Towards Imitation Learning from a Viewpoint of an Internal Observer

Y. Yoshikawa*, M. Asada*† and K. Hosoda*†

* Dept. of Adaptive Machine Systems, † HANDAI Frontier Research Center,
Graduate School of Engineering, Osaka University
{yoshikawa, asada, hosoda}@er.ams.eng.osaka-u.ac.jp

Abstract. How an *internal observer*, that is not given any *a priori* knowledge or interpretation of what its sensors receives, learn to imitate seems a formidable issue from a viewpoint of a constructivist approach towards both establishing the design principle for an intelligent robot and understanding human intelligence. This paper argue two issues towards imitation by an internal observer: one concerns how to construct the self body representation of the robot with vision and proprioception and the other concerns how to construct a mapping of vocalization between agents with different articulation systems. Preliminary results with real robots are given.

1 Introduction

The ability of imitation has been focused in robotics – partially because learning by imitation is regarded as a promising way to accelerate the learning of a robot [1], and partially because it is also one of the most interesting cognitive issues to model human intelligence by a constructivist approach [2]. In the previous work, the designer usually provides specific knowledge to imitate a certain behavior (ex. [3]). However, to model how humans acquire the ability of imitation, we must also address the issue to design a robot that can imitate by itself. In this study, therefore, we assume that the robot is an *internal observer*. An internal observer is defined as an agent that is not given any *a priori* knowledge or interpretation of what the sensory signals it receives mean. By introducing the assumption that the agent can distinguish the different senor modalities, we can start to attack an issue how it can interpret its sensory signals by finding the relationships of its sensory data between different modalities. That is, association of the sensory data from different modalities.

For an internal observer to imitate, constructing a map between the observed demonstrator's body and its own one seems essential for a certain class of imitation where it can imitate through performing the mapped action of the other agent in the coordinate system of its own body based on this map. There are at least two issues to be addressed. First, it must possess the representation of its own body to associate it with other's body. This is not easy because the internal observer does not even know what its body is at the beginning. Another concerns how to construct a map between bodies of different agents without *a priori* knowledge about the relationship between them. To learn the map by itself, the robot needs to find references between them. We must consider the fact that the body of the robot is different from the other agent's one.

In the rest of this paper, we will introduce the preliminary results of our study. Concerning acquiring the representation of the body, we address the problem of finding its body in its uninterpreted sensory data [4]. A cross modal map is proposed as the learning structure based on the idea that the invariance in multisensory data represents the body. Concerning the construction of a mapping between different bodies, we address the problem of acquiring common vowels with the caregiver who has different articulation parameters from the robot [5]. We propose a model of interaction that guides a robot to acquire articulation to vocalize.

2 Acquisition of body representation [4]

One of the fundamental problems of acquiring a representation of the body is how to find the body in the receptive field without *a priori* knowledge from a viewpoint of external observer. Some previous studies proposed methods by which an agent distinguishes the body of the other agent and its own one based on the correlation between its motion and the motion-induced optical flow (e.g. [6, 7]). However, the agent could not distinguish its body from the environment without *a priori* knowledge how its motor system affects its vision. Although another study proposed a method by which an agent finds the boundary of its tactile sensor in the vision based on experience of collision [8], the agent needs to be taught which object in the vision collides with its body.

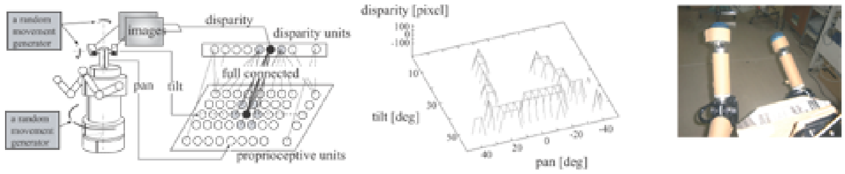
Sensation of its body is considered to be invariant with its posture. For example, when it fixates one object in the environment, the view changes depending on the environmental changes. However, when it fixates its body, the view is independent of the environment. Therefore, it is suggested that such invariance in multimodal sensors can be used to define its body. The robot can find the invariance through the experiences of taking various postures.

As a structure to find the invariance in multiple modalities, we introduce a fully-connected network called *cross modal map* (see Fig. 1(a)). A cross modal map consists of various sensor nodes that are hardwired to real sensors and are activated when the hardwired sensors receive something in their receptive fields. After Hebbian learning, only the weights between the nodes that are simultaneously activated during a certain period of time increase. Since the same pair of nodes are considered to be simultaneously activated in the sensation of self-body, the connections which have large synaptic weights are regarded to represent the body.

2.1 Experiment

A preliminary experiment to learn to represent the body surfaces of the robot by the cross modal mapping between 15 nodes for binocular vision (disparity at the center region of the left camera) and 20 × 15 nodes for proprioception (joint angles) was performed. Fig. 1 (b) is a section of the acquired cross modal map in which the arms have a certain postural configuration after about six minutes learning. During the learning process, the robot keeps changing its posture at random. It shows which disparity node has the largest connection with which posture node as a function of the disparity with respect to pan and tilt angles of the camera head. The shape of the function resembles an egocentric view of the robot (see Fig. 1 (c)). The fixation areas of which disparity node have strong connections (large weights) to the posture nodes were parts of the robot body. Therefore, the robot succeed in learning the cross modal map that represents the body surface of the learner.

Since the sensors of the robot are embedded on its own rigid body, the sensation of self body is constrained to be invariant with its proprioception. However, by using the representation of the invariance, the robot can only judge whether the fixated point in the vision is its body or not. We should extend the proposed method for the concept of body part. Then, we should address many issues such as representing kinematics/dynamics, representing the reachable region by the robot movements, and the establishment of the correspondence between its own body and the other's.



(a) The architecture

(b) A section of the acquired body representation

(c) A real, egocentric view

Fig. 1. The architecture of the robot with a cross-modal map (a), a section of the acquired cross modal map (b) and an example of the egocentric view of the robot (c)

3 Acquisition of common vowels [5]

Infants, who are internal observers, seem to acquire the phonemes of adults without *a priori* knowledge about the correspondence between its vocalization and the phonemes. Previous studies showed that computer simulated agents with a vocal tract and cochlea can acquire shared vowels in population by self-organization through interaction with other agents [9, 10]. Although they didn't assume *a priori* knowledge about vowel, there was an assumption that the agents can reproduce the similar sounds of other agents' so that "imitation game [9]" or "magnet effect [10]" makes self-organized vowels shared in population. However, infants face with more difficult situations. First, they cannot reproduce the caregiver's utterances as they are because their vocalization system is not mature. Furthermore, even if they can imitate the adult phoneme, they perceive the reproduced sounds differently from the caregiver's original sounds because the sound wave of the former travels inside the body to the infant's auditory sensors. In this case, imitation cannot be equated with raw sensory similarity. To take infant's immaturity into account for modeling the infant's acquisition process of vowels, we use a robot that consists of an artificial articulatory system with a 5-DoFs mechanical system that can deform a silicon-made vocal tract connected to an artificial larynx (see Fig. 2). It vocalizes some sounds which can be interpreted as human vowels but are different from the human vowels from a viewpoint of low-level signal similarity.

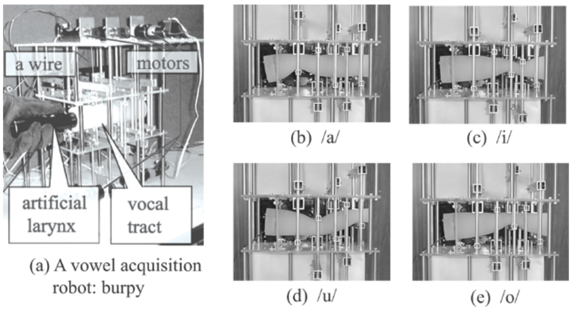
It is reported that maternal imitation of a three-month-old infant's cooing (i.e., parrot-like utterances) increases the vocalization rates [11], and the infant's speech-like cooing tends to lead the mother's utterances [12]. Based on these observations, we conjecture that the caregiver's imitation of the robot's vocalizations plays an important role in the vowel acquisition process – in other words, a regular reaction (a parrot-like behavior), which can be regarded as action invariance, make it possible to acquire vowels instead of actions that produce similar sensory information. As a preliminary, constructive model of our conjecture, we design a random articulation mechanism and embed it in the robot so that an interaction can emerge between the robot and the caregiver who produces its own corresponding vowel when the robot's articulated sounds can be heard as the vowel.

The learning mechanism consists of auditory and articulation layers and connections between them. The auditory layer clusters formants (i.e., sound features) of the caregiver by self-organization while the articulation layer clusters its own articulation parameters. The connections between them are updated according to Hebbian learning. The robot learns through interaction to match its articulation with audition, that is, it acquires the vowel sounds of the caregiver. However, interactions may connect multiple articulation units with a corresponding vowel since the caregiver will interpret some vocalizations caused by different articulations as the same vowel. To match a listened vowel with a unique articulation, we introduce *subjective criteria*, that are evaluated only in terms of the robot's state, into the learning rule — that is, the articulation vectors with less torque and less intensity of deformation changes obtains stronger connection from auditory layer and vice versa.

3.1 Experiment

We examined whether the robot can acquire Japanese vowel sounds by interacting with a caregiver. After the robot vocalizes by a random articulation vector, the human caregiver determines whether the robot's vowel corresponds to the

Japanese vowel and utters the corresponding vowel. The robot calculates formants of the caregiver's vocalization and updates connections between nodes to represent the caregiver's utterances and ones to represent the robot's articulation by Hebbian learning with and without subjective criteria. Each element of a code vectors in the articulation layer is quantized into five levels; these elements are the motor commands of the random articulation mechanism. Fig.2(b)–(e) shows the acquired articulations. The vocalized sound produced by these articulations can be interpreted as being Japanese vowels.



(a) A vowel acquisition robot: burpy

(b) /a/

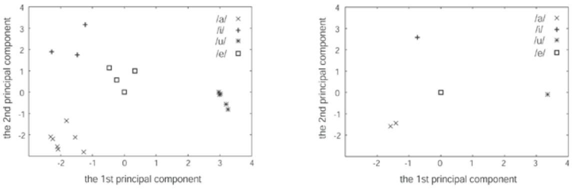
(c) /i/

(d) /u/

(e) /o/

Fig. 2. The appearance of the test-bed robot (a) and the acquired vowels (b)–(e)

We observed which units in the articulation layer are activated by the propagation of the activation in the auditory layer when the caregiver utters one of vowels. The activated unit in the articulation layer can be regarded as the matched vowels with the caregiver's one. Fig. 3(a) shows the distribution of the matched articulation acquired by the normal learning rule without subjective criteria, while Fig. 3(b) shows one by the learning rule with subjective criteria. We can see that fewer articulations are selected in the learning with subjective criteria. Therefore, we confirmed that the subjective criteria decreased the number of units in the articulation layer that are activated by the auditory layer. The selected articulation were more facile to articulate.



(a) without subjective criteria

(b) with subjective criteria

Fig. 3. The acquired clusters without subjective criteria (a) and with it (b)

4 Conclusion

As a preliminary work on understanding the mechanism of imitation by an internal observer, we studied the issues of acquiring the vowel sounds of a caregiver and acquiring a body representation based on constructing mappings between different modalities. Although the robots explored at random to construct the mappings in the both proposed model, they had better utilize their developing mappings to accelerate the learning process. Furthermore, they should learn to use the acquired mappings toward various cognitive functions. Therefore, how to motivate the robot to learn and use mapping is one of our future topic.

Acknowledgment This study is partially supported by "the Advanced and Innovational Research program in Life Sciences" and "The 21st Century COE Program (Project: Center of Excellence for Advanced Structural and Functional Materials Design)" both of which are programs of the Ministry of Education, Culture, Sports, Science and Technology of Japan, and also by Research Fellowships for Young Scientists, a program of Japan Society for Promotion of Science.

References

1. Schaal, S.: Is imitation learning the route to humanoid robots? Trends in Cognitive Science **3** (1999) 233–242
2. Asada, M., MacDorman, K.F., Ishiguro, H., Kuniyoshi, Y.: Cognitive developmental robotics as a new paradigm for the design of humanoid robots. Robotics and Autonomous System **37** (2001) 185–193
3. Kuniyoshi, Y., Inaba, M., Inoue, H.: Learning by watching: Extracting reusable task knowledge from visual observation of human performance. IEEE Transaction on R&A **10** (1994) 799–821
4. Yoshikawa, Y., Hosoda, K., Asada, M.: Does the invariance in multi-modalities represent the body scheme? - a case study with vision and proprioception -. In: Proc. of the 2nd Intl. Symp. on Adaptive Motion of Animals and Machines. (2003)
5. Yoshikawa, Y., Asada, M., Hosoda, K., Koga, J.: A constructive approach to infant's vowel acquisition through mother-infant interaction. Connection Science **15** (2003) 245–258
6. Fitzpatrick, P., Metta, G.: Toward manipulation-driven vision. In: Proc. of the IROS'02. (2002) 43–48
7. Asada, M., Uchibe, E., Hosoda, K.: Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development. Artificial Intelligence **110** (1999) 275–292
8. MacDorman, K.F., Tatani, K., Miyazaki, Y., Koeda, M.: Proto-sysmbol emergence. In: Proc. of the Intl. Conf. on Intelligent Robot and Systems. (2000) 1619–1625
9. de Boer, B.: Self-organization in vowel systems. J. of Phonetics **28** (2000) 441–465
10. Oudeyer, P.Y.: Phonemic coding might result from sensory-motor coupling dynamics. In: Proc. of the 7th intl. conf. on simulation of adaptive behavior. (2002)
11. Peláez-Nogueras, M., Gewirtz, J.L., Markham, M.M.: Infant vocalizations are confitioned both by maternal imitation and motherese speech. Infant behavior and development **19** (1996) 670
12. Masataka, N., Bloom, K.: Accoustic properties that determine adult's preference for 3-month-old infant vocalization. Infant Behavior and Development **17** (1994) 461–464