Supporting Information

Vouloumanos et al. 10.1073/pnas.0906049106

SI Text

Auditory Stimuli for All Experiments. Speech tokens were 2 Japanese words, *nasu* and *haiiro*, spoken by 2 native female speakers. (One early participant heard *neko* instead of *nasu*.) Human nonspeech vocalizations were 2 sounds produced by 2 native female speakers of English. One sound, "mmm," is commonly associated with agreement. The other sound, "haha," is most commonly associated with laughter. Rhesus vocalizations included a coo call associated with food and group movement, and a gekker call used in affiliative interactions, typically between mother and infant. Duck calls were composed of 2 quacks, quack 1 (from http://www.georgetown.edu/faculty/ballc/animals/duck.html) and quack 2 (http://amazingsounds.iespana.es/aseanimal.htm). A squeaky toy provided sound for the pretest.

Using Praat 4.3.35 (58), we extracted length, pitch, and formant frequency measurements (maximum formant setting was at 5,500 Hz for human and duck stimuli, and 8,000 Hz for rhesus stimuli; window length was 0.025 ms). We report the length and mean, minimum and maximum pitch values in Table 1, and mean formant frequencies for manually selected regions in Table 2. We created 40-s sound files for each sound. 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.

Experiment 1

Participants. Twelve (6 female) full-term infants (M = 5 mo, 0 d; range 4 mo, 17 d to 5 m, 10 d) born to predominantly English- and French-speaking parents participated in the study. Data for 6 additional infants were excluded from analysis due to fussiness (n = 3) or equipment malfunction (n = 3). Given that we presented words spoken in Japanese (see below), we ensured that infants had no prior exposure to Japanese. Parents gave informed consent on behalf of their infants. All procedures had been approved by the institutional research ethics boards at McGill University and New York University.

Complete Study Design. The study design consisted of 2 identical blocks of 8 trials, which contained a set of 4 human face trials and a set of 4 rhesus face trials. Each block was prefaced by a checkerboard pretest trial, which helped familiarize infants with the procedure. Within the human face set, infants saw the 2 human faces, each presented twice, once with a human vocalization and once with a rhesus vocalization. Thus an infant would see one human face (HF1) paired with either a monkey sound (MS1) or human sound (HS1) and see the other human face (HF2) paired with the other monkey sound (MS2) or human sound (HS2). In the next set, infants would see the 2 monkey faces (MF1 and MF2), each presented twice, and hear the same 4 sounds (HS1, MS1, HS2, MS2). Sounds and faces were alternated so that infants would never hear vocalizations of the same species or see the same face twice in a row. All sounds were presented at 65 ± 5 dB. Seven of the infants saw HF first, and 5 saw MF first (order was not significant; see findings). For half the infants, the first trial was a match (HFHS or MFMS) and for the other half, the first trial was a mismatch (HFMS or MFHS). Once initiated, a trial continued until the infant terminated the trial by looking away from the screen for more than 2 s (26) or until they had looked for the maximum trial length of 40 s.

Every infant's looking time during each trial was coded offline. Infants' average looking times for each trial were calculated based on frame-by-frame offline coding (30 frames per second) of infant looks toward the monitor during each test trial by a coder blind to the experimental condition being tested. To ensure reliability in the looking measures, a second observer coded 25% of trials, randomly selected (Pearson's $r = 0.998$, $P < 0.001$). To normalize for variability in infant looking times across the different faces, raw looking times (RLT) were converted to proportion looking time (PLT) within each block. Specifically, for each block of faces (human or monkey) we derived a PLT that represents the relative amount of time infants looked at a given species' face when it was presented with a matching vocalization versus a nonmatching vocalization. For example, the following equation was used for the human match PLT = $[\mathrm{RLT}_{\mathrm{HFHS}}/(\mathrm{RLT}_{\mathrm{HFMS}} + \mathrm{RLT}_{\mathrm{HFHS}})]$, and the following equation yielded the monkey match PLT = $[\mathrm{RLT}_{\mathrm{MFMS}}/(\mathrm{RLT}_{\mathrm{MFMS}} + \mathrm{RLT}_{\mathrm{MFHS}})]$.

Experiment 2

Participants. Fifteen healthy, full-term infants (8 female; range: 4 mo, 19 d to 5 mo, 16 d) with no experience with Japanese participated in the study. The data for 7 additional infants were excluded from analysis due to fussiness (n = 2) or inattention (n = 5).

Complete Study Design. Infants were tested in the same apparatus and infant-controlled procedure as in experiment 1 in which they viewed a series of trials composed of a static face and a repeating sound stimulus. Experiment 2 consisted of 2 blocks of 6 trials; blocks contained one set of 3 human trials and one set of 3 monkey trials to allow each species' face to be paired with each of the 3 species' vocalizations. For example, the human set alternated between the 2 human faces, presented in tandem with a speech token, a rhesus call, and a duck call in turn. Although within each set one face would occur twice and the other only once, across both blocks each face occurred 3 times in total, once with each of the 3 vocalizations. In each block, the order of the pairings was identical, but pairings were instantiated over different tokens of the stimuli. For example, if block 1 consisted of pretest, MF2-HS2, MF1-DS2, MF2-MS2, HF2-HS2, HF1-DS2, and HF2-MS2, block 2 would consist of pretest, MF1-HS1, MF2-DS1, MF1-MS1, HF1-HS1, HF2-DS1, and HF1-MS1. Thus, in total, MF2 would have been presented with one human vocalization (block 1), one monkey vocalization (block 2), and one duck vocalization (block 2). Eight infants saw HF first, and 7 saw MF first (order was not significant). Five infants heard DS first, 5 heard HS first, and 5 heard MS first (again, order was not significant).

Infants' looking time was coded offline to establish total looking time for each trial, and 25% of the trials, randomly selected, were coded for reliability (Pearson's $r = 0.999$, $P < 0.001$). Raw looking times were then converted to proportion looking times as for experiment 1. For example, the following equation was used for the human match PLT = $[\mathrm{RLT}_{\mathrm{HFHS}}/(\mathrm{RLT}_{\mathrm{HFMS}} + \mathrm{RLT}_{\mathrm{HFHS}} + \mathrm{RLT}_{\mathrm{HFDS}})]$.

Experiment 3

Participants. Fifteen healthy, full-term infants (7 female; range: 4 mo, 20 d to 5 mo, 15 d) with no experience with Japanese participated in the study. The data for 6 additional infants were excluded from analysis due to fussiness (n = 3), inattention (n = 2), or only looking away from the visual display twice during the entire study (n = 1).

Complete Study Design. Infants viewed a series of trials composed of a static face and a repeating sound stimulus. Experiment 3 had exactly the same structure as experiment 2 except that duck faces were shown instead of monkey faces. Infants' looking time was coded offline to establish total looking time for each trial, and 25% of the trials, randomly selected, were coded for reliability (Pearson's $r = 0.995$, $P < 0.001$).

Experiment 4

Participants. Twelve healthy, full-term infants (6 female; range: 4 mo, 25 d to 5 mo, 17 d) with no experience with Japanese participated in the study. Data for 4 additional infants were excluded from analysis due to crying ($n = 1$), parent stopping the study ($n = 1$), or experimenter error ($n = 2$).

Complete Study Design. Infants were tested in the same infant-controlled procedure as in experiments 2 and 3, except that instead of being paired with faces, the 3 types of sounds were presented in tandem with a black and white checkerboard on every trial. The order of sound presentation was the same as in experiments 2 and 3.

Infants' looking time was coded offline to establish total looking time for each trial, and a random selection of 25% of the trials were coded for reliability (Pearson's $r = 0.991$, $P < 0.001$).

Table S1. Length (in seconds) and mean, minimum, and maximum pitch (in Hz) of auditory stimuli

Sound category	Sound token	Length, s	Mean pitch, Hz	Minimum pitch, Hz	Maximum pitch, Hz
Human speech	Haiiro	0.663	194	168	209
	Nasu	0.478	248	198	289
Human nonspeech	Agreement	0.688	191	149	292
	Laughter	0.520	228	75	314
Rhesus calls	Coo	0.508	330	304	484
	Gekker	0.659	Undefined	Undefined	Undefined
Duck calls	Quack1	0.487	289	256	304
	Quack2	0.602	237	217	245

Table S2. Mean formant frequencies for vocalic-like regions of auditory stimuli

Sound category	Sound token	Start time, s	End time, s	Duration, s	Formant 1, Hz	Formant 2, Hz	Formant 3, Hz	Formant 4, Hz
Human speech	Haiiro	0.139	0.214	0.075	429	2,344	2,978	4,372
	Nasu	0.065	0.128	0.063	765	1,785	3,101	4,695
Human nonspeech	Agreement	0.171	0.221	0.050	337	1,818	3,107	3,828
	Laughter	0.232	0.276	0.044	278	1,783	2,622	3,835
Rhesus calls	Coo	0.216	0.278	0.063	767	2,636	4,309	5,959
	Gekker	0.238	0.306	0.069	2,511	3,681	4,194	5,713
Duck calls	Quack1	0.077	0.127	0.050	606	1,656	2,143	2,999
	Quack2	0.065	0.147	0.081	1,188	1,819	2,392	3,549

The start time and end time (in seconds) mark the beginning and end points within the sound, and duration (in seconds) of the segment for which frequency means for the first 4 formants were calculated (in Hz).