Anim Cogn (2007) 10: 17–21
DOI 10.1007/s10071-006-0025-8

ORIGINAL ARTICLE

Ikuma Adachi · Hiroko Kuwahata · Kazuo Fujita

Dogs recall their owner's face upon hearing the owner's voice

Received: 2 February 2006 / Revised: 12 May 2006 / Accepted: 12 May 2006 / Published online: 21 June 2006

Abstract We tested whether dogs have a cross-modal representation of human individuals. We presented domestic dogs with a photo of either the owner's or a stranger's face on the LCD monitor after playing back a voice of one of those persons. A voice and a face matched in half of the trials (Congruent condition) and mismatched in the other half (Incongruent condition). If our subjects activate visual images of the voice, their expectation would be contradicted in Incongruent condition. It would result in the subjects' longer looking times in Incongruent condition than in Congruent condition. Our subject dogs looked longer at the visual stimulus in Incongruent condition than in Congruent condition. This suggests that dogs actively generate their internal representation of the owner's face when they hear the owner calling them. This is the first demonstration that nonhuman animals do not merely associate auditory and visual stimuli but also actively generate a visual image from auditory information. Furthermore, our subject also looked at the visual stimulus longer in Incongruent condition in which the owner's face followed an unfamiliar person's voice than in Congruent condition in which the owner's face followed the owner's voice. Generating a particular visual image in response to an unfamiliar voice should be difficult, and any expected images from the voice ought to be more obscure or less well defined than that of the owners. However, our subjects looked longer at the owner's face in Incongruent condition than in Congruent condition. This may indicate that dogs may have predicted that it should not be the owner when they heard the unfamiliar person's voice.

I. Adachi (✉) · H. Kuwahata · K. Fujita
Department of Psychology, Graduate School of Letters, Kyoto University, Yoshida-honmachi, Sakyo,
Kyoto 606-8501, Japan
e-mail: iadachi@emory.edu
Tel.: +81-75-753-2759
Fax: +81-75-753-2759

Present address:
I. Adachi
Yerkes National Primate Research Center,
954 Gatewood Road, Atlanta, GA 30329, USA

Keywords Cross-modal representation · Natural concepts · Dogs

Introduction

In their pioneering study, Herrnstein and Loveland (1964) showed that pigeons could be trained to peck at photos containing humans and not to peck at those without humans. Herrnstein and Loveland suggested that this behavior was evidence of pigeons' forming a "concept" of humans. Later experiments have shown that several avian and primate species form natural concepts of a variety of objects such as water, trees, oak leaves, monkeys, and so on (Herrnstein et al. 1976; Cerella 1979; Yoshikubo 1985). Thus, perceiving distinguishable stimuli as a group is widespread in the animal kingdom. After these experiments, studies with well-controlled stimuli have demonstrated how nonhuman animals discriminate classes of stimuli. Some argue that the discrimination might reflect a simple stimulus generalization from a few memorized exemplars (Cook et al. 1990); whereas others suggest that animals may learn a prototype of the stimuli (Aydin and Pearce 1994). Some propose that animals may learn to attend to a few physical dimensions of the stimuli (Lea and Ryan 1993), whereas others argue that categorization sometimes transcends perceptual resemblance, for example, when dissimilar stimuli give rise to the same behavior (Wasserman et al. 1994).

However, those studies are limited in several ways. First, because of the perceptual resemblance among stimuli in previously used categorical discrimination tasks, studies have not demonstrated nonhuman animals' abilities for categorization controlled by conceptual mechanisms that are independent of the perceptual characteristics of the stimuli or associations with motor responses. Second, each exemplar of natural concepts that we humans have, may lead us to generate a specific or typical representation of that concept. This aspect of interchanging information has not received much attention. Comparative cognitive approaches to such aspects are essential for understanding how abilities for categorization might have evolved.

In the present study, we focused on the cross-modal nature of concepts. Clearly, exemplars in different sensory modalities should not share any perceptual characteristics. For instance, our concept of "dogs" contains not only their various shapes but also their vocalizations, smells, etc. with no perceptual resemblance between their vocalizations and their appearance. Furthermore, when we hear the vocalization of a dog, we may activate visual images of dogs. This is the interchanging information aspect of concepts described above. Such interchanging information across sensory modalities would be useful to animals because the modality available at one time may be unavailable at other times. For example, if an animal uses vision as the primary channel for controlling its behavior, transforming information from other modalities to vision would be advantageous, as seems to be the case for humans.

Recent reports show that nonhuman primates can form associations across sensory modalities. Using an auditory–visual matching to sample procedure, Hashiya and Kojima (1999, 2001) demonstrated that a female chimpanzee successfully matched noises of objects or voices of familiar persons with corresponding photographs. In their experiment, the subject was given visual choice stimuli after a sound was played back. She was required to select the appropriate photograph of an object such as a castanet following its sound, or photographs of familiar trainers following their voices. After a training phase, she showed good performance to a new set of familiar objects. Such multi-modal association is not limited to apes; Ghazanfar and Logothetis (2003) reported that rhesus monkeys are also able to detect the correspondence between conspecific facial expressions and the voices. They used a preferential-looking technique. Subjects were seated in front of two LCD monitors and shown two side-by-side 2-s videos, synchronized to an audio track, played in a continuous loop for 1 min, 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 more at the videos showing the facial expression that matched the simultaneously presented vocalization than at those that did not. More recently, Evans et al. (2005) also used a preferential looking procedure with tufted capuchin monkeys and found that they also are able to detect the correspondence between appropriate visual and auditory events.

These results imply that those species generate visual images when they hear sounds or vocalizations. However, the tasks used, involving simultaneous presentation of two visual stimuli to the subject, allow the latter to choose one by judging which stimulus is more associated with the auditory stimulus after comparing them. Thus, in the strict sense, it is still unclear whether they 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 required on how widespread such intermodal transformation of representations might be in the animal kingdom.

To explore this issue, we used an expectancy violation procedure, often used to test human infants for inferences about external events. Typically, the subject is shown one event several times followed by a second event. It is assumed that the subject should look longer at the second event if the latter contradicted the outcome expected based on the first. In our procedure, we present a vocalization followed by only one photograph of a face, either matching (Congruent condition) or mismatching (Incongruent condition) in personal identity with the vocalization. If the subject activates a visual image and expects it on hearing the voice, the expectation should be contradicted if a mismatching face then appears. Thus, the subject should look longer at the photographs in the Incongruent condition than the Congruent condition. However, if the subject does not activate any visual images, looking behaviors toward a visual image in the two conditions should not differ.

We used a nonprimate species, the domestic dog (*Canis familiaris*) as a subject. The domestic dog is a highly social species that shares many characteristics of complex social systems known in nonhuman primates. Recent molecular genetic analyses have shown that dogs diverged from the common ancestor with wolves somewhere between 35,000 and 100,000 years ago (Vilá et al. 1997). During their long history of domestication and close cohabitation with humans, there has been selection for sophisticated skills for interacting and communicating with humans.

It is already known that dogs can use human attentional cues (body/head orientation) in a food-choice situation (Soproni et al. 2002). They are also able to direct human attention toward the location of hidden food (Miklósi et al. 2000). Hare et al. (2002) showed that dogs may even be more skillful than great apes at a number of tasks in which they must read human communicative signals that indicate the location of hidden food. In that study, wolves raised by humans performed poorly in such tasks, whereas dog puppies raised by their biological mothers did well. This suggests that during the process of domestication, dogs have been selected for a set of social-cognitive abilities that enable them to communicate efficiently with humans.

Close cohabitation may also enhance other cognitive skills. For example, flexible formation of categories could be advantageous for dogs sharing the same environments with humans. A recent study demonstrated that a dog learned names of more than 200 objects. He even learned the names of novel items using the exclusion principle after a single trial, and correctly retrieved those items up to 4 weeks after the initial exposure. (Kaminski et al. 2004, but also see, Markman and Abelev 2004; Fischer et al. 2004).

In this study, we investigated whether dogs would actively recall the visual representation of a person when they hear the voice of that person. We used a cross-modal version of the expectancy violation procedure. We played back an owner's or a stranger's voice and then presented a face which either matched or mismatched with the preceding voice. We hypothesized that, if the subject activated the visual image of the

Table 1 Dogs breed and their age in this experiment

Dogs' breed	Number of subjects	Age of dogs (years)
American Cocker Spaniel	2	4, 7
Cavalier King Charles Spaniel	5	5, 5, 5, 5, 8
French Bulldog	1	1
Golden Retriever	3	2, 4, 5
Miniature Dachshund	5	1, 1, 1, 1, 3
Mixed	1	7
Shiba dog	1	4
Welsh Corgi	8	1, 1, 2, 3, 3, 6, 6, 9
West Highland White Terrier	2	6, 11

caller upon hearing the caller's voice, the subject would be surprised at the mismatched combinations, and thus would look at the photo longer than they do in matched conditions.

Experiment

Subjects

We recruited dogs and their owners through a personal acquaintance network from the following four cities and prefectures: Kyoto, Kyoto; Otsu, Shiga; Nishiwaki, Hyogo; Inuyama, Aichi. Participation in the test was voluntary. The dogs included various breeds (see Table 1); there were 15 males and 13 females. All dogs had lived with human families since they were puppies. They were naive with respect to any behavioral experiments.

Stimuli and apparatus

Four test stimuli were prepared for each animal. These were: a photo of the owner against an ivory-colored background (PO); a photo of another person of the same sex unfamiliar to the dog (PN); the voice of the owner calling the dog's name (VO); the voice of the unfamiliar person calling the dog's name (VN). Other characteristics such as clothes, hairstyles, and ages were uncontrolled. We asked each person to call the dog's name and recorded the call on a mini-disk recorder, then stored the digitized voice on computer in WAV format. The sampling rate was 44,100 Hz and the sampling resolution was 16-bit. The duration of the paired voices was approximately the same because they called the same dog's name. The amplitude of the voices appeared equivalent to human ears. We also took a digital photo of a full face of each person smiling and stored the photo on computer in JPEG format of the size 450 (W) × 550 (H) pixels, or ca. 16 cm × 20 cm on the 18.1-in. LCD monitor (SONY SDM-M81) we used. The background and the size of the paired photos was the same.

We set up a transportable apparatus as shown in Fig. 1 at the house of the owner or a friend of the owner. Briefly, the LCD monitor was located about 1 m from the subject's nose. At the beginning of the test, a black opaque screen (90 cm × 90 cm) was placed in front of the monitor to prevent the subject from seeing the monitor. Two digital camcorders (SONY DCR-TRV-30), one located behind the subject and the other behind the LCD monitor, recorded the subject's behavior. Presentation of the stimuli was controlled via a Visual Basic 5.0 program on a laptop personal computer (Dell Inspiron 4100, with Pentium III 1.2 GHz) located in an adjacent room.

Procedure

We used an expectancy violation procedure based on one often used to test human infants for inferences about external events. Typically, the subject is shown one event several times followed by a second event. It is assumed that the subject should look longer at the second event if the latter contradicts the outcome expected based on the first event.

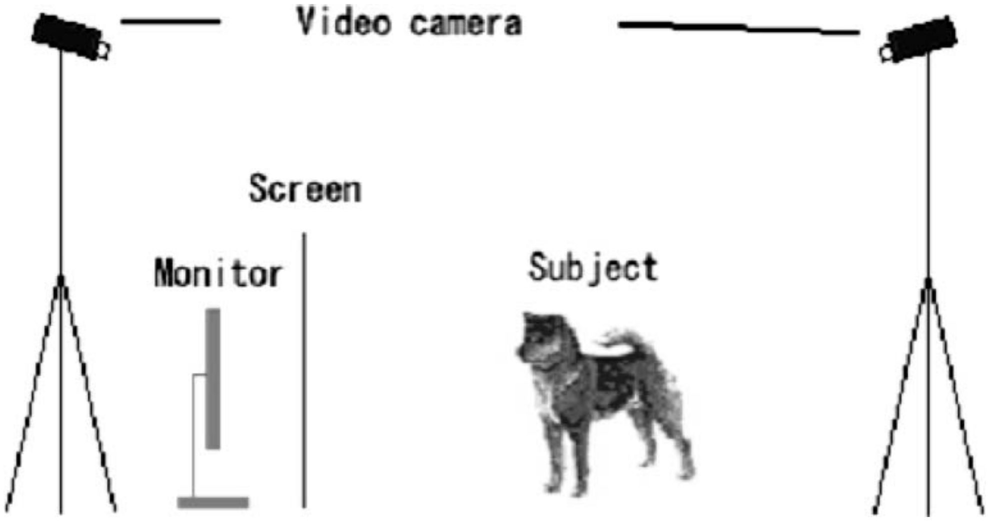

Fig. 1 A schematic drawing of the apparatus

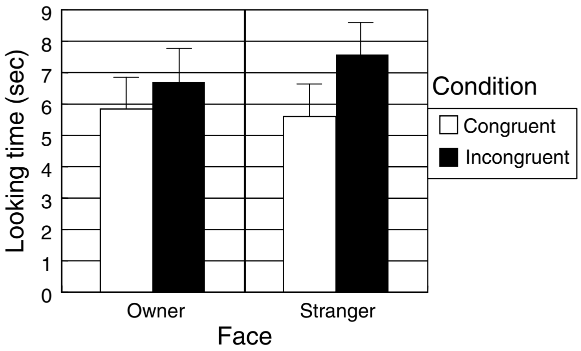

Fig. 2 The mean duration of looking at the monitor in the photo phase for each condition averaged for all subjects. The two bars located in the middle indicate the results in the Incongruent condition. The two bars located at both ends indicate the results in the Congruent condition

We extended the expectancy violation procedure to test multi-modal recognition of correlated events in the dogs. Each trial consisted of the following events: The dog either sat or lay down in front of the LCD monitor. We asked one of the owner's family or friends, familiar to the dog, to lightly restrain the animal either with a harness or by simply holding the dog. The restrainer was ignorant of the purpose of the experiment, and was instructed not to speak to the dog or make any movements that might influence the dog's attention or behavior during the test. An experimenter, who observed the subject via a 4-in. television monitor connected to the camcorder behind the LCD monitor, started the trial when the subject looked at the center of the screen. Each trial consisted of two phases. The first was the voice phase and the second was the photo phase. In the voice phase, one of the two voices was played back from the speakers installed in the monitor every 2 s, for a total of five presentations. The duration of the voice was about 750 ms but varied slightly depending on the subject's name and the caller. The photo phase began immediately after the final call. In the photo phase, the opaque screen was removed by a second experimenter to reveal the photo of a face on the LCD monitor. The experimenter was always positioned behind the screen and was ignorant of the photo shown on the monitor. Screen removal was done calmly and smoothly to avoid disturbing or frightening the subject. The photo phase lasted 30 s. The subject's looking behaviors toward the visual stimulus in this phase were video-recorded for later analyses.

Each dog was given the following four test trials: In the VO-PO trial, the owner's photo appeared after the owner's voice. In the VN-PN trial, the photo of an unfamiliar person appeared after his/her voice. In the VN-PO trial, the owner's photo followed the voice of the unfamiliar person. In the VO-PN trial, the photo of the unfamiliar person followed the owner's voice. The voice and the photo matched in the former two trials (Congruent trials), whereas they mismatched in the latter two (Incongruent trials).

These four test trials were presented in pseudo-random order with the restriction that the same voice was not repeated on consecutive trials. The inter-test trial interval was about 10 min.

Results

After the experiments, the videos of trials were captured on a personal computer and converted into MPEG file format (30 frames per second). A coder who was blind to the stimuli recorded the duration of subjects' looking at the monitor in the photo phase. A second coder scored eight randomly sampled subjects to check the reliability of coding. Total looking time for each trial rated by the second coder was compared with that by the first coder. The correlation was highly positive (Pearson's $r = 0.952$, $n = 32$, $p < 0.01$).

We calculated total looking time for each trial on each subject (Fig. 2). Those looking times were analyzed by a 2 × 2 repeated-measures analysis of variance with faces (Owner or Unfamiliar person) and conditions (Congruent or Incongruent) as factors. This analysis showed a significant main effect of conditions ($F(1,27) = 6.489, p = 0.017$), but no significant main effect of faces ($F(1,27) = 0.708$, $p = 0.408$) nor was there a significant interaction between conditions and faces ($F(1,27) = 0.099$, $p = 0.756$). These results indicate that the subjects' looking time in the two Incongruent trials was significantly longer than those in the two Congruent trials regardless of face stimuli.

Discussion

We tested dogs with an expectancy violation procedure in which we presented a photograph, either of the owner's face or an unfamiliar person's face, after playing back a voice. The voice and face matched in half of the trials (Congruent condition) and mismatched in the other half (Incongruent condition). We found that subjects' looking times toward the monitor were longer in the Incongruent condition than those in the Congruent condition. This result suggests that dogs actively generate an internal representation of the owner's face when they hear the owner calling them. This is the first demonstration that nonhuman animals do not merely associate auditory and visual stimuli but also actively generate a visual image from auditory information.

Interestingly, the dogs also looked at the visual stimulus longer in the Incongruent condition in which the owner's face followed an unfamiliar person's voice (VN-PO trial) than in the Congruent condition in which the owner's face followed the owner's voice (VO-PO trial). Generating a particular visual image in response to an unfamiliar voice should be difficult, and any expected image from the voice ought to be more obscure or less well defined than that of owners. However, our subjects looked more at the owner's face in the Incongruent condition than in the Congruent condition. This may indicate that dogs use exclusion rules; that is, they may have predicted that the owner should not appear after hearing the unfamiliar person's voice. This explanation is supported by evidence that dogs can fast map words on novel items (Kaminski et al. 2004). Fast mapping also needs the ability to use the exclusion rule. Here, the dog must understand that a new name should not be for familiar objects.

One possible confounding factor might be that while restraining the dog, the owner's family member or friends unwittingly provided the dog with cues. If we used a choice task, then a "Clever Hans" effect could have affected the subjects' behaviors. But in our task the restrainers were ignorant of the hypothesis, and they were ignorant of the stimuli presented on the monitor because they were asked to look at the small monitor on the cam coder to keep the dogs on film.

The present finding reminds us of the impressive recognition of alarm calls reported in vervet monkeys and Diana monkeys. In vervets, Seyfarth and Cheney (1992) showed that habituation to one type of alarm call transferred to another call with a similar meaning but quite different acoustic features. Those authors suggested that the monkeys recognized the referent (i.e., meaning) of the call upon hearing it. Zuberbühler (2000a, b, c) obtained similar results in Diana monkeys. These studies suggest that monkeys might generate representations of predators when they hear the corresponding alarm calls.

The present demonstration in dogs is similar to those described above in monkeys but is different in one important aspect: alarm calls work as a signal (an index), with the predator as a referent. Thus they are not regarded as an exemplar but as a label for a certain category. This indexical function is different from one modality-specific exemplar recalling another of the same category. This function of exemplars has never been demonstrated in nonhumans. We suggest that conceptual representations by nonhuman organisms may be much more like that of humans than has previously been assumed.

In order to understand the evolution and function of cognitive skills, we need comparative studies. The domestic dog is a promising species for studying the evolutionary emergence of cognitive abilities (Miklósi et al. 2003). They are known to show highly sophisticated social skills in interspecific interactions with humans (Miklósi et al. 1998; Soproni et al. 2002; Virányi et al. 2004), even in comparisons with their ancestors, the wolf (Miklósi et al. 2003), and with apes (Soproni et al. 2002; Hare et al. 2002). Accordingly, the question arises if other cognitive skills of dogs have also been enhanced through close cohabitation with humans. Kaminski et al. (2004) showed that a dog can learn names of more than 200 objects and also can map fast. Here, we provide additional evidence that dogs have the sophisticated cognitive skills required to form categories.

Acknowledgments We wish to thank all the owners of the dogs who volunteered to participate in our experiments. This study was supported by the Research Fellowships of the Japan Society of the Promotion of Science (JSPS) for Young Scientists to Ikuma Adachi and Hiroko Kuwahata, the Grants-in-Aid for Scientific Research Nos. 13410026, 14651020, and 17300085 from JSPS to Kazuo Fujita, and by the 21st Century COE Program, D-10, from Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan, to Kyoto University. We also thank James R. Anderson for his editing of the manuscript.

References

Aydin A, Pearce JM (1994) Prototype effects in categorization by pigeons. J Exp Psychol ABP 20:264–277

Cerella J (1979) Visual classes and natural categories in the pigeon. J Exp Psychol HPP 5:68–77

Cook RG, Wright AA, Kendrick DF (1990) Visual categorization by pigeons. In: Commons ML, Herrnstein RJ, Kosslyn SM, Munford DB (eds) Quantitative analysis of behavior, vol 8. Hillsdale, NJ: Erlbaum, pp 187–214

Evans TA, Howell S, Westergaard GC (2005) Auditory-visual cross-modal preception of communicative stimuli in tufted capuchin monkeys (Cebus apella). J Exp Psychol ABP 31:399–406

Fischer J, Call J, Kaminski J (2004) A pluralistic account of word learning. Trends Cogn Sci 8:481

Ghazanfar AA, Logothetis NK (2003) Neuroperception: Facial expressions linked to monkey calls. Nature 423:937–938

Hare B, Brown M, Williamson C, Tomasello M (2002) The domestication of social cognition in dogs. Science 298(5598):1634–1636

Hashiya K, Kojima S (1999) Auditory-visual intermodal matching by a chimpanzee (*Pan troglodytes*). Primate Res 15:333–342

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

Herrnstein RJ, Loveland DH (1964) Complex visual concept in the pigeon. Science 146:549–551

Herrnstein RJ, Loveland DH, Cable C (1976) Natural concepts in pigeons. J Exp Psychol ABP 2:285–302

Kaminski J, Call J, Fischer J(2004) Word learning in a domestic dog: Evidence for "fast mapping". Science 304:1682–1683

Lea SEG, Ryan CME (1993) Featural analysis of pigeons' acquisition of discrimination between letters. In: Commons ML, Herrnstein RJ, Wagner AR (eds) Quantitative analyses of behavior, vol 4. Cambridge, MA: Ballinger, pp 239–253

Markman EM, Abelev M (2004) Word learning in dogs? Trends Cogn Sci 8:479–481

Miklósi Á, Kubinyi E, Topál J, Gácsi M, Virányi Z, Csányi V (2003) A simple reason for a big difference: wolves do not look back at humans but dogs do. Curr Biol 13:763–767

Miklósi Á, Polgárdi R, Topál J, Csányi V (1998) Use of experimenter-given cues in dogs. Anim Cogn 1:113–121

Miklósi Á, Polgárdi R, Topál J, Csányi V (2000) Intentional behaviour in dog—human communication: experimental analysis of 'showing' behaviour in dogs. Anim Cogn 3:159–166

Seyfarth RM, Cheney DL (1992) Meaning and mind in monkeys. Sci Am 78–84

Soproni K, Miklósi A, Topál J, Csányi V (2002) Dogs' (*Canis familiaris*) responsiveness to human pointing gestures. J Comp Psychol 116:27–34

Vilá C, Savolainen C, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, Crandall KA, Lundeberg J, Wayne RK (1997) Multiple and ancient origins of the domestic dog. Science 276:1687–1689

Virányi Z, Topál J, Gácsi M, Miklósi Á, Csányi V (2004) Dogs can recognize the behavioural cues of the attentional focus in humans. Behav Proc 66:161–172

Wasserman EA, DeVolder CL, Coppage DJ (1994) Non-similarity-based conceptualization by pigeons via secondary or mediated generalization. Psychol Sci 3:374–379

Yoshikubo S (1985) Species discrimination and concept formation by rhesus monkeys (*Macaca mulatta*). Primates 26:285–299

Zuberbühler K (2000a) Referential labeling in Diana monkeys. Anim Behav 59:917–927

Zuberbühler K (2000b) Causal cognition in a non-human primate: Field playback experiments with Diana monkeys. Cognition 76:195–207

Zuberbühler K (2000c) Causal knowledge of predators' behaviour in wild Diana monkeys. Anim Behav 59:209–220