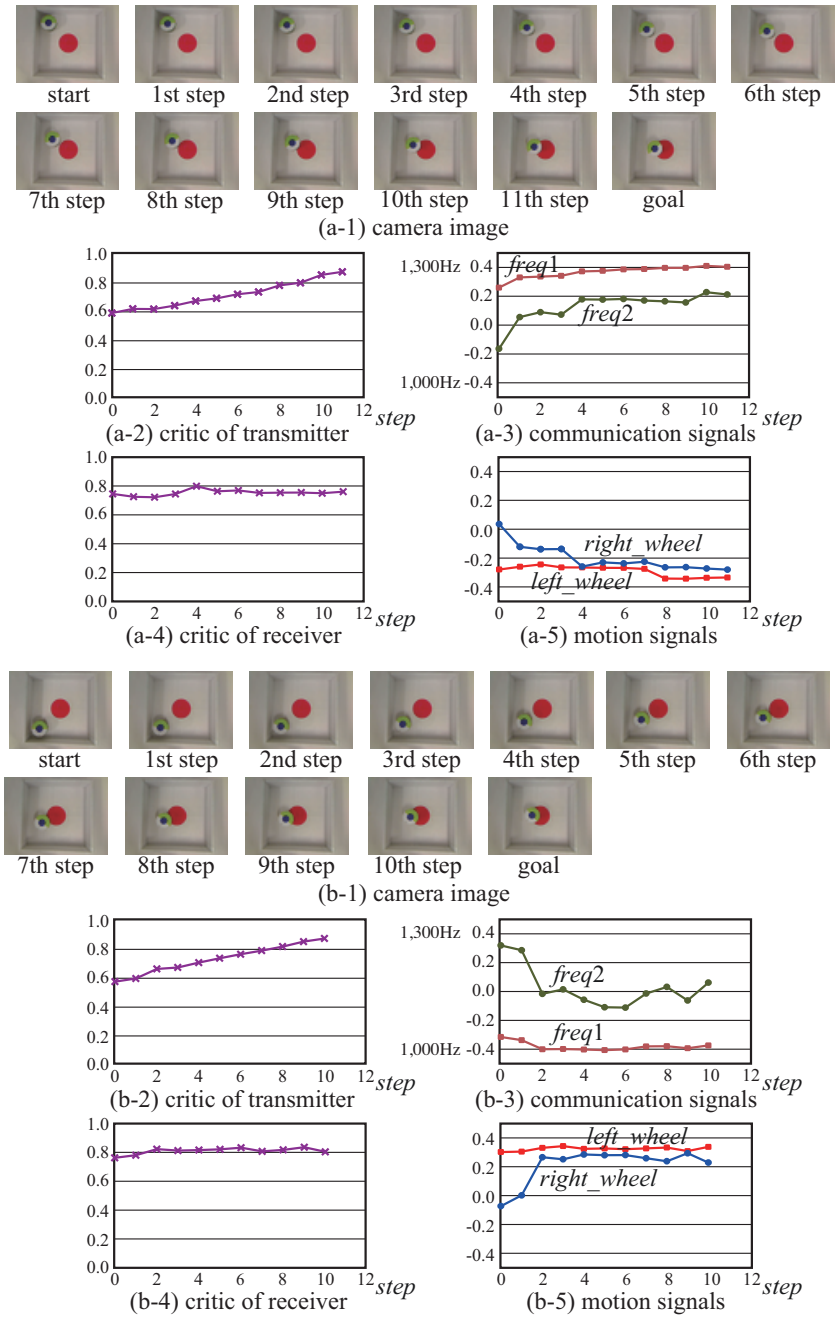


Fig. 5. The robot behavior and transmitter's and receiver's output changes in two sample episodes. Since the communication signals represent only appropriate motions and no value of state or action, the critic output does not increase in the receiver.

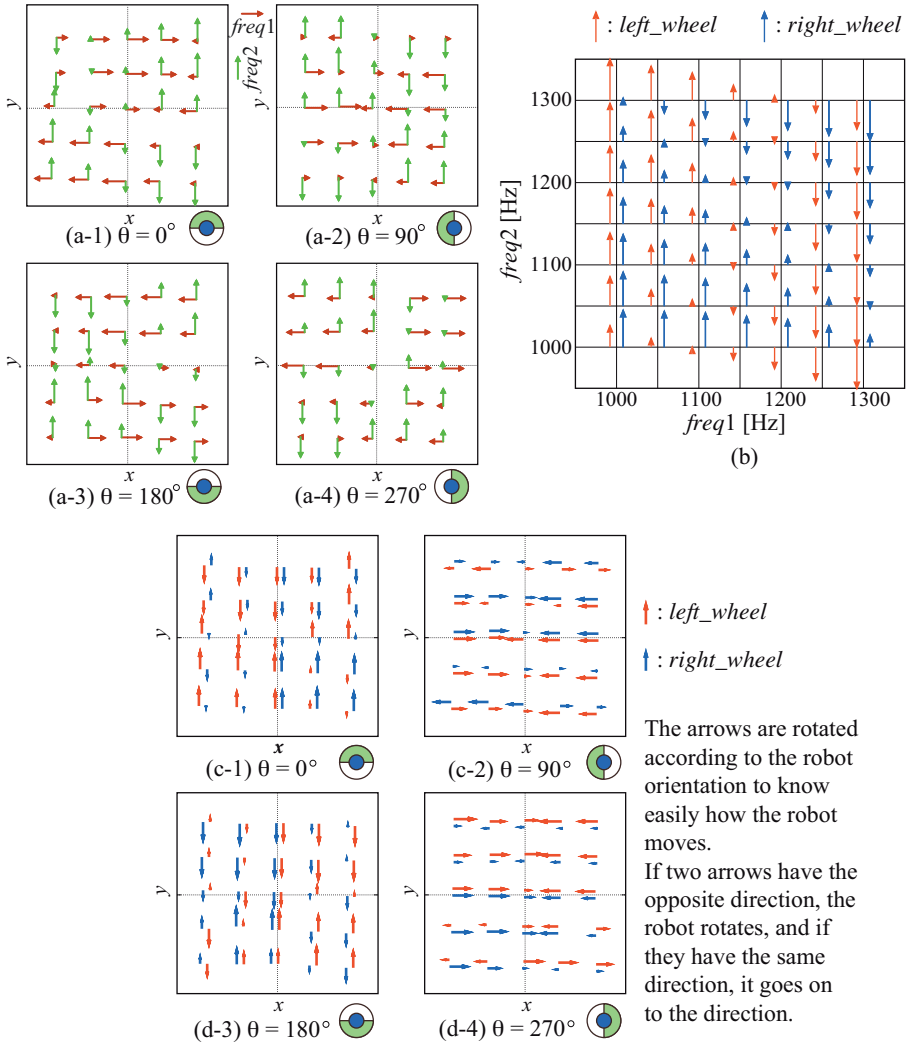


Fig. 6. (a)The frequency of communication signals $(freq1, freq2)$ (transmitter's actor outputs) for some robot locations (x, y). The position of the arrows indicates the robot location on the field. The robot orientation θ is different among (a-1, 2, 3, 4). That is also shown in the small robot image beside each figure. The pair of horizontal brown $(freq1)$ and vertical green $(freq2)$ arrow lengths shows the frequencies of the two signals. (e.g. 1,000Hz: longest in the upper or right direction, 1,150Hz: length is 0, 1,300Hz: longest in the lower or left direction) (b)The motion commands $(left, right)$ (receiver's actor outputs) for some combinations of two communication signals $(freq1, freq2)$. (c)The motion commands $(left, right)$ for some robot locations (x, y) and orientation θ pairs. The motion commands for each state is represented by a pair of red and blue arrows. The red arrows show the motion command for the left wheel, while the blue arrows show that for the right wheel.

the camera image, to the receiver, who generated the motion commands to the robot, could be established through reinforcement learning only from a reward and punishment. It is also claimed that in the communication learning, actual control experience in the transmitter, and also the correlation between the transmitted communication signal and the final effect are important. In this paper, the communication signals are continuous, and in this meaning, the "Symbol Grounding Problem" has not been solved. However, purposive and grounded communication that includes what should be communicated considering the situation through many sensor signals and also how should the communication signals be reflected on motions was acquired through learning without any specialized learning for communication.

Acknowledgment

This work was supported by JSPS Grant-in-Aid for Scientific Research #19300070 and #23500245.

References

1. Harnad, H.: Symbol Grounding Problem. *Physica D*, **42**, pp.335-346 (1990)
2. Nakano, K., Sakaguchi, Y., Isotani, R. & Ohmori, T.: Self-Organizing System Obtaining Communication Ability. *Biological Cybernetics*, **58**, pp.417-425 (1988)
3. Steels, L.: Evolving grounded communication for robots. *Trends in Cognitive Science*, **7**(7), pp. 308-312 (2003)
4. Werner, G.M. & DyerM.G.: Evolution of Communication in Artificial Organisms. *Proc. of Artificial Life II*, pp.1-47 (1991)
5. Ono, N. et al.: Emergent Organization of Interspecies Communication in Q-Learning Artificial Organs. *Advances in Artificial Life*, pp.396-405 (1995)
6. Shibata, K. & Ito, K.: Emergence of Communication for Negotiation By a Recurrent Neural Network. Proc. of ISADS '99, pp.294-301 (1999)
7. Nakanishi, M. & Shibata, K.: Effect of Action Selection on Emergence of One-way Communication Using Q-learning. *Proc. of AROB 10th* , CD-ROM, GS7-3 (2005)
8. Shibata, K.: Discretization of Series of Communication Signals in Noisy Environment by Reinforcement Learning. *Adaptive and Natural Computing Algorithms*, pp. 486-489 (2005)
9. Mitsunaga, N. et al.: Robovie-IV: A Communication Robot Interacting with People Daily in an Office. *Proc. of IROS'06* , pp. 5066-5072 (2006)
10. Suga, Y. et al.: Development of Emotional Communication Robot, WAMOBEA-3. *Proc. of ICAM'04* , pp.413-418 (2004)
11. Bennewitz, M. et al.: Fritz - A Humanoid Communication Robot. *Proc. of RO-MAN'07* , pp. 1072-1077 (2007)
12. Shibata, K.: Emergence of Intelligence through Reinforcement Learning with a Neural Network. *Advances in Reinforcement Learning*, InTech, pp.99-120 (2011)
13. Barto, A.G. et al.: Neuronlike Adaptive Elements That can Solve Difficult Learning Control Problems'. *IEEE Trans. of SMC*, **13**, pp.835-846 (1983)
14. Rumelhart, D.E. et al.: Learning Internal Representation by Error Propagation. *in Parallel Distributed Processing* (1986)
15. Nakanishi, M. et al.: Occurrence of State Confusion in the Learning of Communication Using Q-leaning. *Proc. of AROB 9th* , **2**, pp.663-666 (2004)