Five-month-old infants' identification of the sources of vocalizations

Athena Vouloumanos[a,1], Madelynn J. Druhen[b], Marc D. Hauser[c], and Anouk T. Huizink[d]

[a]Department of Psychology, New York University, 6 Washington Place, New York, NY 10003; [b]Department of Psychology, University of North Carolina, 1000 Spring Garden Street, Greensboro, NC 27403; [c]Departments of Psychology and Human Evolutionary Biology, Harvard University, Cambridge, MA 02138; and [d]Department of Psychology, McGill University, 1205 Docteur Penfield, Montréal, QC, Canada H3A 1B1

Humans speak, monkeys grunt, and ducks quack. How do we come to know which vocalizations animals produce? Here we explore this question by asking whether young infants expect humans, but not other animals, to produce speech, and further, whether infants have similarly restricted expectations about the sources of vocalizations produced by other species. Five-month-old infants matched speech, but not human nonspeech vocalizations, specifically to humans, looking longer at static human faces when human speech was played than when either rhesus monkey or duck calls were played. They also matched monkey calls to monkey faces, looking longer at static rhesus monkey faces when rhesus monkey calls were played than when either human speech or duck calls were played. However, infants failed to match duck vocalizations to duck faces, even though infants likely have more experience with ducks than monkeys. Results show that by 5 months of age, human infants generate expectations about the sources of some vocalizations, mapping human faces to speech and rhesus faces to rhesus calls. Infants' matching capacity does not appear to be based on a simple associative mechanism or restricted to their specific experiences. We discuss these findings in terms of how infants may achieve such competence, as well as its specificity and relevance to acquiring language.

cognitive development | language acquisition | speech perception | conspecific | evolution

Humans' intuitive understanding of physical aspects of the world, such as object identity, continuity, and motion, has its roots in infancy. Carey (1) proposed that children's intuitive understanding of biology is built on their intuitive understanding of psychology, beginning with an understanding of human behavior which is then generalized along a similarity gradient to other animals. Others, such as Keil (2) and Hatano (3), have argued that biology is its own privileged domain of knowledge with dedicated inferential mechanisms. Whichever account is correct, an understanding of the biology of different species requires an understanding of differing properties of those species. Many studies (e.g., 4–6) have shown that infants use visible properties to distinguish between instances of different categories (e.g., dogs vs. cats), but ultimately, categories function to support inferences about unseen properties (7, 8). We investigate what infants might know or infer about the kinds of vocalizations that different species produce.

The vocalizations that animals produce are often species specific: distinctive and functional in providing information about individual and group identity, motivational state, location, and often the triggering contexts (9–13). At the same time, vocalizations contain information about the vocal tract that produced them, effectively functioning as a species-specific signature (14–16). The recognition of vocalizations is especially important because they can help identify and localize conspecifics, predators, and prey, even when those organisms may not be within the line of sight. In the case of humans, spoken language represents the most fundamental vehicle for transmitting ideas and for identifying with a particular culture (8). Speech

is a rich auditory signal modulated in temporal and spectral aspects by precise manipulation of the vocal folds (the source) and the supralaryngeal articulators (the filter), and stored as abstract representations in the form of distinctive features (17, 18). These properties help to distinguish speech from nonspeech human vocalizations and from vocalizations produced by nonhuman animals.

We examine whether human infants have expectations about the nature of vocalizations generated by humans, and in particular, whether they expect humans to produce certain vocalizations, while also expecting other species to produce their own unique vocalizations. More specifically, we began by asking whether infants expected human faces to be associated with human speech, but rhesus monkey faces to be associated with rhesus monkey vocalizations. Humans and rhesus monkeys diverged an estimated 25–30 million years ago (19) and share an estimated 92%–95% of their genetic sequence, leading to significant similarities in facial (20) and vocal characteristics. For example, like human speech, some rhesus vocalizations contain distinct formant frequencies, produced by changing lip configurations and mandibular positions (21–23). Because the vocal folds of rhesus monkeys are lighter than those of humans, rhesus vocalizations encompass a broader range of frequencies than speech. However, rhesus monkeys, like other nonhuman primates, lack the oropharyngeal chamber and thick, curved tongue that human adults use to modulate formant frequencies (21). As a result, rhesus monkeys use a more limited "phonetic" repertoire, although still larger and more varied than many other species with similar vocal tracts (24).

Our specific focus was on 5-month-old human infants. At 5 months, infants are able to participate in sequential looking tasks (e.g., 25, 26), and to match faces with voices on the basis of age or maturity (27), segmental information (28–30), and emotion (for their primary caregiver) (31, 32). However, it is not until after the age of 5 months that many of the effects of experience-dependent learning [e.g., tuning to conspecific faces (33, 34) and to auditory, visual, and multisensory properties of their native language (35–40)] are observed. Although much is known about infants' abilities to match properties of human voices to faces, it is unknown whether infants are able to match vocalizations to the specific species that produces them.

We examined whether 5-month-old infants generate species-specific expectations about the source of vocalizations. We first asked whether infants expect that humans, and not other animals, produce spoken language by testing whether infants match human speech—but not rhesus monkey calls—specifically to

PSYCHOLOGICAL AND COGNITIVE SCIENCES
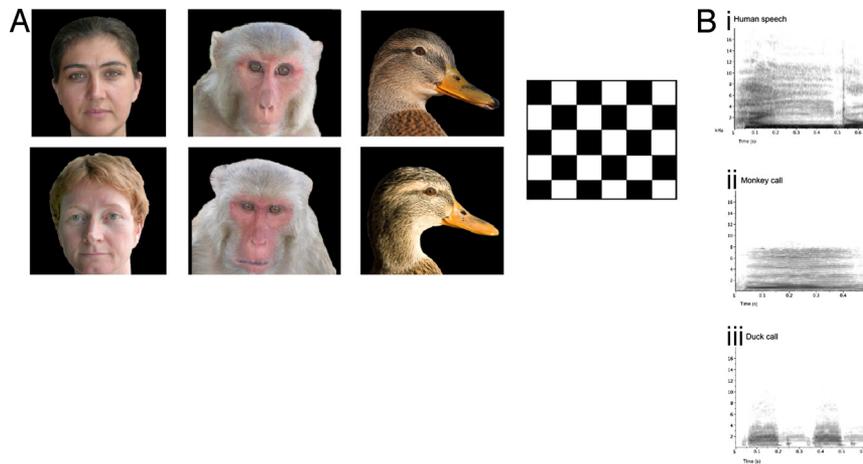
Fig. 1. (*A*) Visual stimuli presented in experiments 1–4. Human and monkey faces were presented individually, paired with either human or monkey vocalizations in experiment 1, and with human, monkey, or duck vocalizations in experiment 2. Human and duck faces were presented individually with human, monkey, or duck vocalizations in experiment 3. A black and white checkerboard was used in experiment 4. (*B*) Auditory stimuli presented in experiments 1, 2, 3, and 4. Human speech (I) was contrasted with rhesus monkey calls (II) in experiment 1. Experiments 2, 3, and 4 included duck calls (III).

human faces as opposed to rhesus monkey faces. We also compared infants' matching of speech to matching of human nonspeech vocalizations to determine whether the process is restricted to speech or extends more generally to all human-generated sounds. Second, we examined whether infants can match voice to face in other species that are visually and acoustically unfamiliar to infants. Specifically, we asked whether infants match rhesus monkey faces with rhesus monkey vocalizations, and duck faces with duck vocalizations, when presented with human speech, monkey calls, or duck calls.

Experiment 1. Twelve 5-month-old infants from English- and French-speaking homes were shown a sequence of individually presented static human faces and rhesus monkey faces (see Fig. 1) paired either with human speech or with rhesus monkey vocalizations. We examined whether infants preferentially attended to the human faces when human vocalizations were presented (2 Japanese single words, *nasu* and *haiiro*), and whether infants preferentially attended to the rhesus monkey faces when rhesus monkey vocalizations (a coo and a gekker call) were presented.

Pilot data using a typical cross-modal matching procedure (e.g., 27, 34, 41), in which 2 faces were presented side by side in tandem with one sound, revealed an overwhelming preference

for human faces that obscured all other potential effects. We therefore adapted the procedure to present single faces sequentially, in tandem with different vocalizations, and compared infant looking time for matched face-voice pairs to mismatched face-voice pairs. Previous research shows that infants successfully extract matching information between auditory and visual stimuli from sequential presentations of stimuli (42, 43). When presented with audiovisual stimuli, infants tend to look longer at matching displays (27, 34, 44–46). We predicted that if infants identified the sources of vocalizations, they would look longer when the vocalizations and faces matched. Infant looking times to trials in which voices and faces matched were compared with looking times to trials with mismatched faces and voices.

Results

Condition 1: Matching Speech to Human Faces. Five-month-old infants looked longer at static human faces when human speech was presented and also looked longer at static rhesus monkey faces when rhesus vocalizations were presented (see Fig. 2). We carried out analyses of variance (ANOVAs) with $P < 0.05$ set as statistical significance, as well as nonparametric tests to confirm the general pattern within the sample of infants. A 2 (congruence: match, mismatch) $\times$ 2 (block: first, second) ANOVA yielded a significant congruence by block interaction $[F(1, 11) =$

Fig. 2. Mean proportion looking time for trials in which the presented faces either matched (dark bars) or did not match (light bars) the vocalizations played for experiments 1–4. Chance performance is in red. Error bars represent standard error.

9.48, $P = 0.01$]. Because a preliminary ANOVA revealed no effects or interactions of sex (F, M), order (human first, monkey first), and pairing (face A-sound A, face A-sound B) ($P > 0.05$), we collapsed across these factors in all analyses. Because in previous cross-modal studies 4- to 5-month-olds did not demonstrate matching of faces to voices until the second presentation block (27), we tested for a block effect and focused our analyses on the second block. Planned comparisons revealed no difference in proportion looking time (PLT) between the match ($M = 0.504$, SE = 0.027) and mismatch ($M = 0.496$, SE = 0.027) in the first block, but a significant difference in PLT for the match ($M = 0.580$, SE = 0.024) and mismatch ($M = 0.420$, SE = 0.024) in the second block [$t(11) = 3.32$, $P = 0.007$]. Consistent with previous cross-modal matching studies in young infants (e.g., 27), matching behavior emerged in the second block. We therefore submitted the data from this second block to subsequent analyses. To examine whether infants succeeded in matching both speech to human faces and monkey calls to monkey faces, we performed one-way ANOVAs and found a significant effect of congruence for both human stimuli [$M_{match} = 0.604$, SE = 0.045, $F(1, 22) = 10.57$, $P = 0.004$] and monkey stimuli [$M_{match} = 0.557$, SE = 0.029, $F(1, 22) = 7.76$, $P = 0.011$]. Finally, 10 of 12 infants looked longer toward the matching display (binomial test, $P < 0.039$). Thus, by 5 months, infants match human speech to human faces, and also match rhesus monkey vocalizations to rhesus monkey faces.

Condition 2: Matching Human Nonspeech Vocalizations to Human Faces. To determine whether the matching results in condition 1 were mediated by speech in particular or human-generated voices more generally, we ran a second condition with human nonspeech vocalizations (e.g., laughter). Infants looked equally at human and monkey faces when either human nonspeech vocalizations or monkey vocalizations were played (see Fig. 2). A 2 (congruence: match, mismatch) × 2 (block: first, second) ANOVA yielded no significant effects and no interactions ($P > 0.05$) on PLT. PLT between the match ($M = 0.504$, SE = 0.019) and mismatch ($M = 0.496$, SE = 0.019) was equivalent in both blocks: block 1 match ($M = 0.506$, SE = 0.025), block 1 mismatch ($M = 0.494$, SE = 0.025), block 2 match ($M = 0.502$, SE = 0.025), block 2 mismatch ($M = 0.498$, SE = 0.025). A repeated-measures mixed ANOVA with congruence (match, mismatch) as a within-subject factor, and sex, order, or pairing as between-subject factors, also revealed no other effects or interactions ($P > 0.05$). Finally, an equivalent number of infants looked longer to the match ($n = 7$) and the mismatch ($n = 5$) (binomial test, $P > 0.05$).

An explicit comparison of conditions 1 and 2 revealed a significant interaction between condition and congruence in the second block [$F(1, 22) = 5.13$, $P = 0.034$], demonstrating that infants' matching performance in condition 1 was reliably better than in condition 2. Thus, infants matched spoken language, but not human nonspeech vocalizations, to human faces.

Experiment 2. Did infants in experiment 1 specifically match rhesus monkey calls to rhesus monkey faces, or did they simply match nonhuman faces with nonhuman vocalizations, effectively matching through a type of disjunctive syllogism [i.e., the unfamiliar sound (monkey call) must match either face A (human) or face B (monkey); if it doesn't match face A (human), then it must match face B (monkey)]? To further test infants' matching of monkey faces with voices, we conducted a second experiment in which the same static human and monkey faces were shown to 5-month-old infants, but this time in tandem with 3 types of vocalizations: human speech, monkey calls, and duck calls. If infants match nonhuman faces with nonhuman voices, then the calls of ducks and monkey should be considered equally likely matches for the monkey faces. If, however, infants generate expectations about monkey vocalizations, then they might be able to match the monkey faces specifically to the monkey calls,

or more modestly, reject duck calls in favor of monkey calls. Fifteen 5-month-old infants were tested using the same infant-controlled looking procedure as in experiment 1, in which they viewed a series of trials composed of a static human or monkey face, but now these faces were presented in tandem with a human speech sound, a monkey call, or a duck call.

Results

In this 3-way matching task, 5-month-old infants once again looked longer at the human face when it was presented with human speech, as compared with monkey or duck vocalizations. They also looked longer at the monkey face when it was presented with monkey vocalizations, as compared with human or duck vocalizations (see Fig. 2). A 2 (congruence: match, mismatch) × 2 (block: first, second) ANOVA yielded a significant effect of congruence [$F(2, 28) = 7.23$, $P = 0.003$, $p_{rep} = 0.97$] with longer PLT to match ($M = 0.406$, SE = 0.020) than the primate mismatch ($M = 0.316$, SE = 0.023) or the duck mismatch ($M = 0.278$, SE = 0.016). Based on our findings from experiment 1 in which matching was only evident in block 2, we analyzed block 1 and block 2 separately. In these separate analyses, as in experiment 1 and previous studies (27), only block 2 showed a significant effect of congruence with longer PLT to match ($M = 0.438$, SE = 0.034) than the primate mismatch ($M = 0.303$, SE = 0.028) or the duck mismatch [$M = 0.259$, SE = 0.024, $F(2, 28) = 6.99$, $P = 0.003$, $p_{rep} = 0.97$]. Eleven of 15 infants looked longer toward the match than either of the mismatches in block 2 (binomial test, $P < 0.002$). The data from block 2 were submitted to subsequent analyses.

To examine whether infants succeeded in matching both the monkey faces and the human faces to the congruent vocalizations, separate one-way ANOVAs revealed a significant effect of congruence for human stimuli [$F(1, 22) = 10.99$, $P < 0.001$], with PLT to match ($M = 0.466$, SE = 0.040) higher than primate mismatch [$M = 0.305$, SE = 0.039, least squares difference (LSD), $P < 0.003$] and higher than duck mismatch ($M = 0.230$, SE = 0.028, LSD, $P < 0.001$). There was also a significant effect of congruence for monkey stimuli [$F(1, 22) = 3.71$, $P = 0.033$] with PLT to match ($M_{match} = 0.409$, SE = 0.038) higher than the human mismatch ($M = 0.304$, SE = 0.030, LSD, $P < 0.036$) and higher than the duck mismatch ($M = 0.287$, SE = 0.034, LSD, $P = 0.016$). This experiment confirmed infants' ability to match human faces specifically to speech, replicating and extending the findings of experiment 1. Crucially, infants matched monkey faces preferentially to monkey calls, perceiving rhesus monkey calls as a better match for rhesus monkey faces than either human speech or duck calls.

Experiment 3. In experiment 2 we showed that infants matched rhesus monkey vocalizations to rhesus monkey faces, rejecting duck calls as a match for human faces or for rhesus monkey faces. One explanation for these findings is that infants were able to match monkey vocalizations and faces based on their prior experience with ducks. That is, infants might have been able to reject the monkey face as a possible match for the duck vocalization based on having previously associated the duck vocalizations with duck faces. To consider infants' familiarity with ducks, and to further examine the specificity of infants' capacity to match faces and vocalizations, we tested whether infants match duck vocalizations to duck faces. We presented fifteen 5-month-old infants with exactly the same acoustic stimuli—human speech, monkey calls, or duck calls—as in experiment 2 but now in tandem with the faces of humans and ducks. If infants are able to match duck vocalizations to duck faces, then this capacity may have contributed to infants' matching of monkeys in the previous experiment. If infants are unable to match duck vocalizations to duck faces, then previous experience with ducks

cannot account for infants' matching of monkey faces and vocalizations.

Results

In this 3-way matching task with human and duck faces, 5-month-old infants did not look systematically longer at the duck face when it was presented with a duck vocalization. Instead, infants looked longest during human speech trials, regardless of which face was presented (see Fig. 2). A 2 (congruence: match, monkey mismatch, other mismatch) × 2 (block: first, second) ANOVA yielded no significant effect of congruence, block, and no interactions (P = n.s.), reflecting equal PLT to match (M = 0.328, SE = 0.021), monkey mismatch (M = 0.320, SE = 0.020), or other species mismatch (M = 0.352, SE = 0.017). Three of 15 infants looked longer toward the match than either of the mismatches in block 2 (binomial test, P = n.s.).

To examine whether infants' performance was equivalent for the duck faces and the human faces, we performed an additional repeated-measures ANOVA examining the PLT for congruent pairs (congruence: match, monkey mismatch, other mismatch) for each of the 2 species (human, duck) combined across blocks. There was a significant interaction between congruence and species [$F(2, 28)$ = 3.478, P = 0.045], which was driven by infants looking longer during speech trials in the presence of the human faces (M = 0.385, SE = 0.028) and also looking longer during speech trials in the presence of the duck faces (M = 0.401, SE = 0.028). Post hoc tests confirmed that PLT was longer overall for human speech trials (M = 0.393, SE = 0.020) than for duck vocalizations [M = 0.287, SE = 0.022, $t(14)$ = 2.690, P = 0.018] and for monkey vocalizations [M = 0.320, SE = 0.20, $t(14)$ = 2.054, P = 0.059]. Infants thus looked longer overall during the presentation of speech, both for human faces and for duck faces, and showed no evidence of matching duck calls to duck faces.

Experiment 4. To control for the possibility that infants' failure to match duck faces to duck vocalizations in experiment 3 stems from an intrinsic aversion to duck calls, we presented exactly the same sounds in the same order as experiment 2 and 3, but instead of presenting the sounds simultaneously with a face, we presented them with a black and white checkerboard display. Because previous studies have shown that young infants prefer speech to other sounds (26, 47, 48), we expected infants to listen longer overall to speech. The crucial test is, therefore, a comparison of the duck calls and the monkey calls. If infants listen less to duck calls than to monkey calls, then the pattern infants demonstrated in experiments 2 and 3 could be explained by an aversion to duck calls, rather than matching behavior per se. If, however, infants listened equally to duck and monkey calls, then the data pattern is best explained by an ability to match rhesus monkey calls to rhesus monkey faces, a capacity that does not extend to matching duck calls to duck faces. We therefore presented twelve 5-month-old infants with the same 3 sound types as in experiments 2 and 3, but with an identical black-and-white checkerboard.

Results

In this test of inherent listening preferences, infants listened equally to monkey and duck calls (see Fig. 2). Infants' overall looking time was analyzed in an ANOVA with sound type (human speech, monkey calls, or duck calls) as the relevant factor. A significant difference in looking time to the 3 sounds was found [$F(2, 22)$ = 5.35, P = 0.013]. Planned comparisons combined across blocks revealed that infants listened proportionally longer to speech (M = 0.393, SE = 0.024) as compared with monkey calls [M = 0.314, SE = 0.015; $t(11)$ = 2.17, P = 0.053] and with duck calls [M = 0.293, SE = 0.016; $t(11)$ = 2.67, P = 0.022] (see Fig. 2). This overall preference for speech was expected, and consistent with previous studies (26, 49). How-

ever, and crucially for the interpretation of experiments 2 and 3, infants listened to monkey calls and duck calls equally [$t(11)$ = 1.06, P = n.s.]. Thus, when tested in a sound preference procedure in the absence of any faces, infants showed no difference in listening time to duck calls and monkey calls. The equal listening times between duck and monkey calls suggest that duck calls and monkey calls were equally salient for infants, and that the difference in looking patterns between monkeys and ducks observed in experiments 2 and 3 is robust.

Discussion

Results from four experiments show that human infants have some expectations about the sources of vocalizations, both for humans and for another primate species. Five-month-old infants matched speech preferentially to human faces, and matched rhesus monkey calls preferentially to rhesus monkey faces. These findings license three conclusions.

First, despite the lack of exposure to rhesus monkey faces and vocalizations, at 5 months, infants do not consider rhesus monkeys to be a likely source for speech: they expect humans, and not monkeys, to produce speech. Although this success might be due to infants' familiarity with human vocalizations, infants were able to match the unfamiliar Japanese speech while failing to match the more familiar nonspeech vocalizations of agreement and laughter. This suggests either an early tuning for some functional properties of speech as opposed to other sounds (48), or that the acoustic properties of human nonspeech sounds provide insufficient information about the vocal source of such sounds. According to this latter possibility, infants may use segmental variations in periodicity and perturbation that are generated by the precise manipulation of the vocal folds and the supralaryngeal articulators as diagnostic of sounds generated by a human vocal apparatus (17). The attenuated segmental variation in human nonspeech vocalizations—relative to speech—may have hindered infants' ability to identify them as human utterances. It remains to be explored whether infants would match speech sounds also used in nonspeech contexts, such as fricatives or clicks (50).

Second, infants did not match the rhesus monkey calls to human faces. In 5 months of postnatal exposure, infants have heard humans emit a wide variety of sounds, including speech, sneezes, and perhaps even some animal sounds associated with story books. It is unlikely that they have experienced every possible human vocalization and yet they have determined that humans are not a likely source of rhesus monkey calls, despite their biological, communicative, and spectrally rich properties (9). At the same time, human nonspeech vocalizations that have the same vocal source and encompass the same frequency range as speech but that are limited in phonetic repertoire were not matched with human faces. This may be because infants are not sensitive to the categorical differences between nonspeech human vocalizations and rhesus monkey calls. Although it is still unclear which specific cues infants use to identify human and rhesus monkey vocalizations, the presence of a broad phonetic repertoire might provide a cue to a human vocal source.

Third, with little or no prior exposure to rhesus monkey faces or voices, human infants match monkey faces preferentially to monkey calls, and not duck calls or human speech. One possibility is that infants were relying on previous experience with ducks. This explanation, however, is undermined by the findings from experiments 3 and 4 in which infants failed to match duck vocalizations to duck faces, and attended equally to duck and monkey calls in the absence of a face. Infants' matching of rhesus monkey calls was not based on their capacity to rule out duck vocalizations based on their prior experience. Further, if prior experience was driving these effects, then, if anything, infants should have shown greater competence in matching duck calls to duck faces than monkey calls to monkey faces.

How would naïve human infants generate a specific expectation about the likely sounds produced by rhesus monkeys? Even static neutral faces such as those used in the current study contain many cues about plausible vocalizations (51). Infants may use attributes such as ratios between facial features (27), in combination with the acoustic properties of the calls, to determine that monkey faces are more likely to produce monkey calls than either human speech or duck calls. Young infants show remarkable abilities in linking specific properties in human faces and voices (52, 53), even for aspects of bimodal matching that are unlikely to be learned, such as vowel matching (28, 30). Infants also readily detect correspondences between specific monkey calls and *dynamic* monkey faces that match (34). Why then did infants fail to match the calls and faces of ducks? Human and monkey faces share many features that may function to identify them as faces, such as the placement of the eyes and mouth in the same plane. Ducks, in contrast, have a differently shaped oral region and their eyes do not fall within the same plane. These featural differences might make it more difficult for infants to identify duck faces as faces, and as a result, reduce their capacity to identify the source of duck vocalizations.

A complementary hypothesis is that infants might succeed by making explicit analogies with their knowledge of humans. Infants might notice structural similarities between human and monkey faces, and between human and monkey vocalizations, and then extrapolate the human matching properties to monkeys (1, 54). Indeed, in Carey's (1) studies of biological knowledge, children's attribution of animal properties depended on their making inductive inferences based on similarity judgments between what they knew—familiar people—and unfamiliar organisms. According to this view, infants may generate a similarity gradient originating from humans: by looking more like human faces than say insect faces do, rhesus monkey faces are more likely to produce vocalizations that are, in important ways, similar to human sounds (e.g., in containing formant structure generated by supralaryngeal articulation). Similarly, studies of adult categorical perception of monkey and human faces suggest that human adults may process monkey faces in terms of characteristics they share with human faces (54). Although Carey studied older children, and categorical perception studies tested adults, the same general principles might be at work in early infancy. Although infants in our study were able to reject duck calls as a possible match for rhesus monkeys, they would undoubtedly have more difficulty rejecting the more similar calls of other nonhuman primates. Because duck faces (and vocalizations) share fewer characteristics with humans, infants may not have been able to analyze them in the same way to match them.

Whether by their existing knowledge of invariant cross-modal relations, or their use of inductive processes and similarity gradients, infants were able to match human speech to humans, and rhesus monkey calls to rhesus monkey faces. This finding provides the first evidence that, from a young age, human infants are able to make correct inferences about some correspondences between different kinds of vocalizations and different species. This ability may help infants identify conspecifics even when they are not visible, and allow them to identify the human-produced speech sounds that are relevant for language acquisition.

Materials and Methods

Participants. Infants ranged in age from 4 months, 17 days, to 5 months, 17 days, and were from predominantly English- and French-speaking families. No infant had prior exposure to Japanese. Experiment 1 tested 12 infants in each of 2 conditions ($M = 5$ months, 0 days), experiments 2 and 3 tested 15 infants ($M = 4$ months, 30 days and $M = 5$ months, 5 days, respectively), and experiment 4 tested 12 infants ($M = 5$ months, 4 days). Complete participant and procedure information is reported in *SI Text*.

Stimuli. The auditory stimuli consisted of 2 Japanese words (*nasu* and *haiiro*), 2 nonspeech vocalizations ["mmm" (agreement) and "haha" (laughter)], 2 rhesus monkey vocalizations (a coo call associated with food and group movement, and a gekker call used in affiliative interactions), and 2 duck calls (see Fig. 1). Each sound file began with 1 s of silence and contained 13 repetitions of that particular sound, with a stimulus onset asynchrony of 3 s. Stimulus properties are reported in (Tables S1 and S2).

Visual stimuli consisted of a checkerboard (3-cm squares) and static color photographs of 2 adult female humans, 2 adult female rhesus monkeys, and 2 adult ducks (all 20 cm and 900 pixels in height; see Fig. 1). Photographs were cropped and placed on a black background using Adobe Photoshop.

Apparatus and Procedure. Infants sat on a parent's lap in front of a cinema display monitor in a sound-attenuated testing room. Sounds played from 2 speakers placed on either side of the monitor. An experimenter seated outside the testing area observed the child over a closed circuit and controlled the presentation of stimuli using Habit X software (Leslie Cohen, University of Texas at Austin). Infant reactions were videotaped. Parents wore headphones playing masking music.

The study design consisted of 2 identical blocks of 8 trials (experiment 1) or 6 trials (experiments 2, 3, and 4) that contained equal numbers of face trials from each species tested. Within each species' face set, infants saw the faces presented with each of the vocalizations being contrasted. Sounds and faces were alternated so that infants never heard the same vocalization or saw the same face consecutively. Sounds were presented at 65 ± 5 dB. In each experiment, the order of faces was counterbalanced, except for experiment 1, in which, due to a computer error, 7 of the infants saw a human first, and 5 saw a monkey first (order was not significant; see *Results*). For all experiments, half the infants saw a match first and half a mismatch. Infants terminated trials by looking away from the screen for more than 2 s or when trials reached the maximum length of 40 s. Infant looking-time was coded offline (reliability, $r > 0.991$) and was converted to proportion looking time (PLT), within each block. Specifically, for each set of faces within each block we derived a PLT for the relative amount of time infants looked at a given species' face when it was presented with a matching vocalization versus nonmatching vocalizations.

1. Carey S (1985) *Conceptual Change in Childhood* (MIT Press, Cambridge, MA).
2. Keil FC (1989) *Concepts, Kinds, and Cognitive Development* (MIT Press, Cambridge, MA).
3. Inagaki K, Hatano G (2002) *Young Children's Naïve Thinking About the Biological World* (Psychology Press, New York).
4. Quinn PC, Eimas PD, Rosenkrantz SL (1993) Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception* 22:463–475.
5. Baillargeon R, Wang SH (2002) Event categorization in infancy. *Trends Cogn Sci* 6:85–93.
6. Welder AN, Graham SA (2006) Infants' categorization of novel objects with more or less obvious features. *Cognit Psychol* 52:57–91.
7. Gelman SA (2003) *The Essential Child: Origins of Essentialism in Everyday Thought* (Oxford Univ Press, Oxford).
8. Kinzler KD, Dupoux E, Spelke ES (2007) The native language of social cognition. *Proc Natl Acad Sci USA* 104:12577–12580.
9. Hauser MD (2000) A primate dictionary? Decoding the function and meaning of another species' vocalizations. *Cog Sci* 24:445–475.
10. Owren MJ, Rendall D (2003) Salience of caller identity in rhesus monkey (Macaca mulatta) coos and screams: Perceptual experiments with human (Homo sapiens) listeners. *J Comp Psychol* 117:380–390.
11. Rendall D, Owren MJ, Rodman PS (1998) The role of vocal tract filtering in identity cueing in rhesus monkey (Macaca mulatta) vocalizations. *J Acoust Soc Am* 103:602–614.
12. Seyfarth RM, Cheney DL (2003) Signalers and receivers in animal communication. *Annu Rev Psychol* 54:145–173.
13. Ryan MJ, Rand AS (1995) Female responses to ancestral advertisement calls in Túngara frogs. *Science* 269:390–392.
14. Fitch WT, Reby D (2001) The descended larynx is not uniquely human. *Philos Trans R Soc Lond B Biol Sci* 268:1669–1675.
15. Fitch W, Hauser MD (1995) Vocal production in nonhuman primates: Acoustics, physiology, and functional constraints on "honest" advertisement. *Am J Primatol* 37:191–219.
16. Owren MJ, Seyfarth RM, Cheney DL (1997) The acoustic features of vowel-like grunt calls in chacma baboons (Papio cynocephalus ursinus): Implications for production processes and functions. *J Acoust Soc Am* 101:2951–2963.

17. Ladefoged P (1996) *Elements of Acoustic Phonetics* (Univ of Chicago Press, Chicago).
18. Chomsky N, Halle M (1968) *The Sound Pattern of English* (Harper & Row, New York).
19. Stewart CB, Disotell TR (1998) Primate evolution—in and out of Africa. *Curr Biol* 8:R582–R588.
20. Pascalis O, Petit O, Kim JH, Campbell R (2000) *Picture Perception in Primates: The Case of Face Perception* (Psychology Press, New York).
21. Hauser MD, Evans CS, Marler P (1993) The role of articulation in the production of rhesus monkey, Macaca mulatta, vocalizations. *Anim Behav* 45:423–433.
22. Fitch WT (1997) Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *J Acoust Soc Am* 102:1213–1222.
23. Hauser MD, Schon-Ybarra M (1994) The role of lip configuration in monkey vocalizations: Experiments using xylocaine as a nerve block. *Brain Lang* 46:232–244.
24. Fitch WT (2000) The evolution of speech: A comparative review. *Trends Cogn Sci* 4:258–267.
25. Cooper RP, Abraham J, Berman S, Staska M (1997) The development of infants' preference for motherese. *Infant Behav Dev* 20:477–488.
26. Vouloumanos A, Werker JF (2004) Tuned to the signal: The privileged status of speech for young infants. *Dev Sci* 7:270–276.
27. Bahrick LE, Netto D, Hernandez-Reif M (1998) Intermodal perception of adult and child faces and voices by infants. *Child Dev* 69:1263–1275.
28. Patterson ML, Werker JF (2003) Two-month-old infants match phonetic information in lips and voice. *Dev Sci* 6:191–196.
29. Kushnerenko E, Teinonen T, Volein A, Csibra G (2008) Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proc Natl Acad Sci USA* 105:11442–11445.
30. Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. *Science* 218:1138–1141.
31. Montague DR, Walker-Andrews AS (2002) Mothers, fathers, and infants: The role of person familiarity and parental involvement in infants' perception of emotion expressions. *Child Dev* 73:1339–1352.
32. Walker-Andrews AS (1997) Infants' perception of expressive behaviors: Differentiation of multimodal information. *Psychol Bull* 121:437–456.
33. Pascalis O, de Haan M, Nelson CA (2002) Is face processing species-specific during the first year of life? *Science* 296:1321–1323.
34. Lewkowicz DJ, Ghazanfar AA (2006) The decline of cross-species intersensory perception in human infants. *Proc Natl Acad Sci USA* 103:6771–6774.
35. Jusczyk PW, Luce PA, Charles-Luce J (1994) Infants' sensitivity to phonotactic patterns in the native language. *J Mem Lang* 33:630–645.
36. Pons F, Lewkowicz DJ, Soto-Faraco S, Sebastian-Galles N (2009) Narrowing of intersensory speech perception in infancy. *Proc Natl Acad Sci USA* 106:10598–10602.
37. Jusczyk PW, Cutler A, Redanz NJ (1993) Infants' preference for the predominant stress patterns of English words. *Child Dev* 64:675–687.
38. Weikum WM, et al. (2007) Visual language discrimination in infancy. *Science* 316:1159.
39. Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B (1992) Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255:606–608.
40. Werker JF, Tees RC (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav Dev* 7:49–63.
41. Jordan KE, Brannon EM (2006) The multisensory representation of number in infancy. *Proc Natl Acad Sci USA* 103:3486–3489.
42. Graf Estes K, Evans JL, Alibali MW, Saffran JR (2007) Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychol Sci* 18:254–260.
43. Brookes H, et al. (2001) Three-month-old infants learn arbitrary auditory-visual pairings between voices and faces. *Infant Child Dev* 10:75–82.
44. Patterson ML, Werker JF (2002) Infants' ability to match dynamic phonetic and gender information in the face and voice. *J Exp Child Psychol* 81:93–115.
45. Kuhl PK, Meltzoff AN (1984) The intermodal representation of speech in infants. *Infant Behav Dev* 7:361–381.
46. Guihou A, Vauclair J (2008) Intermodal matching of vision and audition in infancy: A proposal for a new taxonomy. *Eur J Dev Psychol* 5:68–91.
47. Vouloumanos A, Hauser MD, Werker JF, Martin A (2009) The tuning of human neonates' preference for speech. *Child Dev*, in press.
48. Vouloumanos A, Werker JF (2007) Listening to language at birth: Evidence for a bias for speech in neonates. *Dev Sci* 10:159–164.
49. Colombo JA, Bundy RS (1981) A method for the measurement of infant auditory selectivity. *Infant Behav Dev* 4:219–223.
50. Best CT, McRoberts GW, Goodell E (2001) Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *J Acoust Soc Am* 109:775–794.
51. Yehia H, Itakura F (1996) A method to combine acoustic and morphological constraints in the speech production inverse problem. *Speech Commun* 18:151–174.
52. Bahrick LE, Hernandez-Reif M, Flom R (2005) The development of infant learning about specific face-voice relations. *Dev Psychol* 41:541–552.
53. Teinonen T, Aslin RN, Alku P, Csibra G (2008) Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition* 108:850–855.
54. Campbell R, Pascalis O, Coleman M, Wallace SB, Benson PJ (1997) Are faces of different species perceived categorically by human observers? *Proc Biol Sci* 264:1429–1434.