

Generation of Robot Motions from Environmental Sounds Using Inter-modality Mapping by RNNPB

Tetsuya Ogata[†], Yuya Hattori[†], Hideki Kozima[‡], Kazunori Komatani[†], and Hiroshi G. Okuno[†]

[†]*Graduate School of Informatics, Kyoto University, Kyoto, Japan*

{ogata, yuya, komatani, okuno}@kuis.kyoto-u.ac.jp

[‡]*National Institute of Information and Communication Technology, Kyoto, Japan*

xkozima@nict.go.jp

Abstract

This paper proposes mapping between different sensory modalities for a robot system to generate motion expressing auditory signals or sounds from the movements of objects. Since all correspondences between auditory signals and visual signals in the world are hard to memorize, the ability to generalize is indispensable. We adopted a neural circuit model called RNNPB, which has good generalization ability, for the learning model. We implemented the proposed system on the robot “Keepon.” We taught it horizontal reciprocating or rotating motions with fricative sounds and falling or overturning motion with the sounds of collision by manipulating a box object. Keepon behaved not only from learned events but also from unknown events. It could also generate various sounds according to observed motions.

1. Introduction

Various kinds of robot systems that interact with humans have recently received a great deal of attention, represented by increased interest in humanoid robots (H. Ishiguro, et al., 2001), pet robots, and other human assistant robots. These robots have to react to multi-modal sensory input to execute tasks and communicate with human operators. Most conventional studies so far have handled these multi-modal sensory data independently. After information processing has been done for each modality, they have been synchronized and integrated. However, it is generally quite difficult to design this process of integration. This paper proposes a novel method that enables robots to handle multi-modal information simultaneously.

Humans can deal with ‘cross-modal information’. For example, they can express auditory information (e.g. sounds of collision) by

using visual expressions like gestures (e.g. move the hand fast and stop it sharply). We call this ‘inter-modality’ mapping.

The final goal of this study is to design a technological method for inter-modal mapping. From the aspect of engineering, the method is expected to be able to be applied to a robot’s motion generator from various sound signals, and an effective sound generator corresponding to various images. From the aspect of recognition science, we expect that the method would provide various findings concerning *synaesthesia* as will be described in the Discussion section.

Section 2 presents the outline of our idea in which a robot acquires the relations between different items of modal information by observing various events. Section 3 introduces the neural network model used for association/translation between inter-modalities, and the generalization of multi-modal sensory dynamics obtained from experience with observation. Section 4 describes the concrete implementation of our system with a small robot called Keepon. Section 5 presents the experimental results on inter-modality mapping. Section 6 discusses the characteristics of our method and its relation with findings from cognition science. Section 7 summarizes our work.

2. Model of Inter-modality Mapping

As mentioned in Section 1, all sensor modalities are processed separately in almost all conventional robot systems. However, all the sensory information we experience is actually input simultaneously. We proposed a procedure in the following that is an interpretation of inter-modality mapping (Hattori, Y. et al. 2005). It is mainly divided into two phases, the first is

the “learning phase” and the second is the “interaction phase”.

a) Learning Phase (Look Sounds)

In this phase, the robot observes some events with various kinds of sounds, such as the sound of collision, sound of friction, continuous sound, and rhythmical sound (See Fig. 1). The robot memorizes these sounds with the motions of the sound source. We figuratively call this the “robot looking at sounds” phase.

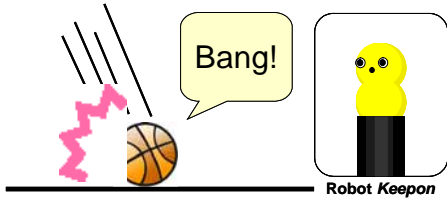


Fig. 1 Robot looking at sounds

b) Interactive Phase

In this phase, limited sensory information from a single modality is input to the robot. The robot associates this with different modality information, and expresses it. For example, the robot generates the motion of a sound source from a given sound by using its body motion (Fig. 2). Conversely, the robot outputs the sound from the observed motion (Fig. 3).

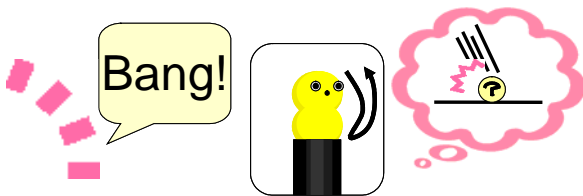


Fig. 2 Mapping from sound to motion

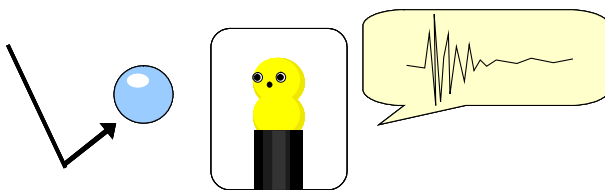


Fig. 3 Mapping from motion to sound

3. Neural Network Model

a) Introduction of RNNPB model

This section introduces a method that enables the robot to deal with multi-modal information simultaneously. There are various sounds in the environment around us. It is

almost impossible to construct a database that can systematically store all environmental sounds. To achieve the inter-modal mapping, it is indispensable for the robot to have the ability to generalize various sounds from limited sounds it can actually observe. A robot is expected to adapt to unknown stimuli by using this.

Based on the above, we introduced the artificial neural network model called Recurrent Neural Net with Parametric Bias (RNNPB) proposed by Tani (Tani, J. and Ito, M. 2003). The main characteristic of the RNNPB is that chunks of sequence patterns of the sensory-motor flow which can be represented by a vector of small dimensions. This vector plays the role of the bifurcation parameters of nonlinear dynamical systems. In other words, different vector values make the system generate different dynamic patterns. The main advantage of utilizing the parameter bifurcation is that ideally the RNNPB can encode infinite number of dynamic patterns with modulating analog values of the vector of small dimensions.

RNNPB is usually designed as a predictor (forwarding forward model) for which input is a current condition $S(t)$ and for which output is the next condition $S(t+1)$. RNNPB has the same structure as the Jordan type RNN (Jordan, M., 1986) except that it has parametric bias in the input layer (See Fig. 4). Unlike other input nodes, these PB nodes have a constant value throughout each time sequence. The context layer has a loop that inputs current output as input data in the next step. An advantage of this layer is that the RNNPB model can learn the time sequences taking advantage of past contexts.

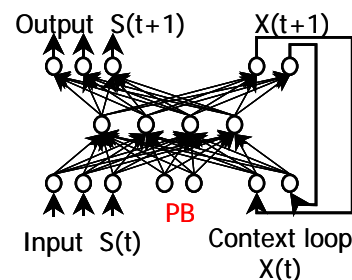


Fig. 4 RNNPB

The RNNPB has three activation modes, i.e., the learning, the prediction, and generation modes.

In the learning mode, the RNNPB updates its weights and the value for the parametric

bias' simultaneously using the BPTT method (Rumelhart, D. et al., 1986) with prediction error. Each update is carried out using the equations below. The step length of a sequence is denoted by l . For each sensory-motor output, back-propagated errors with respect to PB nodes are accumulated and used to update the PB values. The update equations for the i th unit of the parametric bias at the t in sequence are

$$\delta\rho_t = k_{bp} \cdot \sum_{t-1/2}^{t+1/2} \delta_t^{bp} + k_{nb} (\rho_{t+1} - 2\rho_t + \rho_{t-1}), \quad (1)$$

$$\Delta\rho_t = \varepsilon \cdot \delta\rho_t, \quad (2)$$

$$p_t = \text{sigmoid}(\rho_t / \zeta). \quad (3)$$

In Eq. (1), the δ force for updating the internal values of the PB, p_t , is obtained from the summation of two terms. The first term represents the delta error, δ_t^{bp} , back-propagated from the output nodes to the PB nodes: it is integrated over the period from the $t-1/2$ to the $t+1/2$ steps. Integrating the delta error prevents local fluctuations in the output errors from significantly affecting the temporal PB values. The second term is a low-pass filter that inhibits frequent rapid changes in the PB values. Internal value ρ_t is updated using the delta force, as shown in Eq. (2). Further k_{bp} , k_{nb} , and ε are coefficients. The current PB values are then obtained from the sigmoidal outputs of the internal values. After learning time sequences, the RNNPB model self-organizes the PB values at which the specific properties of each individual time sequence are encoded, and can generate a sequence from the corresponding PB values. Our goal is to acquire the specific parameter values corresponding to each event. Therefore, in order to fix the parameter values during the sensing motion, Eq. (1) was simplified in our RNNPB model training as follows.

$$\delta\rho_t = k_{bp} \cdot \sum_0^{L_s} \delta_t^{bp} \quad (4)$$

In the recognition mode, the corresponding PB value for a given sequence can be obtained by using the update rules for the PB values (Eqs. (1) to (3)) without updating the connection weight values. In the generation mode, the PB value for a desired sequence is set to the PB node. The desired sequence is obtained by

carrying out forwarding-forward calculation for RNNPB.

The other important characteristic of the RNNPB model is that the relational structure among training sequences can be acquired in the PB space through the learning process. This capability enables the RNNPB model to generate and recognize unseen sequences without the need for any additional learning.

b) Modality mapping using RNNPB model

In the learning phase discussed in Section 2-(a), a robot with RNNPB learns various events (sensor sequences) by using equations (1), (2), and (3). This section describes how the RNNPB is used in the interaction phase discussed in Section 2-(b) by considering an example of modality translation from sound to motion.

Figure 5 outlines the concept of the BPTT algorithm for RNNPB. When the robot only detects an auditory signal, the PB values are calculated only using the prediction error in the auditory signal. In this case, the input/output layers for the visual signal are handled as the same as the context layer.

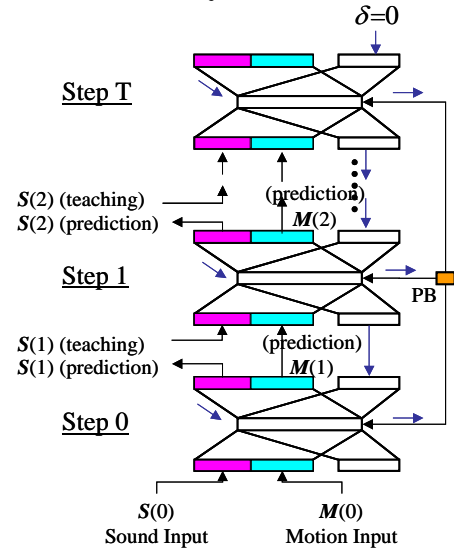


Fig. 5 BPTT algorithm for RNNPB

After obtaining the PB values, the sequence of visual signal is obtained by the following process. The PB values and the real auditory signal are input to the RNNPB in each step and forwarding-forward calculation is carried out. The input/output layer for the visual signal is regarded as the context layer; then, the visual sequence is obtained.

4. Implementation into robot system

a) Interaction robot, Keepon

We used a robot called “Keepon” for our experiments. It was developed at the National Institute of Information and Communication Technology (NICT) mainly for communicative experiments with infants (Kozima, H. et al., 2004). Its body is approximately 12 cm high with 4 degrees of freedom as shown in Fig. 6. It is equipped with two CCD cameras and one microphone on its head.

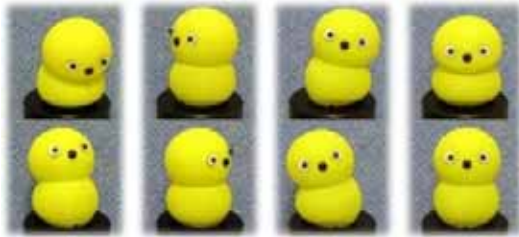


Fig. 6 Robot Keepon and its motions

b) Audio-visual Processing

In the experiments, the robot observed events involving a box manipulated by human. The length of the box is 165 mm, the width is 110 mm and the height is 33 mm. It was made of plastic. In visual processing, the corner positions of the box were detected during operation. In audio processing, we used the values for the Mel Filter Bank multiplied by the four triangular windows in Fig. 7. These values were normalized and synchronized in 50 [ms] for input to RNNPB.

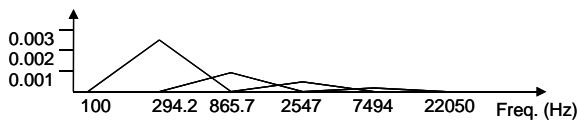


Fig. 7 Triangular windows

c) Generation Processing

An inverse translation process from the audio-visual signal obtained by RNNPB to actual robot motion and the sound is required for modality mapping.

In the motion generation phase, Keepon reproduces the trajectory obtained from RNNPB output using its pitch and yaw axis. In the sound generation phase, Keepon outputs colored noise by multiplying white noise with the Mel Filter Bank value obtained from the RNNPB

output. The actual frequency and time data were obtained using a linear approximation due to poor resolution in the RNNPB output.

5. Experiments

a) Learning and PB space

In the learning phase, we made Keepon observe four kinds of manipulations of the blue box with different types of sound. These were 1) rotation on the wall with the sound of continuous friction, 2) reciprocating on the table with periodic sounds of friction, 3) overturning on the table with the sound of collision, and 4) falling down to the table with the sound of collision. Keepon observed each event three times (the total sequences were $4 \times 3 = 12$), and the RNNPB was trained using these collected data. The RNNPB consisted of 8 neurons in the input layer, 35 neurons in the middle layer, 25 neurons in the context layer, and 2 neurons as parametric bias. The training sequence for the RNNPB was segmented when the changed values for all sensory inputs were less than the threshold. The event lengths were 10 to 40 steps (0.5 - 2 sec) in the experiments.

Figure 8 plots the acquired PB space. The two parametric values in the RNNPB correspond to the X-Y axes in the space. We confirmed that the reciprocating motions were mapped in the upper area, the rotating motions were mapped in the left area, the overturning motions were mapped in the left-bottom area, and the falling downs motions were mapped in the right area. The distance between the areas for rotating motions and overturning motions is close. This is because overturning motion can be regarded as part of rotating motion.

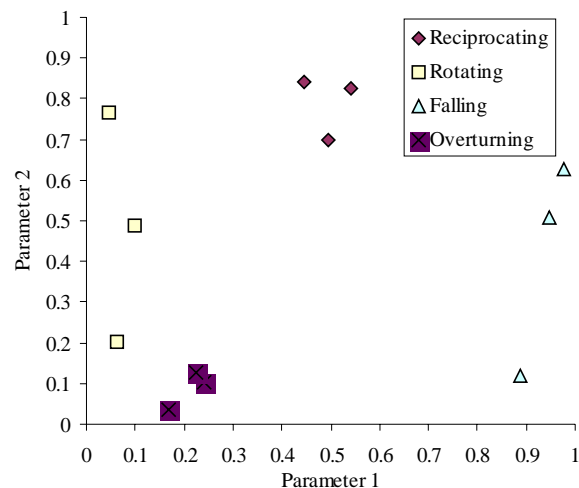


Fig. 8 PB space acquired in learning phase

b) Mapping from sounds to motions

Figure 9 plots the sound recognition results for the RNNPB. The circled plots in Figure 9 denote the four events described in Section 5-(a). These were not used to train the RNNPB. We also investigated the PB values corresponding to completely novel sounds generated by 1) a spray jet, 2) clapping, and 3) a plastic bag being shaken randomly a few times. The PB values corresponding to these events are also plotted in Fig. 9.

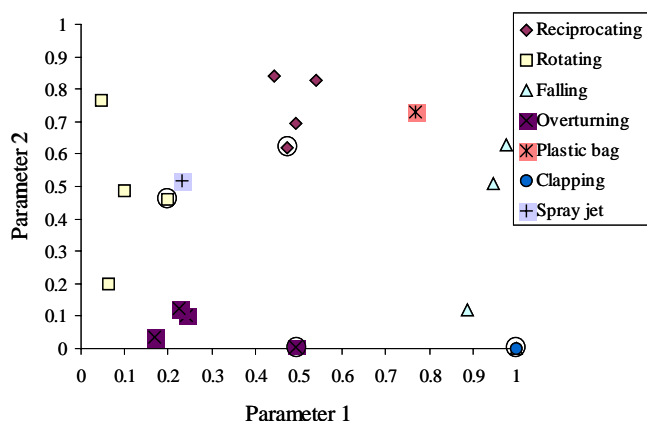


Fig. 9 Sound recognition results

Fig. 10 plots the motion trajectories corresponding to 'known' events. We confirmed that Keelson could generate trajectories that were similar to the manipulations of the blue-box in the learning phase.

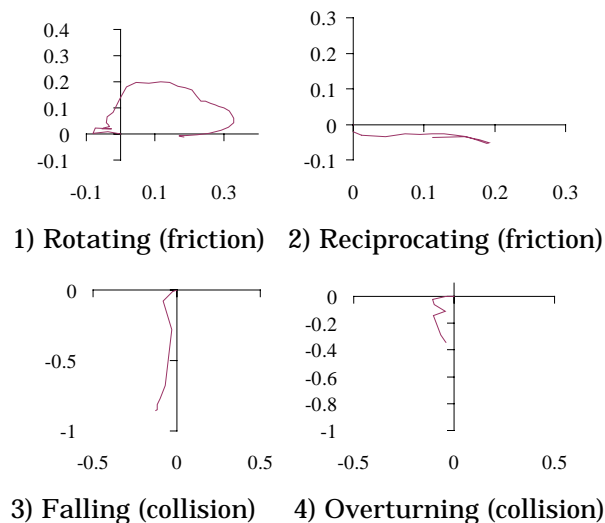


Fig 10 Trajectories generated by sounds corresponding to four events

Figures 11, 12, and 13 show the 'unknown' sounds and results for generated motion. The generated motion for the clapping sound was similar to that for the falling manipulation. This might be because they were common in the sense they were 'collision sounds'. Similarly, the generated motion for the sound of the spray jet was similar to that for the rotating manipulation. This was because they were common in the sense they were 'continuous friction sounds'.

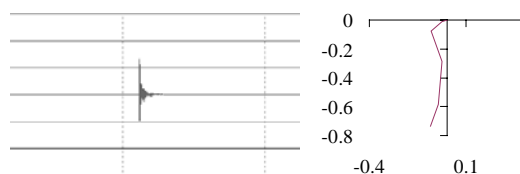


Fig. 11 Clapping sound and generated motion

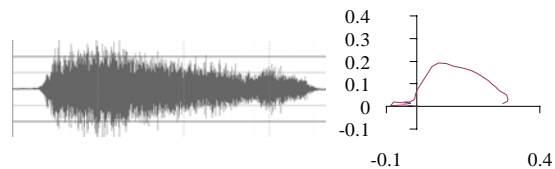


Fig. 12 Sound of spray jet and generated motion

It should be noted that the generated motion from the sound of the plastic bag being shaken was not as simple as the previous two examples (see Fig. 13). This sound contains not only friction/collision but has various other features. The motion reflected such complex characteristics as shaking up and down and small rotations. The ability of the RNNPB to generalize can generate such novel motion patterns from unknown sounds.

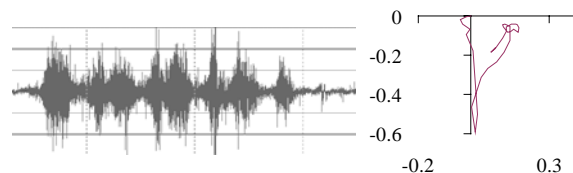


Fig. 13 Sound of plastic bag being shaken and generated motion

c) Mapping from motions to sounds

Figure 14 plots the motion recognition results for the RNNPB. The circled plots in the PB space denote the four events described in Section 5-(a). These were not used to train the RNNPB. We also investigated the PB values corresponding to novel motions without sound,

such as 1) moving the box quickly horizontally, 2) moving the box slowly horizontally, and 3) shaking the box up and down a few times. The PB values corresponding to these events are also plotted in Fig. 14.

Figure 15 shows the generated sounds corresponding to 'known' events. We confirmed that Keepon could generate sounds that were almost the same as sounds observed in the learning phase. The reason for the two power peaks in 2) and 4) is that the RNNPB learnt the rebound sound from the falling-down event.

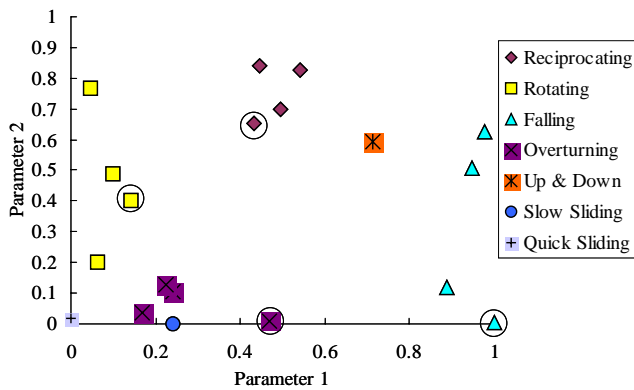


Fig. 14 Motion recognition results

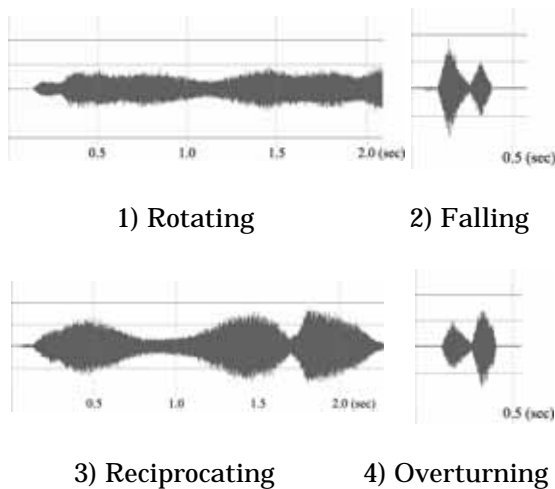


Fig. 15 Sounds generated by motions corresponding to four events

Figures 16, 17, and 18 show the 'unknown' motions and results for the sound generation. The generated sound for the quick motion is similar to that for the falling-down manipulation, whereas there is no sound. This may be because they are common in the sense they are both 'collisions'. Similarly, the generated sound for the slow motion is similar to that for the rotating manipulation. This is may be because they are common in the sense

they are 'continuous motions'.



Fig. 16 Generated sound for quick slide motion

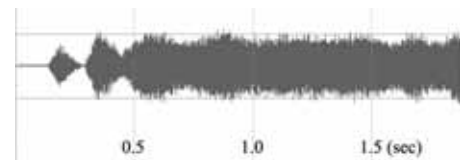


Fig. 17 Generated sound for slow slide motion

We investigated the sound output when a novel motion pattern was observed as the same as for the experiment on motion generation. Figure 18 shows the sound output when Keepon observed the box being shaken up and down without any sound input. The wave cycle is different to that for the reciprocating manipulation used for training in Fig.15. It synchronised the observed motions.

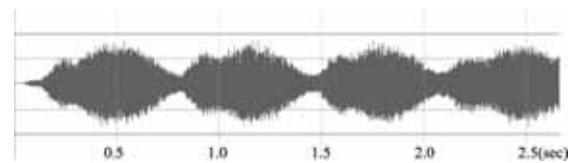


Fig. 18 Generated sound for shaking up and down

6. Discussion

a) Generalization by RNNPB

We confirmed that the RNNPB has generalization capabilities for clustering various events. It can also express relationships between events in the PB self-organized space. This generalization is quite difficult by only using symbol-base expressions like a database and feature mapping methods.

The RNNPB acquires multi-attractors in an overlapping fashion in a single network by changing parameters that represent the boundary condition (distributed expression). In the RNNPB, all neurons and synaptic weights contribute to represent all trained patterns. In such distributed expressions, memory interference will occur since memories share the same network resources. Nevertheless, as a

result of embedding multiple attractors in a distributed network, we obtained a global structure that could handle learned patterns as well as unknown (unlearned) patterns. This may be why RNNPB could exhibit the generalization abilities in recognizing unknown events.

b) Bouba-kiki Effect

The *bouba-kiki effect* is one of the most important for investigating inter-modality mapping from the epigenetic systems perspective. Ramachandran et al. found a sharply-angled shape and a rounded shape as depicted in Fig. 19 and asked subjects “which shape is bouba and which is kiki?” As a result, 95% of them answered the sharply-angled shape was kiki (Ramachandran, V. S. et al., 2003). The words bouba and kiki are meaningless words given as words for the subjects.

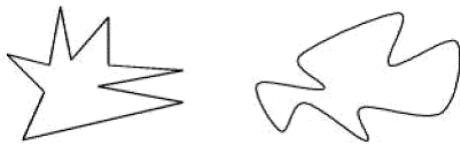


Fig. 19 Test figure in “bouba-kiki” effect

The bouba-kiki effect is known to indicate *synaesthesia*, an innate cross-modality phenomenon. This effect can, however, be interpreted as follows. When objects move along a sharply-angled shape as depicted at the left in Fig. 19, objects make sounds with a rising edge, which are similar to “kiki” in power modulations, on the edge. When objects move along a rounded shape as depicted at the right of Fig. 19, objects make sounds with a gradual edge, which are similar to “bouba”. Based on this consideration, it can be said that the robot Keepon with our method could replay to the Ramachandran question in almost same way as his subjects.

Bahrick et al. proposed a concept of “*amodal*” which is information not specific to a single sense modality, but is completely redundant across more than one sense (Bahrick, L. E. et al. 2004). For example, the face and voice of a person speaking share temporal synchrony, rhythm, tempo, and changing intensity. By selectively attending to these *amodal* properties, perceive can attend to the unitary event, that is, “speaking”. We think our model has a strong relationship with the

concept of “*amodal*”, and it could be an interpretation for explaining “bouba-kiki effect”.

7. Summary

This paper proposed a method of mapping between different sensory modalities for a robot system to generate motions expressing auditory signals or sounds from the movements of objects. Since all correspondences between auditory signals and visual signals in the world are hard to memorize, the generalization ability is indispensable. We adopted a neural circuit model called RNNPB, which has good generalization ability, for the learning model.

We implemented the proposed system on the robot “Keepon.” We made Keepon observe horizontal reciprocating or rotating motion with fricative sounds and falling or overturning motions with the sounds of collision by manipulating a box object. Keepon behaved not only from learned events but also from unknown events. It could also generate various sounds according to observed motions.

An interesting challenge for future work is to apply our method to a humanoid robot with many degrees of freedom. One crucial problem is how to select the joints to express the sounds. The joints available for Keepon discussed in this paper were selected in advance. It is well known that human infants learn what muscle movements achieve a particular goal. This process called ‘body babbling’ enables infants to acquire mapping between movement and the resulting body-part configuration. We will try to introduce a ‘body babbling’ process into our future humanoid robot.

References

- Bahrick, L. E., Lickliter, R., and Flom, R. (2004). Intersensory Redundancy Guides the Development of Selective Attention, Perception, and Cognition in Infancy, American Psychological Society, Vol. 13, No. 5, pp. 99-102.
- Hattori, Y., Ogata, T., Kozima, H., Komatani, K., Okuno, G. H., (2005) “Robot gesture generation from environmental sounds using inter-modality mapping”, International Workshop on Epigenetic Robotics (EpiRob-2005), pp.139-140.
- Ishiguro, H., Ono, T., Imai, M., Maeda, T., Kanda, T., and Nakatsu, R. (2001), “Robovie: an interactive humanoid robot.” International Journal of Industrial Robotics, Vol. 28, No. 6, pp.498-503.

Jordan, M. (1986) "Attractor dynamics and parallelism in a connectionist sequential machine." in Proc. of the Eighth Annual Conference of the Cognitive Science Society (Erlbaum, Hillsdale, NJ), pp. 513-546.

Kozima, H., Nakagawa, C., Yasuda, Y., Kosugi, D. (2004), "A toy-like robot in the playroom for children with developmental disorders", International Conference on Development and Learning (ICDL-2004; San Diego, CA, USA).

Ramachandran, V. S. and Hubbard, E. M. (2003). Hearing Colors, Tasting Shapes, SCIENTIFIC AMERICAN, Vol. 288, No. 5, pp. 53-59.

Rumelhart, D., Hinton, G., and Williams, R. (1986), "Learning internal representation by error propagation", in D.E. Rumelhart and J.L. McClelland, editors, Parallel Distributed Processing (Cambridge, MA: MIT Press).

Tani, J. and Ito, M. (2003) "Self-Organization of Behavioral Primitives as Multiple Attractor Dynamics: A Robot Experiment", IEEE Transactions on Systems, Vol. 33, No. 4, pp. 481-488.