ACTIVE DETERMINATION OF CORRESPONDENCES BETWEEN AUDIO AND VISUAL EVENTS THROUGH ACTIVE MOTION

Kento Nishibori, Yoshinori Takeuchi, Tetsuya Matsumoto, Hiroaki Kudo, Noboru Ohnishi

Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
E-mail: {kent, takeuchi, matumoto, kudo, ohnishi}@ohnishi.m.is.nagoya-u.ac.jp

ABSTRACT

A human understands the objects in an environment by integrating information obtained by the senses of sight, hearing, and touch. In this integration, active manipulation of objects plays an important role. We propose a method for determining the correspondence of audio-visual events by manipulating an object.

1. INTRODUCTION

Infants 18 to 20 weeks old recognize the correspondence between speech sound (audio information) and mouth movement (visual information), and learn pronunciation and language by this ability[1]. In addition, co-occurrence of rhythmic action and vocal behavior contribute to an infant's acquisition of spoken language[2]. Similarly, it is necessary for an artificial intelligent system to correspond and integrate multi-modal sensor information in order to acquire knowledge about objects without any supervisor. In such integration, active movement plays an important role. We propose a method for determining the correspondence of audio-visual events by handling an object. The method uses the general grouping rules in Gestalt psychology, i.e."simultaneity" and "similarity" among motion commands, sound onsets, and motion of objects in images.

2. SENSORY-MOTOR CORRESPONDENCE

This system comprises four components, motor, audio, visual, and integration (Fig. 1). In the motor part (Fig. 1(1)), a computer sends motor commands to the manipulator to handle an unknown object. Then the object emits a sound and causes movement change in images. In the sound part (Fig. 1(2)), sound onset time series are detected in the frequency domain, and similar spectra at each sound onset are grouped into the same sound source of the object. In the visual part (Fig. 1(3)), a moving-object region and its movement are extracted. The motion loci of object regions are calculated and the changes in object movement are recorded. In the integration part (Fig. 1(4)), sampling rates of manipulator commands, audio signals, and camera images are converted to the same rate by re-sampling. Finally, we calculate correlations among audio, visual, and motor signals, and events with high correlation are grouped together.

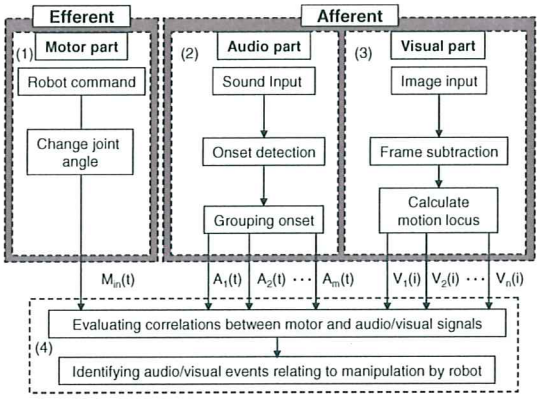

Fig. 1. Integration of afferent and efferent signals.

3. EXPERIMENTS AND DISCUSSION

In experiments, we used a microphone, a camera, and a robot featuring a hand manipulator. The robot grasps an object like a bell and shakes it, or grasps an object like a stick and beats a drum with it. Those motions are periodic or non-periodic. Then the object emits periodical/non-periodical events. To create a more realistic scenario, we put an another event source (a metronome) in the environment. We conducted two trials for each of 40 different conditions (80 trials in total).

Figure 2 shows a motion signal to the manipulator. It's non-periodical time intervals of motion-direction change are derived from the normal distribution with a mean of 0.963 s and a standard deviation of 0.126 s. The robot beats a drum with a stick and we add another event source (a metronome). Sound onset groups (Fig. 3(b) and 3(c)) were extracted from

This research was supported in part by a Grant-in-Aid for 21st Century COE Program "Intelligent Media Integration" of the Ministry of Education, Culture, Sports, Science and Technology Japan.

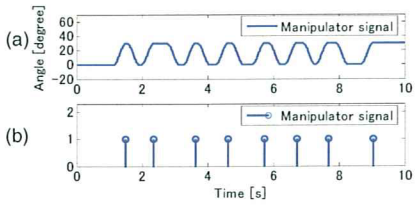

Fig. 2. Motion signal to manipulator.

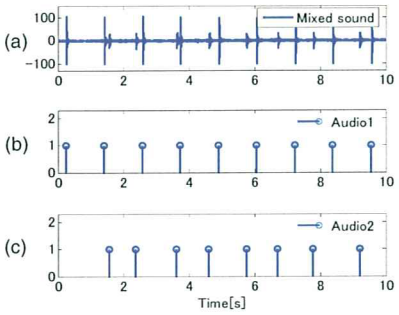

Fig. 3. Result of audio processing.
(a) Mixed sound, (b) Audio group 1, (c) Audio group 2.

the mixed sound shown in Fig. 3(a). In the visual part, three motion areas were extracted from these images (Fig. 4). Figure 5 shows the result, i.e. (b) the onset time series obtained from (a) mixed signals. Group 1 shown in Fig. 5(c) is a power spectrum (upper figure) at a sound onset relating to a motion image (lower figure), which was produced by manipulation. Group 2, shown in Fig. 5(d), is a power spectrum (upper figure) at a sound onset relating to a motion image (lower figure).

Table1 presents the results of correspondence between audio and visual events relating to object manipulation. It shows that the experimental conditions (objects, periodicity, and another event) did not affect the correspondence between audio-visual events. As a result, we have a success rate of 81.3 percent in determining the correspondence between audio-visual events (afferent signal), which were relating to robot motion (efferent signal).

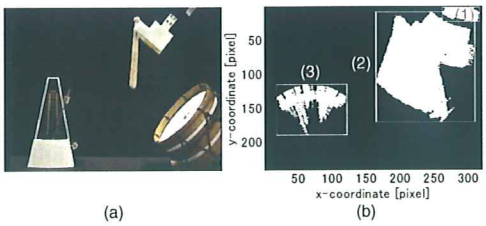

Fig. 4. Extracted motion areas.
(a) One of the input images, (b) Motion area of objects.

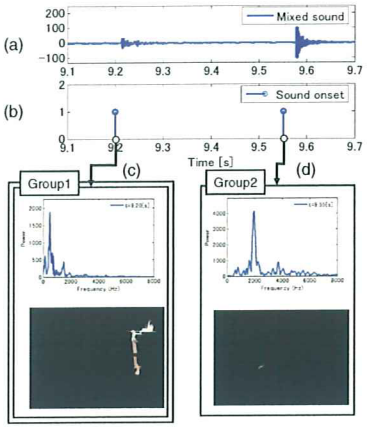

Fig. 5. Correspondence of two movements and two sounds (a drum and a metronome).

Table 1. Correspondence between audio and visual events relating to objects, periodicity, and another event.

Objects	Periodicity	Correlation		Correspondence	
		m:a	m:v	a:v	result
Drum	p.	0.932	0.694	0.737	8/10
	np.	0.904	0.649	0.711	9/10
Drum + Met.	p.	0.867	0.708	0.710	7/10
	np.	0.868	0.696	0.726	8/10
Bell	p.	0.757	0.719	0.715	10/10
	np.	0.816	0.587	0.638	10/10
Bell + Met.	p.	0.769	0.717	0.666	5/10
	np.	0.779	0.726	0.706	8/10
Average		0.837	0.687	0.701	65/80

4. CONCLUSION AND FUTURE WORK

We proposed a method for determining the correspondence of audio-visual events through active manipulation. This method is applicable to a practical situation in which conventional methods do not work well. One future work is to realize robustness against object occlusion.

5. REFERENCES

[1] P. Kuhl and A. Meltozoff, "The Bimodal perception of Speech in Infancy," Science, Vol. 218, pp. 1138-1140 (1982).

[2] K. Ejiri and N. Masataka, "Co-occurrence of preverbal vocal behavior and motor action in early infancy," Developmental Science 4, pp. 40-48 (2001).