Berthouze, L., Kaplan, F., Kozima, H., Yano, H., Konczak, J., Metta, G., Nadel, J., Sandini, G., Stojanov, G. and Balkenius, C. (Eds.)
Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems
Lund University Cognitive Studies, 123. ISBN 91-974741-4-2

Robot Gesture Generation from Environmental Sounds Using Inter-modality Mapping

Yuya Hattori* **Hideki Kozima**** **Kazunori Komatani*** **Tetsuya Ogata*** **Hiroshi G. Okuno***

*Graduate School of Informatics, Kyoto University, Japan
{yuya, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

**National Institute of Information and Communication Technology, Japan
xkozima@nict.go.jp

Abstract

We propose a motion generation model in which robots presume the sound source of an environmental sound and imitate its motion. Sharing environmental sounds between humans and robots enables them to share environmental information. It is difficult to transmit environmental sounds in human-robot communications. We approached this problem by focusing on the iconic gestures. Concretely, robots presume the motion of the sound source object and map it to the robot motion. This method enabled robots to imitate the motion of the sound source using their bodies.

1. Introduction

Based on advances in information technologies, the widespread use of robots is expected. Robots interacting with humans in the real world are supposed to do so using diverse modalities as humans do. In particular, we focus on various kinds of non-verbal sounds in our surroundings such as a door-opening sound and the cry of an animal, called "environmental sounds." Since environmental sounds are very important clues to understanding the surroundings, it is very useful to share the information of the environmental sounds with humans and robots. Ishihara et al. developed a sound-to-onomatopoeia translation method for such interactions (Ishihara et al., 2004), for example. In this paper, we focus on gestures in order to interact about environmental sounds.

2. Motion Generation using Inter-modality Mapping

2.1 Definition of Inter-modality Mapping

Humans ordinarily perceive events in the real world as the stimuli of multiple modalities such as vision and audition. Accordingly humans can express the stimuli in each modality. In the real environment, information from all modalities is not obtained properly. Optical information can have occlusions, for example. In such cases, humans can complement losses from properly obtained information. We defined **inter-modality mapping** as mapping from information of properly obtained modalities to the information of modalities not obtained.

In this paper, we focus on mapping from input sounds to motions. It is because visual and auditory information are remarkably important in humans' communications. It has been reported that children often use onomatopoeia and a gesture simultaneously and that the linkup of them is important for the development of multi-modality interactions (Werner and Kaplan, 1963).

2.2 Iconic Gesture Generation

We aim to generate motions expressing environmental sounds by imitating the motions of the sound source objects. It is because the kinds of environmental sounds are closely related to the motion of the sound source. It is known as **iconic gestures from an observer viewpoint** to imitate the motion of objects (McNeill, 1992). Iconic gesture means imitating concrete circumstances or events using one's body.

In our model, robots memorize the motion of the sound source, which is captured by their camera, when they listen to a sound with looking the sound source. After learning the correspondences between the sound and the motion, they can imitate the motion of the sound source when they listen to a sound without looking at the sound source. Namely, they can perform iconic gestures.

3. System Implementation

3.1 Tasks and System Overview

In order to make iconic gestures, the system learns connections between the sound and the motion when a sound occurs. In interaction phase, the system generates a motion when a sound is input.

We use the robot "Keepon" for implementation and experiments in this study. It is a creature-like robot developed at NICT mainly for communicative experiments with infants. Its body is approximately 12 cm high.

3.2 Learning Process

If object velocity, or the norm of the optical flow vector, is higher than the threshold when a sound is input, the system interprets the motion of the object as the reason why the sound occurred. The system memorizes the pair of the spectrogram of the separated sound and the sequence of the optical flow vectors into the mapping memory. This process is shown in Fig. 1. The detailed process in each module is explained in the following paragraphs.

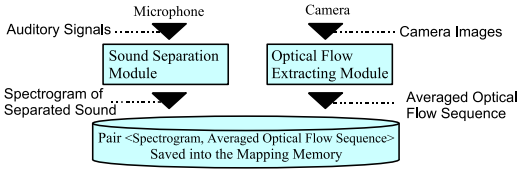

Figure 1: Motion and Sound Learning

Optical Flow Extraction Optical flows are constantly extracted from camera images. We adopted the block matching method for optical flow. Assuming that the camera captures only a sound source object, the system averages the flows of all of the blocks

on every frame. The averaged vector is regarded as the velocity vector of the sound source object on the frame.

Sound Separation In most physical phenomena, one sound is equivalent to one peak in the power envelope of an input audio signal. Therefore, each event is separated by extracting each peak. We adopted the separation method developed by Ishihara et al. (Ishihara et al., 2004). Individual sound separation is done by extracting each local maximum of power envelope which is not regarded as a part of another local maximum.

Learning Correspondences between Sounds and Motions Not only the motion during a sound, but also the motions right before and right after the sound are also parts of the motion which generates the sound. Therefore they are recorded into the mapping memory. The time span $\boldsymbol{T}_i$ of the i-th sound stored in the memory is defined as follows:

$$\boldsymbol{T}_i = [\ \min_t \{\, t \mid V_s < |\boldsymbol{F}(t)|,\ t \leq \min \boldsymbol{T}_s \},\ \max_t \{\, t \mid V_s < |\boldsymbol{F}(t)|,\ \max \boldsymbol{T}_s \leq t \}\ ]$$

where $\boldsymbol{F}(t)$ is the averaged optical flow vector at time t, V_s is the threshold by which the system decides whether the object is moving or not, and $\boldsymbol{T}_s$ is the time span that corresponds to the time span of the sound, which is given by the sound separation module.

3.3 Sound-to-Motion Translation

Fig. 2 illustrates the process of a translation from sounds to robot motions. The auditory distance from input sound to each sound in the mapping memory is computed. The motion learned as a pair with the sound which gives the smallest distance is picked up. Then the robot imitates the motion. If the smallest distance is larger than the threshold, the system does nothing. Motion generation uses the following modules in addition to the above-mentioned modules.

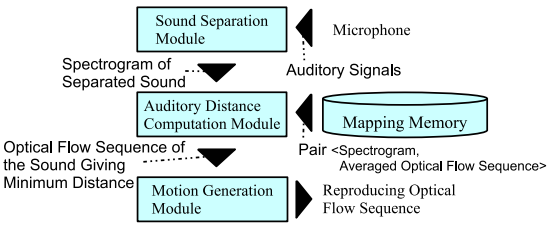

Figure 2: Sound-to-Motion Translation

Auditory Distance Computation Auditory distances are computed by dynamic time warping (DTW) with mel filter bank output. DTW has been reported as one of the most effective methods in recognition of environmental sounds (Cowling and Sitte, 2003).We adopt mel filter bank output, which is made to adapt to human perception, as a feature of each frame in DTW.

Motion Generation In imitating the motion of the object, the body position $\boldsymbol{X}(t')$ at the elapsed time t' from beginning of the gesture is defined as

$$\boldsymbol{X}(t') = C \sum_{s=1}^{t'} \boldsymbol{F}_i(s)$$

for the recorded optical flow sequence $\boldsymbol{F}_i(t)$ and a constant C. Because the motion $\boldsymbol{X}(t')$ is generated on a 2-dimensional plane, 2 degrees of freedom, which can reproduce the generated motion, must have been selected from all of the degrees of freedom of the robot beforehand.

4. Experiments

The motion generation system was implemented on Keepon. We made Keepon learn 4 kinds of sounds, such as the sound of rubbing uneven plastic pieces together and the sound of striking metal boards vertically. After the learning, we made Keepon generate the imitation motion by giving a sound without looking at the sound source. Based on the rubbing sound, Keepon made a right rotation, then a left rotation as shown in Fig. 3 (a). According to the striking sound, Keepon executed an anteflexion then a retroflexion as Fig. 3 (b). The coordinates of these motions are shown as Fig. 4. A video of these experiments can be viewed at `http://winnie.kuis.kyoto-u.ac.jp/members/yuya/demo.avi`.

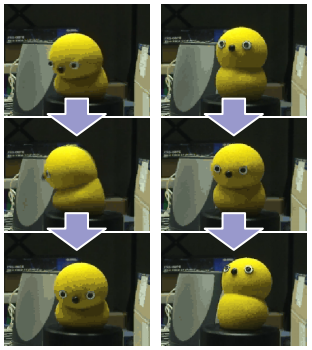

Figure 3: Generated Motions at the Experiment

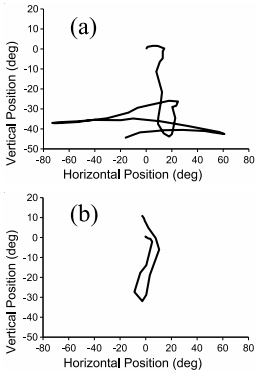

Figure 4: Coordinates of the Generated Motions

5. Conclusion

In this paper, we proposed an interaction model, in which robots make motions to communicate the sound. For this purpose, we developed an iconic gestures generation system on the real robot Keepon. The mappings from motions to sounds and motion generation for unknown sounds are the subjects of our future work. We are also interested in the relation with bouba-kiki effect (Ramachandran and Hubbard, 2001).

Acknowledgments This study was partially supported by the Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Science, Sports and Culture, and by the JPSP 21st Century COE Program.

References

Cowling, M. and Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recogn. Lett.*, 24(15):2895–2907.

Ishihara, K., Nakatani, T., Ogata, T., and Okuno, H. G. (2004). Automatic sound-imitation word recognition from environmental sounds focusing on ambiguity problem in determining phonemes. *PRICAI–2004*.

McNeill, D. (1992). *HAND AND MIND: What Gestures Reveal about Thought*. Univ. of Cicago Pr, Chicago.

Ramachandran, V. S. and Hubbard, E. M. (2001). Synaesthesia — a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34.

Werner, H. and Kaplan, B. (1963). *Symbol Formation: An Organismic-Developmental Approach to the Psychology of Language*. John Wiley and Sons, New York.