Developmental Science

Developmental Science 12:3 (2009), pp 446–452 DOI: 10.1111/j.1467-7687.2008.00780.x

PAPER

Plasticity of ability to form cross-modal representations in infant Japanese macaques

Ikuma Adachi,[1] Hiroko Kuwahata,[1] Kazuo Fujita,[1] Masaki Tomonaga[2] and Tetsuro Matsuzawa[2]

1. *Department of Psychology, Graduate School of Letters, Kyoto University, Japan*
2. *Section of Language and Intelligence, Primate Research Institute, Kyoto University, Japan*

Abstract

*In a previous study, Adachi, Kuwahata, Fujita, Tomonaga & Matsuzawa demonstrated that infant Japanese macaques (*Macaca fuscata*) form cross-modal representations of conspecifics but not of humans. However, because the subjects in the experiment were raised in a large social group and had considerably less exposure to humans than to conspecifics, it was an open question whether their lack of cross-modal representation of humans simply reflected their lower levels of exposure to humans or was caused by some innate restrictions on the ability. To answer the question, we used the same procedure but tested infant Japanese macaques with more extensive experience of humans in daily life. Briefly, we presented monkeys with a photograph of either a monkey or a human face on an LCD monitor after playing a vocalization of one of these two species. The subjects looked at the monitor longer when a voice and a face were mismatched than when they were matched, irrespective of whether the preceding vocalization was a monkey's or a human's. This suggests that once monkeys have extensive experience with humans, they will form a cross-modal representation of humans as well as of conspecifics.*

Introduction

Most objects and events that we experience in the world provide both amodal and arbitrary information to multiple sensory modalities. For instance, when a person speaks, the synchrony of voice and mouth provides amodal information, whereas the pairing of voice and face is arbitrary. Amodal information is usually provided by synchrony of movements incorporating temporal characteristics such as duration and rhythm. These characteristics indicate that information in multiple modalities comes from the same source and helps us to learn arbitrary connections among information in multi-sensory modalities.

One important question is when and how humans develop the ability to process amodal and arbitrary relations among information in multi-sensory modalities. In the past few decades, a growing number of studies have attested to the early development of intermodal perception in human infants (for reviews, see Bushnell & Boudreau, 1991, 1993; Lewkowicz, 2000, 2001; Lickliter & Bahrick, 2001). Currently, most studies support a theoretical view that basic intersensory perceptual abilities are present at birth and become increasingly differentiated and refined with age (Gibson, 1969; Lewkowicz, 2000).

Three-week-old infants have been shown to equate auditory and visual stimuli spontaneously on the basis of intensity (Lewkowicz & Turkewitz, 1980). In addition, infants can detect temporal connections between auditory and visual stimulation, such as temporal synchrony between sights and sounds (Lewkowicz, 1996a) at 4 months. At around the same age, infants can also detect the amodal relation between the common rhythm and duration unifying flashing lights with tones (Allen, Walker, Symonds & Marcell 1977; Lewkowicz, 1986), the tempo of action unifying the sights and sounds of stuffed animals bouncing (Spelke, 1979), and the composition of moving objects (Bahrick, 1987, 1988).

Young infants are also skilled perceivers of amodal information in audiovisual speech and are able to match faces and voices on this basis (see Lewkowicz, 1996b; Walker-Andrews, 1997). By the age of 2 months, infants can detect voice–lip synchrony (Dodd, 1979), and by 4 months infants match audible and visible speech on the basis of spectral information in vowel sounds (Kuhl & Meltzoff, 1984; Patterson & Werker, 1999). Between 5 and 7 months, infants are also able to match faces and voices on the basis of affect (Soken & Pick, 1992; Walker-Andrews, 1997), gender (Walker-Andrews, Bahrick, Raglioni & Diaz, 1991) and age of speaker (Bahrick, Netto & Hernandez-Reif, 1998). These findings demonstrate that infants are excellent perceivers of a wide range of amodal relations uniting faces and voices across a variety of events.

Address for correspondence: Ikuma Adachi, Primate Research Institute, Kyoto University, Kanrin, Inuyama-city, Aichi 484-8506, Japan; e-mail: adachi@pri.kyoto-u.ac.jp

Research indicates that infants detect arbitrary relations in multimodal stimulation but that detection of these relations emerges later than detection of amodal relations in the same events (e.g. Bahrick, 1994, 2001). At 7 months, but not at 3 or 5 months, human infants detected the arbitrary relation between the colour–shape of an object and the pitch of its sound of impact when the sounds and sights of impacts were synchronous (Bahrick, 1994). Similarly, at around the age of 7 months, infants no longer required synchrony for the detection and memory of the collocated object–sound pairings (Morrongiello, Lasenby & Lee, 2003). Amodal relations provide a basis for guiding attention and constraining learning to unitary audiovisual events and subsequently for appropriate generalization of knowledge across a variety of audiovisual events.

Another important question to be answered is how these abilities have evolved. These two aspects have received attention in comparative cognitive science, and species from various taxa have been tested. For example, Parr (2004) demonstrated that chimpanzees (*Pan troglodytes*) are able to match conspecific vocalizations to photographs of corresponding facial expressions in a matching-to-sample procedure. Ghazanfar & Logothetis (2003) reported that rhesus monkeys (*Macaca mulatta*) are able to detect amodal relationships between conspecific facial expressions and the vocalizations. They used a preferential looking technique in which the subjects were shown two side-by-side video clips, synchronized to an audio track of the same conspecific individual ('stimulus animal') articulating two different calls. A sound that corresponded to one of the two facial postures was played through a speaker. The subjects looked longer at the videos showing the facial expression that matched the simultaneously presented vocalization than at those that did not. More recently, Evans, Howell & Westergaard (2005) also used a preferential looking procedure with tufted capuchin monkeys (*Cebus paella*) and found that they also are able to detect the correspondence between appropriate visual and auditory events. Thus this ability to detect amodal relationships between visual and auditory information seems to be shared widely in primates from New World species to humans.

The ability to detect arbitrary relations has also been demonstrated in non-human species. A female chimpanzee was successfully trained to match noises of objects such as castanets or voices of familiar conspecifics or trainers to their respective photographs (Hashiya & Kojima, 2001; Kojima, Izumi & Ceugniet, 2003). These results suggest that chimpanzees can learn associations across sensory modalities. However, these findings are so far limited to one chimpanzee. It remains to be seen how widely this cognitive ability is shared in the animal kingdom.

Here, we raise another important aspect related to the ability to detect arbitrary relations; that is, the interchange of information from one modality to another. We humans form a concept incorporating multi-sensory information after learning arbitrary relations through experience. For instance, our concept of dogs includes their visual appearances, their smells, their vocalizations, etc. Furthermore, each exemplar of a concept naturally may activate other exemplars of the same concept even across sensory modalities: we may activate visual images of dogs when we hear their barking without seeing the animals. Such an interchange of information across sensory modalities is useful because a modality available at one time may be unavailable at another.

However, this aspect, namely cross-modal representation, has not received much attention to date. The studies described above imply that the species in question generate visual images when they hear sounds or vocalizations. However, the tasks used, involving the simultaneous presentation of two visual stimuli to the subject, allow the subject to choose one stimulus by judging which one is more strongly associated with the auditory stimulus. Thus, in the strict sense, it is still unclear whether the subjects actually activate visual images on hearing the vocalization, before the visual stimuli appear. This aspect of intermodal transformation of representations remains to be tested directly. At the same time, further information is needed on how widespread such intermodal transformation of representations might be in the animal kingdom.

Martin-Malivel & Fagot (2001) tested this cross-modal representation in Guinea baboons (*Papio papio*), an Old World monkey species. These authors trained the subjects to discriminate between human and baboon vocalizations. They then introduced in probe trials either human or baboon photos as a brief prime before a stimulus vocalization. The results showed that the presence of the photos of matching species shortened the response time in one of two baboons. This cross-modal priming effect of pictures on auditory discrimination suggested that the subject had formed arbitrary associations of the voices of these species and their appearances. It could be that one of the subjects transformed information from a priming picture to the corresponding auditory information, which in turn helped the subject to solve the task when the prime picture was conceptually associated with the target sound.

Using a similar procedure, we demonstrated that squirrel monkeys (*Saimiri sciureus*), a New World primate species, formed cross-modal representations of their caretakers and transformed auditory information into visual images (Adachi & Fujita, 2007). Briefly, we first trained the subjects to discriminate between photographs of two of their caretakers. After reaching criterion, a voice, either matching (congruent condition) or mismatching (incongruent condition) with the sample photograph, was played back during the delay in probe test trials. The matching accuracies of the subjects in the incongruent condition were significantly lower than those in the congruent condition. This suggests that the preceding vocalization made the subjects activate the corresponding visual image, which interfered with their memory traces for the visual sample stimuli.

We also tested cross-modal representations in non-human animals using an expectancy violation procedure. We presented subjects with a voice followed by a photograph of a face, either matching or mismatching the preceding voice. Our hypothesis was that, if the subjects recall an appropriate representation upon hearing the voice, they should be surprised when a mismatching face follows. This would lead to a longer looking time towards the mismatching photograph than towards the matching one. With this procedure, we demonstrated that dogs (*Canis familiaris*) form cross-modal representations of their owners (Adachi, Kuwahata & Fujita, 2007).

However, the ontogeny of such cross-modal representations is not well known. In a previous study, using an expectancy violation procedure, we successfully demonstrated that infant Japanese macaques formed a cross-modal representation of conspecifics in their first year of life but not of humans (Adachi, Kuwahata, Fujita, Tomonaga & Matsuzawa, 2006). Our subjects in that study were raised in a large compound with other group mates and had substantial experience of conspecifics but not of humans. This limited exposure to humans may have been the cause of the asymmetrical result. Another possibility is that an innate mechanism for species recognition might have restricted the formation of such representations to within their own species. Thus, it remains to be seen whether the formation of cross-modal representations is so general that it can be extended to agents other than those that have special importance for animals, such as conspecisfics.

The main purpose of this study was to answer this question. We thus tested infant Japanese monkeys who were kept in indoor cages and had abundant experience of seeing and hearing humans, using the same expectancy violation procedure that we used in our previous experiment.

Methods

Subjects

We tested five infant Japanese macaques, all kept at the Primate Research Institute, Kyoto University, Inuyama, Japan. They were raised by their biological mothers and lived in indoor monkey cages. In the indoor-cage rooms, there were at least 10 cages (each cage had an adult monkey, a mother and her child, or two infants in it), and consequently the subjects had extensive experience of seeing and hearing conspecifics. Moreover, they had abundant experience of seeing and hearing humans, who brought them their daily feed, cleaned the cages, and so on. The infants were tested longitudinally from 27 days up to 103 days after birth ($M = 64.5$, $SD = 26.9$). We completed 12 sessions in total. The experiments complied with the Guideline for the Care and Use of Laboratory Animals, Primate Research Institute, Kyoto University.

Stimuli and apparatus

We used the same stimuli as we used in the previous experiment: a photograph of an unfamiliar adult Japanese macaque against an ivory-colored background (PM); a photograph of an unfamiliar adult human (PH) against the same background; a vocalization by a Japanese macaque (VM); and a vocalization by a human (VH). The monkey vocalization was a ‘coo-call’, typically used to solicit social contact with other individuals. The human vocalization was ‘oi’, which Japanese people typically use to attract another’s attention. The two vocalizations were of approximately equal duration.

The stimuli were prepared as follows: all vocalizations were stored on a computer in WAV format, with a sampling rate of 44,100 Hz and a sampling resolution of 16 bit. The amplitude of the vocalizations sounded equivalent to human ears. We took a digital full-face photograph of each stimulus individual and stored the photograph on the computer in JPEG format, sized 450 (W) × 550 (H) pixels, or ca. 16 × 20 cm on the 18.1-inch LCD monitor (SONY SDM-M81) used for presentation. We set up the apparatus as shown in Figure 1 in an experimental room at the Primate Research Institute. Briefly, the LCD monitor was located about 50 cm from the subject’s face. At the start of the test, a black opaque screen (50 × 70 cm) was placed in front of the monitor to prevent the subject from seeing the monitor. A digital camcorder (SONY DCR-TRV-30), located behind the

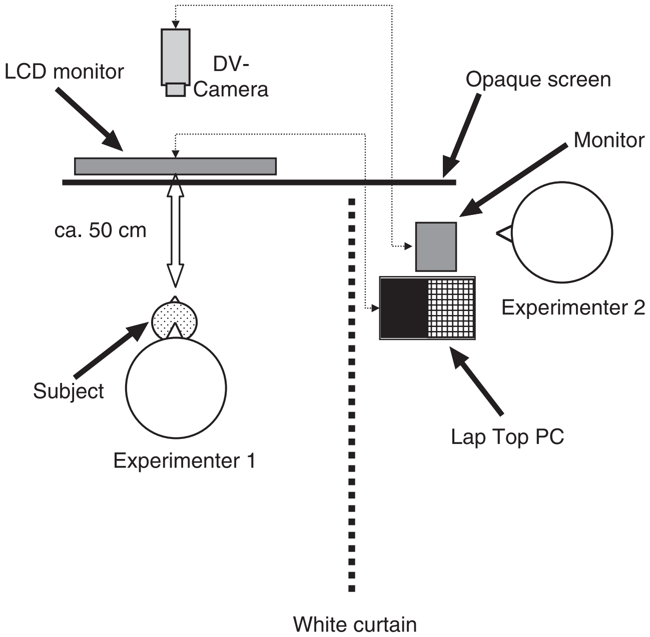

Figure 1 *A schematic drawing of the apparatus. The experimental area was separated into two parts by a white curtain. Experimenter 1 held the subject, while Experimenter 2 operated the personal computer to present stimuli, and removed the screen so that the subject could see the visual stimuli on the monitor.*

monitor, recorded the subject's behaviour. Presentation of the stimuli was controlled by a Visual Basic 5.0 program on a laptop personal computer (Dell Inspiron 4100, with Pentium III 1.2 GHz).

Procedure

We used the same procedure as in the previous experiment, namely an expectancy violation procedure. Our previous experiments on dogs (Adachi, Kuwahata & Fujita, 2007) and Japanese macaques (Adachi *et al.*, 2006) have shown that this procedure is useful for investigating cross-modal representations.

Each trial consisted of the following events: one experimenter (Experimenter 1) held the monkey in a towel on his or her lap in front of the LCD monitor, and remained silent, stationary and passive throughout the trial. A second experimenter (Experimenter 2), who observed the subject via a 2.5-inch television monitor connected to the camcorder behind the LCD monitor, started the trial when the subject appeared calm and alert, and oriented towards the LCD monitor. Each trial consisted of two phases. The first was the voice phase and the second was the photograph phase. In the voice phase, one of the two vocalizations was played through the speakers installed in the monitor, every two seconds for a total of three vocalizations. The duration of each vocalization was about 600 msec. The photograph phase began immediately after the final vocalization. Experimenter 2 smoothly removed the opaque screen to reveal a face on the LCD monitor. This experimenter was always positioned behind the curtain (Figure 1) and was ignorant of the precise face shown on the monitor. The photograph phase lasted for 15 sec. The behaviour of the subject in this phase was video-recorded for later analysis. We used the opaque screen to facilitate the subject's understanding that something was hidden behind the screen. Each subject was given the following four types of test trials: a VM-PM trial, in which the monkey photograph appeared after the monkey vocalization; a VH-PH trial, in which the human photograph appeared after the human vocalization; a VH-PM trial, in which the monkey photograph followed the human vocalization; and a VM-PH trial, in which the human photograph followed the monkey vocalization. The vocalization and the photograph matched in the first two trials but mismatched in the last two.

These four trials were presented in semi-random order, with the restriction that the same vocalization was not repeated on consecutive trials. The inter-trial interval was about 5 min, during which subjects were run on other experiments involving non-social stimuli. We hypothesized that, if the subject generated a visual image of the appropriate species upon hearing a vocalization, it would be surprised at the mismatch in the last two types of test trials (VH-PM and VM-PH trials), and thus would look at the photograph for longer than in the other two types of test trials.

Results

After the experiments, the videos of trials were captured on a personal computer and converted to MPEG file format (30 frames per second). A coder who was blind to the stimuli recorded the duration of subjects' looks at the monitor in the photograph phase. A second coder scored data for four randomly sampled sessions to check the reliability of coding. The correlation between total looking time for each trial measured by the two coders was highly significant (Pearson's $r = 0.995$, $n = 16$, $p < .01$).

We calculated the total looking time in each trial for each subject. Figure 2 shows the mean duration of looking at the monitor in the photograph phase for each condition averaged for all subjects (see also Table 1). It may be noted that our subjects looked at the stimuli for only around 30% of the photograph phase on average; however, it is generally the case that the looking duration of non-human primates towards the objects on the monitor is much shorter than that of humans, and this proved to be the case for the current study too.

Looking times were analysed by means of a 2 × 2 repeated-measures analysis of variance with photographs (Monkey or Human) and conditions (Congruent or Incongruent) as factors. There was a significant main effect of condition ($F(1, 11) = 8.816$, $p = .013$), but no significant main effect of photograph ($F(1, 11) = 0.055$, $p = .850$).

More importantly, in this study, there was no significant interaction between the two factors ($F(1, 11) = 0.437$, $p = .522$). These results suggest that our subjects looked longer at the monitor in Incongruent condition

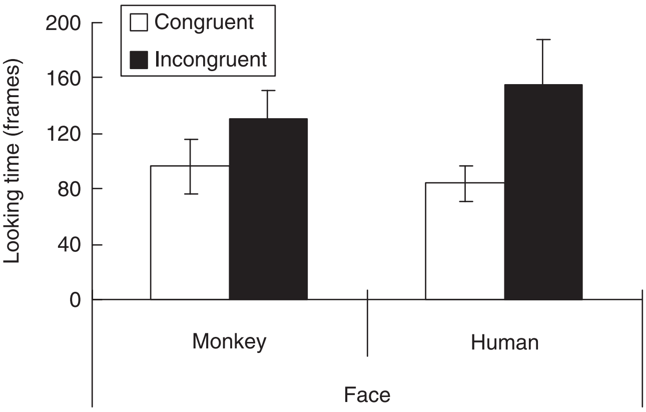

Figure 2 *Duration of looking at the monitor in the photograph phase for each condition averaged for all subjects.*

Table 1 *Mean looking time and* SE *for each condition (in frames)*

	Condition			
	Congruent		Incongruent	
Trial type	VM-PM	VH-PH	VM-PH	VH-PM
Mean	96.2	83.9	154.5	130.9
SE	19.9	12.9	33.4	20.2

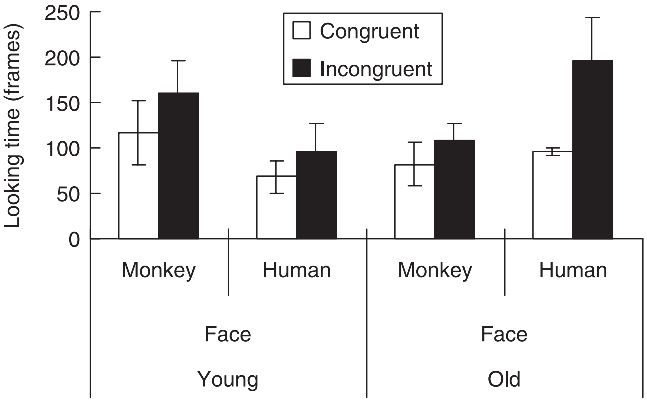

Figure 3 *Duration of looking at the monitor in the photograph phase for each condition for both younger and older age groups.*

than in Congruent condition, irrespective of whether the preceding vocalization was a monkey's or a human's. Thus the subjects in the present study had formed cross-modal representation of both conspecifics and humans.

In order to determine whether there was any developmental change across the ages of the subjects, we conducted another analysis. We divided the results into two groups in light of the age of the subjects when the test was conducted, namely the first two months (number of trials = 5, range: 27–55 days, M = 39.2 days, SD = 13.2) and the second two months (number of trials = 7, range: 62–103 days, M = 82.6 days, SD = 17.3) (Fig. 3). The looking times were analysed by means of a 2 × 2 × 2 analysis of variance with photographs (Monkey or Human), conditions (Congruent or Incongruent) and age groups (the first two or the second two months). The analysis again showed a significant main effect of condition (F(1, 10) = 7.531, p = .021). However, there was no significant main effect of photograph (F(1, 10) = 0.258, p = .623), no significant interaction between the two factors (F(1, 10) = 0.258, p = .623), and no main effect of age (F(1, 10) = 0.189, p = .673). Moreover, any interactions with age groups were not significant. These results mean that the tendency of looking behaviour for each stimulus was not different across the two testing periods. However, we had a small number of subjects for each test period, and so more studies are needed to confirm the result.

Discussion

The present study shows that infant Japanese macaques form cross-modal representations that incorporate visual and auditory senses not only of conspecifics but also of humans. Thus the lack of cross-modal representation of humans in the previous study on monkeys (Adachi *et al.*, 2006) could have been caused mainly by their lower levels of exposure to humans. Our findings suggest that infant Japanese macaques can form cross-modal representations of species other than conspecifics when they are exposed to that species extensively from birth.

One possible confounding factor is that the experimenter, who held the subjects, might have unwittingly provided the subjects with cues. However, this is unlikely because the experimenter was asked to be silent and passive throughout, and was ignorant of the stimuli presented because he or she was asked to look at the small monitor on the camcoder to keep the subjects on film.

Although we found that the infant monkeys formed cross-modal representations not only of conspecifics but also of humans and we did not find any differences across age of the subjects, further studies are called for to investigate their developmental changes in detail. In a previous study (Adachi, Fujita, Kuwahata & Ishikawa, 2003), we found that recognition of biological motion was affected by an innate mechanism for species recognition. Although the monkeys in Adachi *et al.* (2003) recognized biological motion of the species with which they had the most extensive experience, there was a difference in how this recognition developed. Cage-reared monkeys came to prefer human biological motion from the age of 8 to 15 weeks, whereas the enclosure-reared group preferred macaque biological motion at all ages tested from 0 to 25 weeks. This difference suggests that some innate factors in the development of biological motion perception might interact with experience. More recently, Dufor, Pascalis & Petit (2006) showed that face processing was limited to own-species. In their study, they found that both Tonkean macaques (*Macaca tonkeana*) and brown capuchin monkeys (*Cebus apella*) showed advantage for own-species face recognition and showed difficulty for other-species face recognition. Similar innate factors for species recognition may influence monkeys' formation of cross-modal representations of species. In the present study, however, we had too small a number of subjects for each test period to understand developmental changes in detail. Further studies are needed to explore this issue. In particular, it is important to test infants younger than two months old.

More recently, Martin-Malivel & Okada (2007) found that chimpanzees who were exposed to humans more than to chimpanzees from their birth were better at discriminating human faces than at discriminating chimpanzee faces. By contrast, Matsuzawa (1991) reported that a chimpanzee found it more difficult to name photos of humans than photos of chimpanzees even though the subject was exposed to humans for a long time after coming to the institute at one year old. This difference may imply that there is a critical period for chimpanzees during the first year in which they tune their perceptual systems for processing certain objects, a process known as 'perceptual narrowing' in human infants (e.g. Pascalis *et al.*, 2005). In the latest study, Sugita (2008) showed that such 'perceptual narrowing' actually exists in Japanese monkeys. He isolated infant Japanese monkeys immediately after birth and raised them in a face-free environment for 6–24 months. After this deprivation period, monkeys were exposed to either human or monkey faces for a month. Immediately following deprivation, the monkeys could discriminate faces of both species equally well. However, they lost their ability to discriminate non-exposed-species faces after exposure to faces of the other species. Interestingly, the monkeys could not retrieve the loss in discriminatory skills for the non-exposed species even though they were exposed to these species for a year after the experiments. This finding strongly suggests that (1) there is a critical period for developing face perceptions in Japanese macaques, (2) the developmental door is shut after this period, and (3) no re-tuning can occur. The question whether there is a critical period for Japanese macaques to form cross-modal representations of species should also be addressed in future studies.

Another aspect of cross-modal representations also needs to be examined, namely whether or not there is any preferred direction in which to transform cross-modal information. Previous studies showed cross-modal transformation of information in both visual to auditory (Martin-Malivel *et al.*, 2001) and auditory to visual (Adachi & Fujita, 2007) directions. Although only one of two subjects showed evidence of this transformation in the former direction, both monkeys did in the latter. This might imply that monkeys' preferred direction of this transformation is from auditory to visual. If an animal uses vision as the primary channel for controlling its behaviour, the transformation of information from other modalities to vision could be advantageous.

Finally, a weakness of our experiment is that we used only one photo and one vocalization for each of the two stimulus species. Further tests with more exemplars are needed before we can safely conclude that monkeys have cross-modal representation of both species, although such tests may be difficult in practice. However, the ability we have shown could not be individual-specific because both photos and vocalizations were of unfamiliar individuals. Nor can the ability be a consequence of association learning because there was no pairing between stimuli that would result in the formation of such association learning.

Acknowledgements

This study was supported by Research Fellowships from the Japan Society of the Promotion of Science (JSPS) for Young Scientists to Ikuma Adachi and Hiroko Kuwahata, by Grants-in-Aid for Scientific Research Nos 13410026 and 14651020 from JSPS, Japan, to Kazuo Fujita, and by the 21st Century COE Program, D-10, from the Ministry of Education, Culture, Sports, Science, and Technology to Kyoto University. It was also supported by the Cooperative Research Program of the Primate Research Institute, Kyoto University. We thank James R. Anderson for his editing of the manuscript. We would like to express our appreciation to all members of the Center for Human Evolution Modeling Research, Primate Research Institute, Kyoto University for their assistance during this study and for the management of the subjects' health. The experiments complied with the Guide for the Care and Use of Laboratory Primates, Primate Research Institute, Kyoto University.

References

Adachi, I., Fujita, K., Kuwahata, H., & Ishikawa, S. (2003). Perception of biological motion in infant macaques. In M. Tomonaga, M. Tanaka, & T. Matsuzawa (Eds), *Cognitive and behavioral development in chimpanzees AQ comparative approach*. Kyoto University Press, Kyoto, pp. 333–336 (in Japanese).

Adachi, I., Kuwahata, H., Fujita, K., Tomonaga, M., & Matsuzawa, T. (2006). Japanese macaques form a cross-modal representation of their own species in their first year of life. *Primates*, **47**, 350–354.

Adachi, I., Kuwahata, H., & Fujita, K. (2007). Dogs recall their owner's face upon hearing the owner's voice. *Animal Cognition*, **10**, 17–21.

Adachi, I., & Fujita, K. (2007). Cross-modal representation of human caretakers in squirrel monkeys. *Behavioral Processes*, **74**, 27–32.

Allen, T.W., Walker, K., Symonds, S.L., & Marcell, M. (1977). Intersensory and intrasensory perception of temporal sequences. *Developmental Psychology*, **13**, 225–229.

Bahrick, L.E. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behavior and Development*, **2**, 13–17.

Bahrick, L.E. (1988). Intermodal learning on the basis of two kinds of invariant relations in audible and visual events. *Child Development*, **59**, 197–209.

Bahrick, L.E. (1994). The development of infants' sensitivity to arbitrary intermodal relations. *Ecological Psychology*, **6**, 111–123.

Bahrick, L.E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology*, **79**, 253–270.

Bahrick, L.E., Netto, D., & Hernandez-Reif, M. (1998). Intermodal perception of adult and child faces and voices by infants. *Child Development*, **69**, 1263–1275.

Bushnell, E.W., & Boudreau, J.P. (1991). The development of haptic perception during infancy. In M.A. Heller, W. Schiff (Eds), *The psychology of touch*. Hillsdale, NJ: Erlbaum.

Bushnell, E.W., & Boudreau, J.P. (1993). Motor development and the mind: the potential role of motor abilities as a determinant of aspects of perceptual development. *Child Development*, **64**, 1005–1021.

Dodd, B. (1979). Lip reading in infants: attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, **11**, 478–484.

Dufour, V., Pascalis, O, & Petit, O. (2006). Face processing limitation to own species in primates: a comparative study in brown capuchins, Tonkean macaques and humans. *Behavioral Processes*, **73**, 107–113.

Evans, T.A., Howell, S., & Westergaard, G.C. (2005). Visual cross-modal perception of communicative stimuli in tufted Capuchin monkeys (Cebusapella). *Journal of Experimental Psychology: Animal Behavior Processes*, **31**, 399–406.

Ghazanfar, A.A., & Logothetis, N.K. (2003). Neuroperception: facial expressions linked to monkey calls. *Nature*, **423**, 937–938.

Gibson, E.J. (1969). *Principles of perceptual learning and development*. New York: Appleton.

Hashiya, K., & Kojima, S. (2001). Acquisition of auditory–visual intermodal matching to sample by a chimpanzee (*Pan troglodytes*): comparison with visual–visual intramodal matching. *Animal Cognition*, **4**, 231–239.

Kojima, S., Izumi, A., & Ceugniet, M. (2003). Identification of vocalizers by pant hoots, pant grunts and screams in a chimpanzee. *Primates*, **44**, 225–230.

Kuhl, P.K., & Meltzoff, A.N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, **7**, 361–381.

Lewkowicz, D.J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior and Development*, **9**, 335–353.

Lewkowicz, D.J. (1996a). Perception of auditory–visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance*, **22**, 1094–1106.

Lewkowicz, D.J. (1996b). Infants' response to the audible and visible properties of the human face: I. role of lexical/syntactic content, temporal synchrony, gender, and manner of speech. *Developmental Psychology*, **32**, 347–366.

Lewkowicz, D.J. (2000). The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychological Bulletin*, **126**, 281–308.

Lewkowicz, D.J. (2001). The concept of ecological validity: what are its limitations and is it bad to be invalid. *Infancy*, **2**, 437–450.

Lewkowicz, D.J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory–visual intensity matching. *Developmental Psychology*, **16**, 597–607.

Lickliter, R., & Bahrick, L.E. (2001). The salience of multimodal sensory stimulation in early development: implications for the issue of ecological validity. *Infancy*, **2**, 451–467.

Martin-Malivel, J., & Fagot, J. (2001). Cross-modal integration and conceptual categorization in baboons. *Behavioural Brain Research*, **122**, 209–213.

Martin-Malivel, J., & Okada, K. (2007). Human and chimpanzee face recognition in chimpanzees (Pan troglodytes): role of exposure and impact on categorical perception. *Behavioral Neuroscience*, **121**, 1145–1155.

Matsuzawa, T. (1991). *Chimpanzee Kara Mita Sekai*. Tokyo: Tokyo University Press (in Japanese).

Morrongiello, B.A., Lasenby, J., & Lee, N. (2003). Infants' learning, memory, and generalization of learning for bimodal events. *Journal of Experimental Child Psychology*, **84**, 1–19.

Parr, L.A. (2004). Perceptual biases for multimodal cues in chimpanzee (Pan troglodytes) affect recognition. *Animal Cognition*, **7**, 171–178.

Pascalis, O., Scott, L., Kelly, D.J., Shannon, R.W., Nicholson, E., Coleman, M., & Nelson, C.A. (2005). Plasticity of face processing in infancy. *Proceedings of the National Academy of Sciences*, **102**, 5297–5300.

Patterson, M., & Werker, J.F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, **22**, 237–247.

Soken, N.H., & Pick, A.D. (1992). Intermodal perception of happy and angry expressive behaviors by seven-month old infants. *Child Development*, **63**, 787–795.

Spelke, E.S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, **15**, 626–636.

Sugita, Y. (2008). Face perception in monkeys reared with no exposure to faces. *Proceedings of the National Academy of Sciences*, **105**, 394–398.

Walker-Andrews, A.S. (1997). Infants' perception of expressive behaviors: differentiation of multimodal information, *Psychological Bulletin*, **121**, 437–456.

Walker-Andrews, A.S., Bahrick, L.E., Raglioni, S.S., & Diaz, I. (1991). Infant's bimodal perception of gender. *Ecological Psychology*, **3**, 55–75.

Received: 16 July 2007
Accepted: 14 April 2008