

Dogs might associate the voices and faces of human girls and adult women

Hitomi Chijiwa ^{a, b, *} , Yasuhiro Kanakogi ^a 

^a Graduate School of Human Sciences, The University of Osaka, Osaka, Japan

^b Japan Society for the Promotion of Science, Tokyo, Japan

ARTICLE INFO

Article history:

Received 17 October 2025

Initial acceptance 24 February 2026

Final acceptance 22 May 2026

Available online 9 September 2026

MS. number: 25-00625R

Keywords:

cross-modal representation

dog

dog–child interaction

expectancy violation

social cognition

When we hear a person's voice, we can often infer attributes such as age or identity, even without seeing the speaker. This ability suggests the presence of a cross-modal representation that integrates information from both the visual and auditory modalities. Such representations are not limited to interactions within the same species but also occur between different species. Previous studies have shown that dogs, *Canis familiaris*, can match human voices with faces based on familiarity, sex and emotional expression. However, it remains unclear whether they can distinguish humans by age across modalities. Such ability could be important for dogs to adapt their behaviour appropriately towards children versus adults in everyday interactions. This study investigated whether dogs can form cross-modal representations of human adults and children using an expectancy-violation paradigm. We hypothesized that if dogs possess such representations, they would exhibit longer looking times during voice–face mismatched (incongruent) trials compared to voice–face matched (congruent) trials for both adult and child stimuli. The experiment involved seating dogs in front of a monitor. Immediately after hearing the voice of an adult woman or a girl played through a speaker, the dogs were shown a photograph of a face that either matched or mismatched the voice. We then measured gaze duration towards the monitor. Contrary to our prediction, dogs tended to look longer in congruent than in incongruent conditions, although this effect was only marginally significant. This pattern suggests that dogs may discriminate between congruent and incongruent voice–face pairings, but not in the predicted direction. Exploratory analyses further suggested that this tendency was more pronounced among dogs who had lived with children. Overall, these findings provide tentative evidence that dogs may associate human female voices and faces across age classes, and that experience with children could influence this process.

© 2026 The Authors. Published by Elsevier Ltd on behalf of The Association for the Study of Animal Behaviour. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Recognizing other individuals by integrating different sensory modalities is crucial for social species living in complex societies, as it helps them distinguish individuals, avoid conflict, maintain social hierarchies and effectively cooperate. Humans possess cross-modal representations that integrate information from both visual and auditory modalities (Campanella & Belin, 2007). For example, when a familiar singer's song is played on the radio, we can easily bring their face to mind, even without seeing them. Several nonhuman animals are capable of forming auditory–visual representations of conspecifics, including chimpanzees, *Pan troglodytes* (Kojima et al., 2003), rhesus macaques, *Macaca mulatta*

(Sliwa et al., 2011), horses, *Equus caballus* (Proops et al., 2009), lions, *Panthera leo* (Gilfillan et al., 2016), goats, *Capra hircus* (Pitcher et al., 2017) and crows, *Corvus macrorhynchos* (Kondo et al., 2012). These findings suggest that individuals in these species may form integrated representations linking vocalizations and visual features, although the underlying cognitive mechanisms remain unclear.

Cross-modal representations are not limited to conspecific interactions but may also extend to heterospecifics. Some animals can form such associations for familiar humans, including owners or caretakers, as demonstrated in rhesus macaques (Sliwa et al., 2011), squirrel monkeys, *Saimiri boliviensis* (Adachi & Fujita, 2007), dogs, *Canis familiaris* (Adachi et al., 2007), cats, *Felis catus* (Takagi et al., 2019) and horses (Proops & McComb, 2012). For

* Corresponding author.

E-mail address: h-chijiwa0609@hus.osaka-u.ac.jp (H. Chijiwa).

example, [Adachi et al. \(2007\)](#) employed an expectancy violation paradigm and presented dogs with a photo of either their owner's or a stranger's face after playing back one of their voices. Dogs looked longer when the face did not match the voice than when the stimuli did match, suggesting the possibility that dogs may recall their owner's face upon hearing the owner's voice. While some nonhuman primates have shown cross-modal representations of humans, these results typically derive from lab-housed animals. In the wild, they rarely encounter humans; therefore, this ability may not be ecologically relevant for them. In contrast, for companion animals, such as dogs, cats and horses, living with humans is their natural environment. Their survival and wellbeing depend heavily on their interactions with humans. For these animals, distinguishing human characteristics is probably critical in daily life.

Among companion animals, dogs have been studied frequently, and their ability to form cross-modal representations of human traits has been demonstrated empirically. Dogs can match not only human familiarity (owner or stranger, [Adachi et al., 2007](#)) but also human gender ([Ratcliffe et al., 2014](#)) and emotional expressions ([Albuquerque et al., 2016](#)) across modalities. However, whether dogs can distinguish between humans of different age groups remains unclear. Age is an essential characteristic used to categorize others, and recognizing the age class of social partners is important for many species. It allows individuals to adapt their behaviour appropriately, such as avoiding juveniles as mates or showing tolerance towards them. Notably, a recent study reported no evidence for cross-modal representations of conspecific age classes in dogs ([Kawaguchi et al., 2024](#)). However, this null finding may reflect methodological limitations rather than a true absence of the ability. For example, large spatial separation between visual stimuli and pronounced size differences between adult and puppy images may have biased attention towards more salient stimuli, potentially obscuring any congruency effect. Thus, it remains unresolved whether dogs lack cross-modal representations of conspecific age classes or whether such representations were not detectable under the experimental conditions employed. Importantly, the absence of evidence for cross-modal representations of conspecific age classes does not preclude the possibility that dogs may form cross-modal representations of human age classes. Given their long history of domestication and close association with humans, dogs may be particularly attuned to human-related social categories. Supporting this possibility, recent evidence indicates that horses, another domesticated species living in close association with humans, can form cross-modal representations of human age groups, distinguishing between adults and children using vocal and facial cues ([Jardat et al., 2023](#)). Therefore, dogs may be more likely to exhibit such representations for humans than for conspecifics.

Previous studies have suggested that dogs are responsive to human children. For example, their cortisol levels increase in response to hearing an infant crying, relative to other auditory stimuli, indicating sensitivity to infant-related cues, although this study did not directly compare responses to adult vocalizations ([Yong & Ruffman, 2014](#)). They can also form attachment bonds with children and synchronize behaviours with them, although to a lesser degree than with adults ([Wanser et al., 2020, 2021](#)). Despite risks such as biting, child–dog relationships are generally positive ([Giraudet et al., 2022](#)). This might be due in part to dogs' tolerance and patient responses to children's unpredictable behaviour. Such behavioural patterns suggest that dogs may perceive children as distinct from adults, which indicates dogs may have a representation of human age classes. However, it remains unknown whether they integrate visual and auditory cues into coherent 'child' or 'adult' categories. A general examination of

whether dogs can distinguish between human adults and children can help ascertain whether dogs adapt their behaviour appropriately to different human age groups, which has important implications for promoting safe and positive child–dog relationships.

Dogs' sensitivity to human social categories may also depend on their prior experience. For example, dogs exposed to a wider range of men and women show stronger cross-modal sensitivity to human gender cues than dogs with more limited exposure ([Ratcliffe et al., 2014](#)). This suggests that such abilities may be shaped by experience rather than being uniformly expressed. Similarly, dogs that have lived with young children may have greater opportunities to learn and associate the characteristic visual and auditory features of children. Accordingly, these dogs may exhibit stronger cross-modal sensitivity to human age-related cues.

This study investigated whether dogs can form cross-modal representations of human adult and child voices and faces, using the expectancy-violation method. We predicted that if such representations exist, dogs would look longer at mismatched rather than matched auditory–visual (voice–photo) pairings. We further predicted that this effect would be more pronounced in dogs with prior experience of living with young children.

METHODS

Subjects

Sixty-nine domestic dogs (40 males, 29 females; mean age $\pm$ SD = 5.6 ± 3.1 years; various breeds) participated. The initial sample size ($N = 48$) was determined based on practical feasibility and on sample sizes commonly used in comparable studies of canine cognition (approximately 30–50 individuals; e.g. [Kawaguchi et al., 2024](#); [Korzeniowska et al., 2022](#); [Surányi et al., 2025](#)). Because the initial pattern of results was unexpected and weak, the sample was increased to 69 dogs to assess the robustness of the findings. An additional seven dogs were excluded because they failed to look at the stimuli in all four test trials. Owners and their dogs were initially recruited through our laboratory's participant database, which was originally established for developmental studies with infants and children and later expanded to include dog volunteers. Additional participants were recruited via personal contact, community flyers and word of mouth. Detailed information for each dog, including age, sex, breed and child-experience status, is provided in the online Supplementary Material.

Stimuli

Facial photographs were taken against a white background, and voices were recorded from two young girls (6 years old) and two women (48 and 53 years old, mean = 50.5 years), all strangers to the dogs. Speakers were instructed to say a standard sentence in Japanese ('Hello, you are cute'). The acoustic characteristics of the sounds were analysed using Praat version 6.4.18 (<https://www.fon.hum.uva.nl/praat>). The mean fundamental frequency (F0) $\pm$ SD was 310.51 ± 14.17 Hz for children's voices and 262.45 ± 24.79 Hz for adults' voices. The mean F0 range (maximum minus minimum F0) was 137.07 ± 81.75 Hz for children's voices and 148.20 ± 8.70 Hz for adults' voices. Only female stimuli were used, consistent with previous research in horses ([Jardat et al., 2023](#)). Female stimuli were also used since distinguishing between boys and men may be easier due to larger acoustic differences in vocal characteristics such as F0 ([Stathopoulos et al., 2011](#)), whereas the contrast between girls and women provides a more challenging test. Each dog was presented with one child and one adult,

meaning that, across all trials, they were exposed to two individuals. The four possible adult–child pairings (two adults $\times$ two children) were randomly assigned across participants.

Procedure

In each trial, the dog sat in front of a monitor while being gently held by the owner from behind, with hands placed on the dog's shoulders (Fig. 1). No lead was used. Owners were instructed to close their eyes or look downwards during stimulus presentation, and compliance was verified by the experimenter before each trial. Owners were blind to the experimental conditions. The presentation of auditory and visual stimuli was controlled using Visual Studio 2022 (Microsoft, U.S.A.) on a laptop. A voice (child or adult) was played twice through a Bluetooth speaker (Anker SoundCore mini, <https://www.soundcore.com>) positioned behind the monitor. Immediately after the voice was played, a face photograph that either matched or mismatched the voice appeared on the monitor for 10 s. The 28×29 cm photographs were presented on a 27-inch monitor. Four conditions were tested: adult–congruent, adult–incongruent, child–congruent and child–incongruent. Each dog completed one trial per condition, with intertrial intervals of at least 3 min. The number of trials was limited to maintain the dogs' attention and minimize inattentive trials. The order of trials was pseudorandomized to avoid consecutive repetitions of the same voice. Sessions were recorded with three video cameras: two (GoPro HERO 12, <https://gopro.com>; JVC GZ-RX670-A) positioned in front of the dog and another (Sony HDR-CX680) positioned diagonally behind the dog.

To assess dogs' experience with children, owners reported whether their dog had lived with children aged 3–6 years and whether the dog had lived with minors under 18 years of age. The 3–6-year range was selected to match the age of the child stimuli (6-year-old girls) and to avoid including broader age groups with substantially different vocal characteristics. This range

corresponds to preschool-aged children who are beyond infancy and capable of producing relatively fluent speech. In Japan, children aged 6 years are typically still considered preschool-aged, as compulsory schooling begins at age 7; therefore, we regarded ages 3–6 years as a coherent developmental category.

Statistical Analyses

Looking time was defined as the duration for which the dog oriented its head towards the centre of the monitor with gaze directed at the screen. A coder, blind to the conditions, counted the number of frames (30 frames/s) within each 10 s trial in which the dog was looking at the monitor. Looking time was then converted to seconds for analysis. The videos were analysed using Adobe Premiere Pro 2025 (Adobe Inc., U.S.A., <https://www.adobe.com>) software. Trials were excluded if the dog did not look at the monitor (92 trials in total). An additional seven trials were excluded due to camera error (four), intervention by the owner (one) or distraction due to outdoor noise (two). As a result, 177 of 276 trials were included as valid. Table 1 presents the number of valid trials per condition. Inter-coder reliability was assessed on 20% of randomly selected videos, independently coded by a second coder, who was also blind to the conditions. The correlation between the two coders for the looking time was strong (Pearson correlation $r = 0.84$, $N = 58$, $P < 0.001$), indicating good reliability.

All analyses were performed using R version 4.5.1 (R Core Team, 2025). Looking time was analysed using linear mixed models (LMMs) implemented in the lmer function of the lme4 package (Bates et al., 2015). An initial model included the interaction between face (adult/child) and congruency (congruent/incongruent); as this interaction was not significant, the final model included congruency, face, trial (1–4), experience (living with children aged 3–6 years; yes/no) and the interaction between congruency and experience as fixed effects. Participant identity was included as a random effect. The congruency $\times$ experience interaction was included to test whether experience with children altered the pattern of responses to congruent versus incongruent trials, rather than simply affecting dogs' overall attention. Fixed effects were evaluated using Type II F tests via the Anova function in the car package (Fox & Weisberg, 2019), with denominator degrees of freedom estimated using the Kenward–Roger approximation (Kenward & Roger, 1997), which provides more accurate estimates of degrees of freedom in linear mixed models, particularly for unbalanced designs. To assess the robustness of findings, an additional model was fitted using a broader definition of child experience (living with minors under 18 years).

Given our a priori interest in the possible role of experience with children, supplementary analyses were conducted separately for dogs with and without experience of living with children aged 3–6 years to examine group-specific congruency effects. To evaluate the potential effects of trial order on missing data, we analysed the probability of attending to the screen (looking versus not looking) using a binomial generalized linear mixed model (GLMM). As a further robustness check, looking time was analysed as a proportion of total trial duration using a beta-error GLMM. For both GLMMs, Type II Wald χ^2 tests were conducted using the Anova function in the car package (Fox & Weisberg, 2019).

Ethical Note

All procedures were approved by the Animal Experiment Ethics Committee of the Graduate School of Human Sciences, The University of Osaka (approval number: AHB-R6-1-0). Dog owners were registered volunteers from the university's child and pet participant database and participated voluntarily after giving

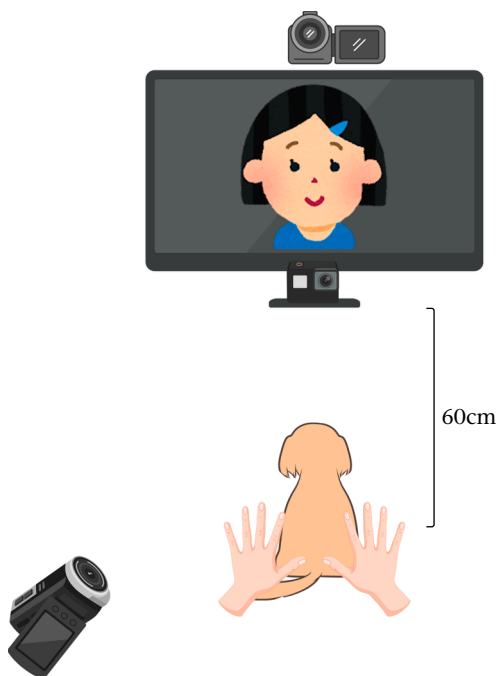


Figure 1. Schematic showing apparatus used. The owner held the dog 60 cm away from the monitor. The monitor was connected to the laptop via an HDMI cable. Three cameras recorded the test, two positioned in front and one diagonally behind the dog, respectively. The speaker was placed behind the monitor.

Table 1
The number of valid data points representing the number of trials with dogs looking at the monitor in each condition

	Adult face		Child face	
	Congruent voice	Incongruent voice	Congruent voice	Incongruent voice
Trials	51	42	38	46

informed consent. The experiment was entirely noninvasive and conducted in a quiet room. Dogs were free to leave the testing area at any time if they showed signs of discomfort. No food or water restriction was applied, and sessions were kept brief to minimize potential stress.

RESULTS

Fig. 2 shows the average looking time at the monitor for all dogs. Mean looking times $\pm$ SD (s) for each condition were as follows: adult–congruent, 3.99 ± 3.22 ; adult–incongruent, 3.08 ± 2.30 ; child–congruent, 4.31 ± 3.21 ; and child–incongruent, 3.64 ± 3.28 .

The LMM revealed a marginal effect of congruency ($F_{1,131.38} = 3.20$, $P = 0.076$, partial $\eta^2 = 0.03$), with dogs showing a tendency to look longer at the monitor in the congruent condition than in the incongruent condition. There were no significant effects of face ($F_{1,126.64} = 1.22$, $P = 0.272$, partial $\eta^2 = 0.009$), trial ($F_{1,125.84} = 0.01$, $P = 0.917$, partial $\eta^2 < 0.001$) or experience ($F_{1,65.66} = 0.49$, $P = 0.485$, partial $\eta^2 = 0.006$). The interaction between congruency and experience was also not significant ($F_{1,135.32} = 1.26$, $P = 0.265$, partial $\eta^2 = 0.006$). Using the broader definition of child experience (living with minors under 18 years of age) yielded a similar pattern of results (Table S1 in the Supplementary Material). Specifically, there was again a marginal effect of congruency ($F_{1,130.43} = 3.09$, $P = 0.081$, partial $\eta^2 = 0.03$), whereas neither the main effect of experience nor its interaction with congruency was significant ($P_s > 0.300$).

Separate analyses for dogs with and without experience of living with children aged 3–6 years showed that the tendency to look longer in congruent than in incongruent conditions was observed only in dogs with such experience (Fig. S1 in the Supplementary Material). In this group, congruency had a significant effect ($F_{1,33.93} = 4.32$, $P = 0.045$, partial $\eta^2 = 0.11$), whereas no such effect was observed among dogs without child experience ($P = 0.352$; Table S2 in the Supplementary Material).

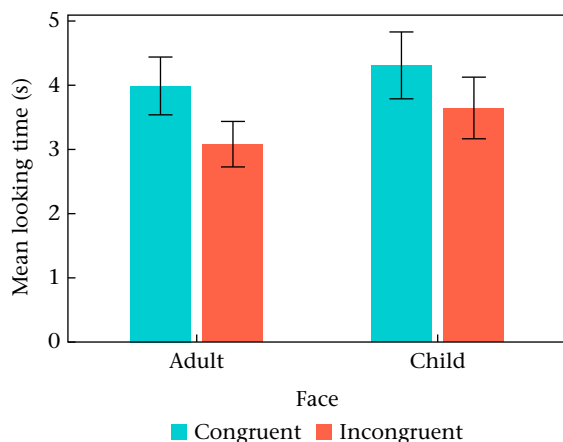


Figure 2. Mean looking time at the monitor across conditions ($N = 69$). Error bars indicate standard errors. Blue bars represent congruent conditions and red bars incongruent conditions.

A binomial GLMM examining the probability of attending to the screen (looking versus not looking) showed that dogs were less likely to look at the screen in later trials ($\chi^2_1 = 6.00$, $P = 0.014$), indicating a habituation effect. However, neither congruency nor its interaction with trial significantly affected the likelihood of looking ($P_s > 0.880$), suggesting that the exclusion of trials with no looking did not systematically bias the comparison between congruent and incongruent conditions (Table S2 in the Supplementary Material).

Finally, a beta-error GLMM analysing looking time as a proportion of total trial duration yielded the same overall pattern of results, with a marginal effect of congruency ($\chi^2_1 = 2.83$, $P = 0.093$), indicating that dogs tended to look longer in congruent than incongruent conditions (Table S2 in the Supplementary Material).

DISCUSSION

Our results showed that dogs tended to look longer in congruent than in incongruent conditions for both adult and child stimuli. This effect was marginal when considering the full sample. Exploratory analyses indicated that the effect reached significance only in dogs with experience of living with children aged 3–6 years. However, because the interaction between congruency and experience was not significant in the main model, this finding should be interpreted with caution. Nevertheless, the observed pattern raises the possibility that direct experience with young children may enhance dogs' sensitivity to auditory–visual consistency between children's and adults' faces and voices.

These findings support the view that dogs' perceptions of human social categories are flexible and may be shaped by their developmental environment and everyday experiences with humans (Ratcliffe et al., 2014). This contributes to ongoing debates regarding whether canine sociocognitive abilities primarily reflect evolutionary processes associated with domestication (Hare et al., 2002; Hare & Tomasello, 2005) or are shaped by learning and experience (Udell et al., 2008, 2010). Although the present study does not directly address this distinction, as it lacks a comparative approach, the exploratory findings suggest that direct experience with young children may play a role in shaping dogs' sensitivity to human age-related cues.

We predicted that dogs would look longer in incongruent conditions than in congruent conditions. However, the findings showed the opposite trend: dogs spent more time looking in congruent trials. One explanation is that the unfamiliarity of the stimuli prevented expectancy violations from occurring, leaving dogs in an early processing stage regarding the representation between the voices and faces of children and adults. In developmental research, human infants show 'familiarity effects' when stimuli are novel or insufficiently encoded, while 'novelty effects' emerge only after learning or repeated exposure (Hunter & Ames, 1988; Kidd et al., 2012; Powell & Spelke, 2018). Similarly, dogs in this study may have been more attentive to congruent pairings because they appeared more coherent when identity cues were absent. Adachi et al. (2007) found expectancy violations occurred when dogs were presented with their owner's voice and face, highlighting the importance of familiarity. In contrast, in our

study, the absence of familiar voices and faces may have limited the possibility of expectancy-violation responses. The relatively short overall looking times observed in the present study may further reflect the limited salience of unfamiliar stimuli. However, the use of unfamiliar individuals was necessary to isolate responses to age categories rather than to specific familiar individuals. Future studies could test this interpretation by examining dogs with greater experience of particular age groups or using more naturalistic auditory–visual stimuli, such as dynamic videos of speaking individuals rather than static photographs paired with recorded voices. Such stimuli may allow dogs to make fuller use of age-related cues.

Interestingly, similar effects have been reported in cats for object permanence tasks using a violation-of-expectancy paradigm: contrary to expectations, cats showed greater interest in knowledge-consistent events than in violation events (Forman *et al.*, 2025). This indicates that our findings are not unique to dogs and that animals may sometimes respond to expectancy violations in unexpected ways, warranting further research into the motivational and cognitive explanations for this response. Nevertheless, substantial variation was observed in the present study: among dogs with at least one valid congruent and one valid incongruent trial, 23 showed longer looking in congruent than in incongruent trials, whereas 24 showed the opposite pattern. This variability suggests heterogeneity in the extent to which individual dogs form or use cross-modal expectations about human age categories.

Although a previous study did not find evidence that dogs form cross-modal representations of conspecific age classes (Kawaguchi *et al.*, 2024), the present findings suggest that they may form such representations for human age classes, as indicated by their sensitivity to whether a face matched the age implied by a voice. Domestication may have shifted dogs' social attention. Marshall-Pescini, Schwarz, *et al.* (2017) demonstrated that wolves outperform dogs in conspecific cooperation tasks, suggesting that dogs' motivation or cognitive skills for interacting with their own species may have diminished through domestication. Moreover, compared to their wild counterparts, dogs rely less on pack members for both foraging and pup rearing, which may have led to less sensitivity to puppy cues (Marshall-Pescini, Cafazzo, *et al.*, 2017). However, dogs are highly attuned to human communicative signals, such as understanding pointing gestures and gaze following (Miklósi *et al.*, 1998; Soproni *et al.*, 2001; Téglás *et al.*, 2012). Consequently, conspecific age information may have lost social relevance, while human social categories may have become more central to dogs' everyday interactions.

This study has some limitations. The first limitation lies in methodological choice. We employed a single-stimulus design rather than the preferential-looking paradigm (presenting two pairs). Preferential looking has been widely adopted in cross-modal research, with studies showing that animals, including dogs, tend to look longer at congruent auditory–visual pairings (e.g. Albuquerque *et al.*, 2016; Proops & McComb, 2012; Quaranta *et al.*, 2020; Ratcliffe *et al.*, 2014, but see Jardat *et al.*, 2023; Trösch *et al.*, 2019). However, while preferential looking enables direct comparison between two simultaneous stimuli, it may also create attentional competition, which may obscure subtle differences in cognitive processing. By presenting one stimulus at a time, we sought to isolate the dog's responses to each individual pairing without the confounding influence of competing visual stimuli. This enabled us to more clearly assess the degree of attentional engagement elicited by each condition. Whether similar results can be replicated using a preferential-looking paradigm remains an important question.

Second, we examined only two broad age categories (adults versus children) using a limited set of stimuli, comprising only female individuals and only two identities per age category. Thus, the findings cannot be readily generalized to broader human populations. It is also unknown whether dogs can discriminate finer distinctions in human age, such as infants versus older children or middle-aged versus older adults. Thus, we did not explore whether dogs perceive changes in human age as a continuous quantity. Future studies using a larger and more diverse set of stimuli are needed to determine whether dogs form more general cross-modal representations of human age.

Third, we used a binary measure of whether dogs had lived with a child as an indicator of 'experience with children'. Dogs coded as 'no experience' may nevertheless have had incidental exposure to children in daily life, such as on walks or in other social situations. Moreover, we did not take into account how long the dogs had cohabited with children. In future studies, the frequency, quality and duration of contact with children should be considered.

Finally, although owners were instructed to close their eyes and were unaware of the experimental conditions, they could hear the auditory stimuli. Therefore, subtle unintentional cueing cannot be entirely excluded. This possibility could be minimized in future studies by requiring owners to wear sound-attenuating headphones.

In summary, our findings provide tentative evidence that dogs may associate human voices with corresponding facial age categories. However, this effect was weak at the group level and should be interpreted with caution, particularly given the limited number of stimuli used in the study. The results suggest that dogs' sensitivity to human age-related cues may be influenced by individual experience, although this effect was not robust across all analyses.

Author Contributions

Hitomi Chijiwa: Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.
Yasuhiro Kanakogi: Writing – review & editing, Validation, Supervision, Software, Methodology, Conceptualization.

Data Availability

Data and R script are available as part of the Supplementary Material.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the authors used OpenAI's ChatGPT in order to improve language and readability. No content was generated or altered beyond language editing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of Interest

The authors have no competing interests to declare.

Acknowledgments

This study was financially supported by Grants-in-Aid for Scientific Research (KAKENHI) no. 24KJ0150 to H. Chijiwa from the Japan Society for the Promotion of Science (JSPS). We thank all the dogs and their owners who participated in this study. We also wish

to thank S. Takagi for coding the video data and H. Yamamoto for providing valuable advice on statistical analyses.

Supplemental Material

Supplementary material associated with this article is available at: <https://doi.org/10.1016/j.anbehav.2026.123710>.

The Supplementary Material has not been copy-edited and the content, formatting and file functionality are the sole responsibility of the author(s) from the copy-edited files when preparing the proof.

References

- Adachi, I., & Fujita, K. (2007). Cross-modal representation of human caretakers in squirrel monkeys. *Behavioural Processes*, 74(1), 27–32. <https://doi.org/10.1016/j.beproc.2006.09.004>
- Adachi, I., Kuwahata, H., & Fujita, K. (2007). Dogs recall their owner's face upon hearing the owner's voice. *Animal Cognition*, 10(1), 17–21. <https://doi.org/10.1007/s10071-006-0025-8>
- Albuquerque, N., Guo, K., Wilkinson, A., Savalli, C., Otta, E., & Mills, D. (2016). Dogs recognize dog and human emotions. *Biology Letters*, 12(1), Article 20150883. <https://doi.org/10.1098/rsbl.2015.0883>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, 11(12), 535–543. <https://doi.org/10.1016/j.tics.2007.10.001>
- Forman, J., Rowe, J. S., & Leavens, D. A. (2025). Object permanence in domestic cats (*Felis catus*) using violation-of-expectancy by owner and stranger. *PLoS One*, 20(7), Article e0312225. <https://doi.org/10.1371/journal.pone.0312225>
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage <https://www.john-fox.ca/Companion>.
- Gillfillan, G., Vitale, J., McNutt, J. W., & McComb, K. (2016). Cross-modal individual recognition in wild African lions. *Biology Letters*, 12(8), Article 20160323. <https://doi.org/10.1098/rsbl.2016.0323>
- Giraudet, C. S. E., Liu, K., McElligott, A. G., & Cobb, M. (2022). Are children and dogs best friends? A scoping review to explore the positive and negative effects of child-dog interactions. *PeerJ*, 10, Article e14532. <https://doi.org/10.7717/peerj.14532>
- Hare, B., Brown, M., Williamson, C., & Tomasello, M. (2002). The domestication of social cognition in dogs. *Science*, 298(5598), 1634–1636. <https://doi.org/10.1126/science.1072702>
- Hare, B., & Tomasello, M. (2005). Human-like social skills in dogs? *Trends in Cognitive Sciences*, 9(9), 439–444. <https://doi.org/10.1016/j.tics.2005.07.003>
- Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, 5, 69–95.
- Jardat, P., Ringhofer, M., Yamamoto, S., Gouyet, C., Degrande, R., Parias, C., Reigner, F., Calandreau, L., & Lansade, L. (2023). Horses form cross-modal representations of adults and children. *Animal Cognition*, 26(2), 369–377. <https://doi.org/10.1007/s10071-022-01667-9>
- Kawaguchi, Y., Virányi, Z., Faragó, T., Huber, L., & Völter, C. J. (2024). Cross-modal perception of puppies and adult conspecifics in dogs (*Canis familiaris*). *Journal of Comparative Psychology*, 138(4), 246–258. <https://doi.org/10.1037/com0000385>
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3), 983–997. <https://doi.org/10.2307/2533558>
- Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One*, 7(5), Article e36399. <https://doi.org/10.1371/journal.pone.0036399>
- Kojima, S., Izumi, A., & Ceugniet, M. (2003). Identification of vocalizers by pant hoots, pant grunts and screams in a chimpanzee. *Primates*, 44(3), 225–230. <https://doi.org/10.1007/s10329-002-0014-8>
- Kondo, N., Izawa, E. I., & Watanabe, S. (2012). Crows cross-modally recognize group members but not non-group members. *Proceedings of the Royal Society B: Biological Sciences*, 279(1735), 1937–1942. <https://doi.org/10.1098/rspb.2011.2419>
- Korzeniowska, A. T., Simner, J., Root-Gutteridge, H., & Reby, D. (2022). High-pitch sounds small for domestic dogs: Abstract crossmodal correspondences between auditory pitch and visual size. *Royal Society Open Science*, 9(2), Article 211647. <https://doi.org/10.1098/rsos.211647>
- Marshall-Pescini, S., Cafazzo, S., Virányi, Z., & Range, F. (2017). Integrating social ecology in explanations of wolf–dog behavioral differences. *Current Opinion in Behavioral Sciences*, 16, 80–86. <https://doi.org/10.1016/j.cobeha.2017.05.002>
- Marshall-Pescini, S., Schwarz, J. F. L., Kostelnik, I., Virányi, Z., & Range, F. (2017). Importance of a species' socioecology: Wolves outperform dogs in a conspecific cooperation task. *Proceedings of the National Academy of Sciences*, 114(44), 11793–11798. <https://doi.org/10.1073/pnas.1709027114>
- Miklósi, Á., Polgárdi, R., Topál, J., & Csányi, V. (1998). Use of experimenter-given cues in dogs. *Animal Cognition*, 1(2), 113–121. <https://doi.org/10.1007/s100710050016>
- Pitcher, B. J., Briefer, E. F., Baciadonna, L., & McElligott, A. G. (2017). Cross-modal recognition of familiar conspecifics in goats. *Royal Society Open Science*, 4(2), Article 160346. <https://doi.org/10.6084/m9.figshare.c.3679390>
- Powell, L. J., & Spelke, E. S. (2018). Human infants' understanding of social imitation: Inferences of affiliation from third party observations. *Cognition*, 170, 31–48. <https://doi.org/10.1016/j.cognition.2017.09.0079>
- Proops, L., & McComb, K. (2012). Cross-modal individual recognition in domestic horses (*Equus caballus*) extends to familiar humans. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741), 3131–3138. <https://doi.org/10.1098/rspb.2012.0626>
- Proops, L., McComb, K., & Reby, D. (2009). Cross-modal individual recognition in domestic horses (*Equus caballus*). *Proceedings of the National Academy of Sciences*, 106(3), 947–951. <https://doi.org/10.1073/pnas.0809127105>
- Quaranta, A., d'Ingeo, S., Amoroso, R., & Siniscalchi, M. (2020). Emotion recognition in cats. *Animals*, 10(7), 1107. <https://doi.org/10.3390/ani10071107>
- R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>.
- Ratcliffe, V. F., McComb, K., & Reby, D. (2014). Cross-modal discrimination of human gender by domestic dogs. *Animal Behaviour*, 91, 127–135. <https://doi.org/10.1016/j.anbehav.2014.03.009>
- Sliwa, J., Duhamel, J. R., Pascalis, O., & Wirth, S. (2011). Spontaneous voice–face identity matching by rhesus monkeys for familiar conspecifics and humans. *Proceedings of the National Academy of Sciences*, 108(4), 1735–1740. <https://doi.org/10.1073/pnas.1008169108>
- Soproni, K., Miklósi, Á., Topál, J., & Csányi, V. (2001). Comprehension of human communicative signs in pet dogs (*Canis familiaris*). *Journal of Comparative Psychology*, 115(2), 122–126. <https://doi.org/10.1037/0735-7036.115.2.122>
- Stathopoulos, E. T., Huber, J. E., & Sussman, J. E. (2011). Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age. *Journal of Speech, Language, and Hearing Research*, 54(4), 1011–1021. [https://doi.org/10.1044/1092-4388\(2010\)10-0036](https://doi.org/10.1044/1092-4388(2010)10-0036)
- Surányi, K., Gábor, A., Somogyi, R., Gyenes, B., Morvai, B., Boros, M., Faragó, T., & Andics, A. (2025). Individual level recognition of familiar human speakers in dogs. *Animal Behaviour*, 219, Article 123016. <https://doi.org/10.1016/j.anbehav.2024.10.030>
- Téglás, E., Gergely, A., Kupán, K., Miklósi, Á., & Topál, J. (2012). Dogs' gaze following is tuned to human communicative signals. *Current Biology*, 22(3), 209–212. <https://doi.org/10.1016/j.cub.2011.12.018>
- Takagi, S., Arahori, M., Chijiwa, H., Saito, A., Kuroshima, H., & Fujita, K. (2019). Cats match voice and face: Cross-modal representation of humans in cats (*Felis catus*). *Animal Cognition*, 22(5), 901–906. <https://doi.org/10.1007/s10071-019-01265-2>
- Trösch, M., Cuzol, F., Parias, C., Calandreau, L., Nowak, R., & Lansade, L. (2019). Horses categorize human emotions cross-modally based on facial expression and non-verbal vocalizations. *Animals*, 9(11), 862. <https://doi.org/10.3390/ani9110862>
- Udell, M. A. R., Dorey, N. R., & Wynne, C. D. L. (2008). Wolves outperform dogs in following human social cues. *Animal Behaviour*, 76(6), 1767–1773. <https://doi.org/10.1016/j.anbehav.2008.07.028>
- Udell, M. A. R., Dorey, N. R., & Wynne, C. D. L. (2010). What did domestication do to dogs? A new account of dogs' sensitivity to human actions. *Biological Reviews*, 85(2), 327–345. <https://doi.org/10.1111/j.1469-185X.2009.00104.x>
- Wanser, S. H., MacDonald, M., & Udell, M. A. R. (2021). Dog–human behavioral synchronization: Family dogs synchronize their behavior with child family members. *Animal Cognition*, 24(4), 747–752. <https://doi.org/10.1007/s10071-020-01454-4>
- Wanser, S. H., Simpson, A. C., MacDonald, M., & Udell, M. A. R. (2020). Considering family dog attachment bonds: Do dog–parent attachments predict dog–child attachment outcomes in animal-assisted interventions? *Frontiers in Psychology*, 11, Article 566910. <https://doi.org/10.3389/fpsyg.2020.566910>
- Yong, M. H., & Ruffman, T. (2014). Emotional contagion: Dogs and humans show a similar physiological response to human infant crying. *Behavioural Processes*, 108, 155–165. <https://doi.org/10.1016/j.beproc.2014.10.006>